**Evaluation and Prediction of Heart Disease**

A project work done in partial fulfilment of the “Certificate course on Data Analytics & Business Intelligence”



Submitted by:

Saraansh Chikara (Roll No:DA-10/15)

Certificate Course on Data Analytics & Business Intelligence Batch-10 Shaheed Sukhdev College of Business Studies

May, 2024

**Acknowledgement**

I would like to express my deepest gratitude to several individuals who have contributed significantly to the completion of this project. First and foremost, I extend my sincere appreciation to my project supervisor, Professor Rishi Ranjan. His guidance, unwavering support, and wealth of knowledge have been invaluable throughout the entire journey. Professor Rishi Rajan's mentorship not only enriched the technical aspects of the project but also provided me with valuable insights and perspectives that have undoubtedly shaped the project's success. I am also thankful to the faculty members of Data Analysis and Business Intelligence Course at SSCBS, who have fostered an environment of academic excellence and provided me with a platform to explore and apply my skills in real-world projects. I would like to express my gratitude to the researchers and authors whose work served as a foundation for my project. Their contributions have been instrumental in shaping the theoretical framework and methodology employed in this study. Lastly, but certainly not least, I am thankful to my family for their unwavering support and understanding throughout this academic endeavour. Their encouragement has been a driving force behind my perseverance and commitment to this project. This project would not have been possible without the collective support of these individuals. Thank you all for being a part of this academic journey.

-Saraansh Chikara

**Declaration**

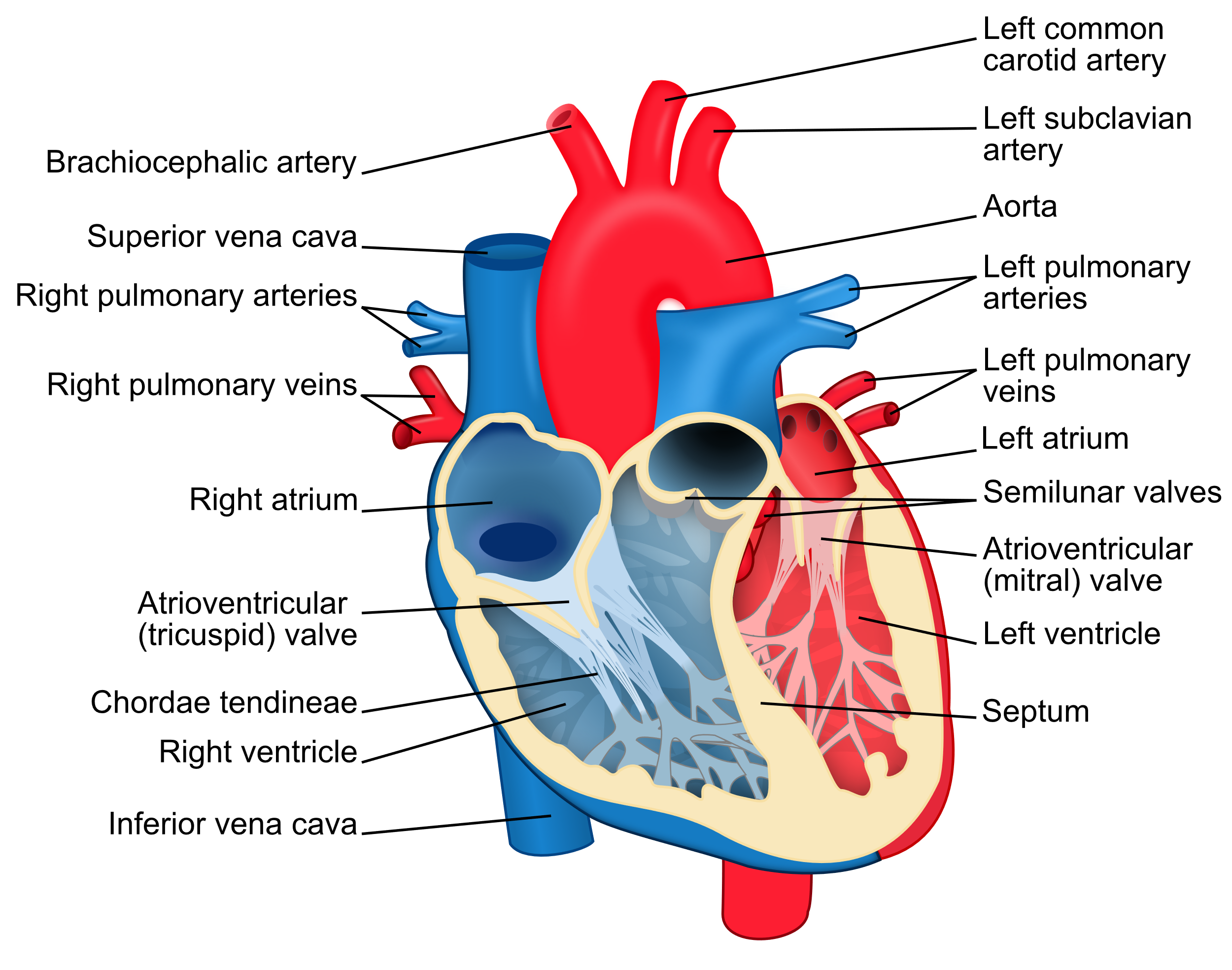
I, Saraansh Chikara, declare that this project titled “Heart Disease Prediction” is the original work done by me under the guidance of Dr. Rishi Rajan Sahay Professor, Shaheed Sukhdev College of Business Studies, University of Delhi.

