TOPIC : HEART DISEASE PREDICTION USING LOGISTIC REGRESSION

Name: Tanwistha Guha
ID: 231001001437
Team 1– PBEL-AI & ML (IBM)

Problem Statement:

I. Heart disease remains one of the leading causes of death worldwide, and early detection is difficult without proper analysis of patient health data.

II. Raw healthcare datasets often contain missing, inconsistent, or noisy data, which reduces the accuracy of predictive models.

III. Understanding the relationship between multiple health factors such as age, cholesterol, and blood pressure is complex and difficult to interpret manually.

Solution Proposed:

I. Develop a backend system that preprocesses and analyzes patient medical data to identify early risk patterns using data analytics techniques.

II. Implement data cleaning, normalization, and exploratory data analysis (EDA) using **NumPy** and **Pandas** to prepare high-quality input data for heart disease prediction.

III. Use **Matplotlib** and **Seaborn** to visualize correlations and trends among key attributes, enabling better insights and feature selection for the predictive model.
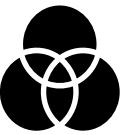
Why logistic regression?

- Logistic Regression was chosen for this project because it is one of the most effective and interpretable algorithms for **binary classification problems**, such as predicting whether a patient has heart disease (**Yes/No**). It models the probability of an event occurring based on input variables like age, cholesterol, and blood pressure.