Any existing work, which has been referred to, while making this project, has been stated and cited explicitly under the References/Bibliography of the Project.

I further declare that it is the original work made by me as a part of my Certificate Course on Data Analytics and Business Intelligence

-Saraansh Chikara

**Introduction**



Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels. Heart attacks and strokes are usually acute events and are mainly caused by a blockage that prevents blood from flowing to the heart or brain

Cardiovascular diseases (CVDs) are the leading cause of death globally. An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke. Over three quarters of CVD deaths take place in low- and middle-income countries. Out of the 17 million premature deaths (under the age of 70) due to noncommunicable diseases in 2019, 38% were caused by CVDs.It is important to detect cardiovascular disease as early as possible so that management with counselling and medicines can begin.

Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol.

Cessation of tobacco use, reduction of salt in the diet, eating more fruit and vegetables, regular physical activity and avoiding harmful use of alcohol have been shown to reduce the risk of cardiovascular disease.

Heart disease describes a range of conditions that affect the heart. Heart diseases include:

* Blood vessel disease, such as coronary artery disease
* Irregular heartbeats (arrhythmias)
* Heart problems you're born with (congenital heart defects)
* Disease of the heart muscle
* Heart valve disease

Heart attacks and strokes are usually acute events and are mainly caused by blockage that prevents blood from flowing to the heart or brain. The most common reason for this is a build-up of fatty deposits on the inner walls of the blood vessels that supply the heart or brain

At least three-quarters of the world's deaths from CVDs occur in low- and middle-income countries. People living in low- and middle-income countries often do not have the benefit of primary health care programmes for early detection and treatment of people with risk factors for CVDs. People in low- and middle-income countries who suffer from CVDs and other noncommunicable diseases have less access to effective and equitable health care services which respond to their needs. As a result, for many people in these countries detection is often late in the course of the disease and people die at a younger age from CVDs and other noncommunicable diseases, often in their most productive years.

 In this Project I have predicted the presence (1) or absence (0) of heart disease, which is represented in the ‘target’ column. This is a classic example of a binary classification problem, and we can predict the values using various machine learning models such as Logistic Regression and Decision Trees.

**Logistic Regression**



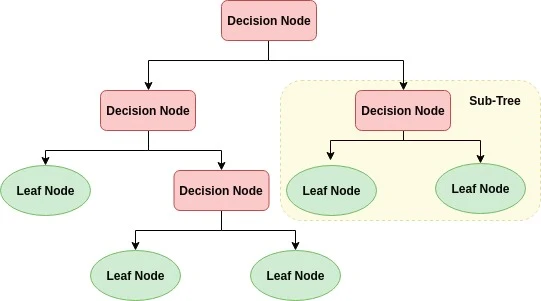
Logistic regression is a statistical model that is commonly used in machine learning for binary classification problems, which is exactly what i have here: predicting the presence (1) or absence (0) of heart disease.

In the context of this dataset, logistic regression can be used to understand the relationship between several independent variables (age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal) and the dependent variable (target), which represents the presence or absence of heart disease.

The logistic regression model can provide valuable insights into which factors are most influential in predicting the presence of heart disease, as the magnitude of each coefficient 𝑏𝑖bi​ indicates the importance of the corresponding feature 𝑋𝑖Xi​.

After training the model on this dataset, it could be used to predict the likelihood of heart disease in new patients based on their medical measurements.

**Decision Tree**



A decision tree is a type of machine learning model that makes decisions based on a series of questions asked about the data, each question narrowing down the possibilities until a final decision is made. It’s called a “tree” because of the way these questions are structured in a hierarchical manner, similar to the branches of a tree.It is a flowchart-like structure used to make decisions or predictions. It consists of nodes representing decisions or tests on attributes, branches representing the outcome of these decisions, and leaf nodes representing final outcomes or predictions. Each internal node corresponds to a test on an attribute, each branch corresponds to the result of the test, and each leaf node corresponds to a class label or a continuous value.

The process of creating a decision tree involves:

1. Selecting the Best Attribute: Using a metric like Gini impurity, entropy, or information gain, the best attribute to split the data is selected.
2. Splitting the Dataset: The dataset is split into subsets based on the selected attribute.
3. Repeating the Process: The process is repeated recursively for each subset, creating a new internal node or leaf node until a stopping criterion is met (e.g., all instances in a node belong to the same class or a predefined depth is reached).

**Main Work/Contribution**

The main task with this dataset was to use these features to build a predictive model that can accurately determine whether a new patient has heart disease or not. This is an example of a binary classification problem, and it can be done with various machine learning algorithms such as Logistic Regression, Decision Trees. We considered Heart Disease Dataset from Kaggle.

**1)Dataset**

Contents of the Dataset include:

* **age**: The person’s age in years.
* **sex**: The person’s sex (1 = male, 0 = female).
* **cp**: The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic).
* **trestbps**: The person’s resting blood pressure (mm Hg on admission to the hospital).
* **chol**: The person’s cholesterol measurement in mg/dl.
* **fbs**: The person’s fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false).
* **restecg**: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy).
* **thalach**: The person’s maximum heart rate achieved.
* **exang**: Exercise induced angina (1 = yes; 0 = no).
* **oldpeak**: ST depression induced by exercise relative to rest (‘ST’ relates to positions on the ECG plot).
* **slope**: The slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping).
* **ca**: The number of major vessels (0-3).
* **thal**: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect).
* **target**: Heart disease (0 = no, 1 = yes).

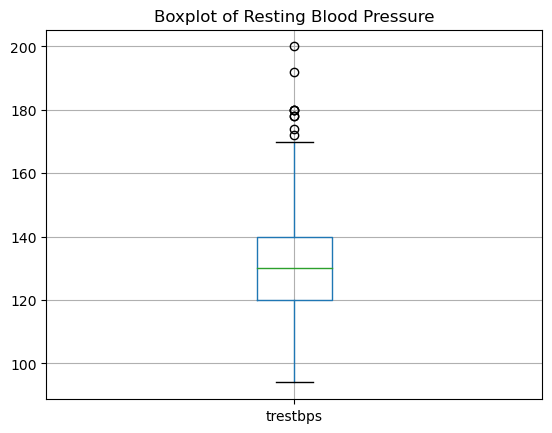
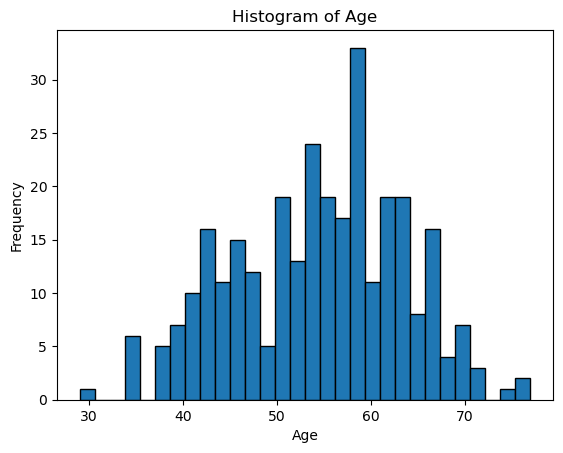
**2)Data Cleaning**

Data cleaning is an important step in any data analysis process. It involves preparing the data for analysis by removing irrelevant data or modifying data that is incorrect, incomplete, duplicated. In our dataset, the number of null values was 0 so there was no need for data cleaning.

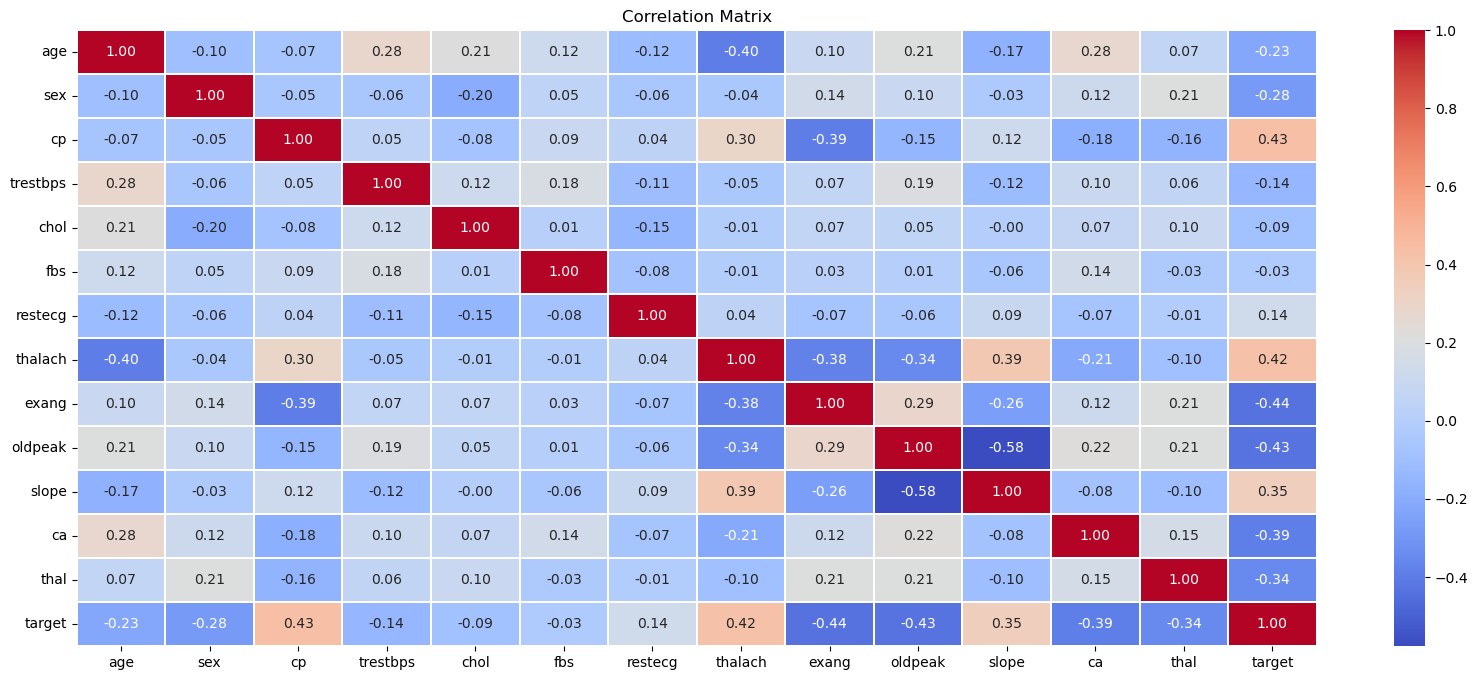
**3)Data Visualisation**

Data visualization is the graphical representation of information and data. It uses visual elements like charts, graphs, and maps to provide an accessible way to see and understand trends, outliers, and patterns in data. Here are some visualisation like Histogram(of Age),Box Plot (for Resting Blood Pressure) and Corelation Matrix.

1)Histogram of Age 2) Boxplot of Resting Blood Pressure



3)Corelation Matrix



We observe that target and cp(chest pain) have a high corelation as chest pain is most important in identifying heart diseases.

**4)Data Modelling**

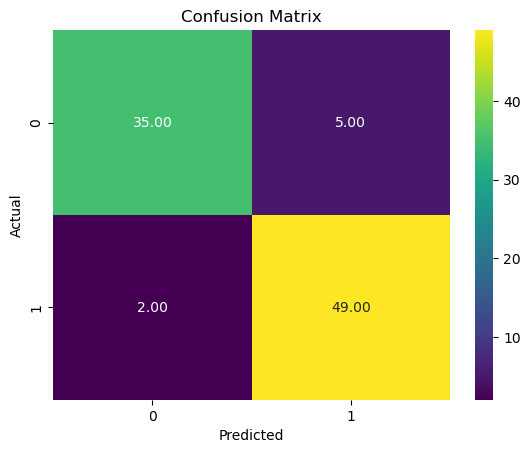
The classification algorithm used is Logistic Regression and Decision Tree. When we train the model, the dataset was split into a 70:30 ratio, allocating 70 percent of the data for training the model and reserving the remaining 30 percent for evaluating model accuracy

**Logistic Regression**

1)Accuracy score is :0.9230769230 4)Precision score is:0.9074074074074

2)Training score is :0.83490566 5)Testing score is :0.92307692

3)Recall score is :0.96078431 6)F1 score is :0.93333333

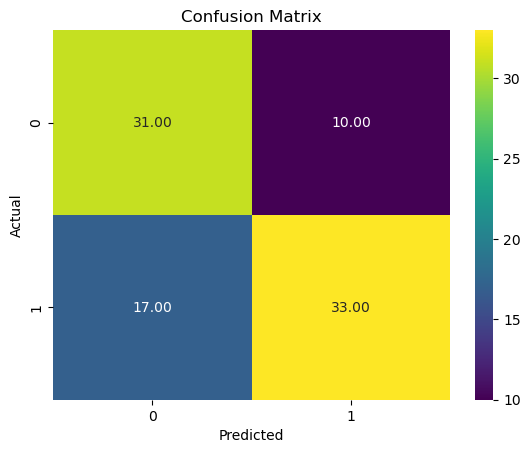


**Decision Tree**

1)Accuracy score is :0.7032967032 4) Precision score is:0.76744186046511

2)Training score is :0.85377358490 5) Recall score is :0.9607843137254

3)Testing score is :0.846153846153 6) F1 score is :0.9333333



**Conclusion**

**Logistic Regression** gives a statistical analysis of the relationships between the features and the target variable, whereas **Decision Tree** model provides a clear visualization of the decision-making process, showing how different conditions on the features lead to the final prediction.

Here we can see that chest pain , thalach (maximum heart rate achieved), exang (exercise induced edema) have high corelation with the output (output being a person has heart disease or not).Increase in these levels ,increases a persons chances of having heart diseases.

According to our models, Logistic Regression is the best approach in predicting the outcomes of heart diseases as it has an accuracy score of 0.9230769230 which is considered to be very good, but since the data is limited, we need to gather more data for improving out results.

**References**

1)<https://www.mayoclinic.org/>

2)<https://www.who.int/>

3) <https://www.kaggle.com/>

4) <https://chatgpt.com/>

.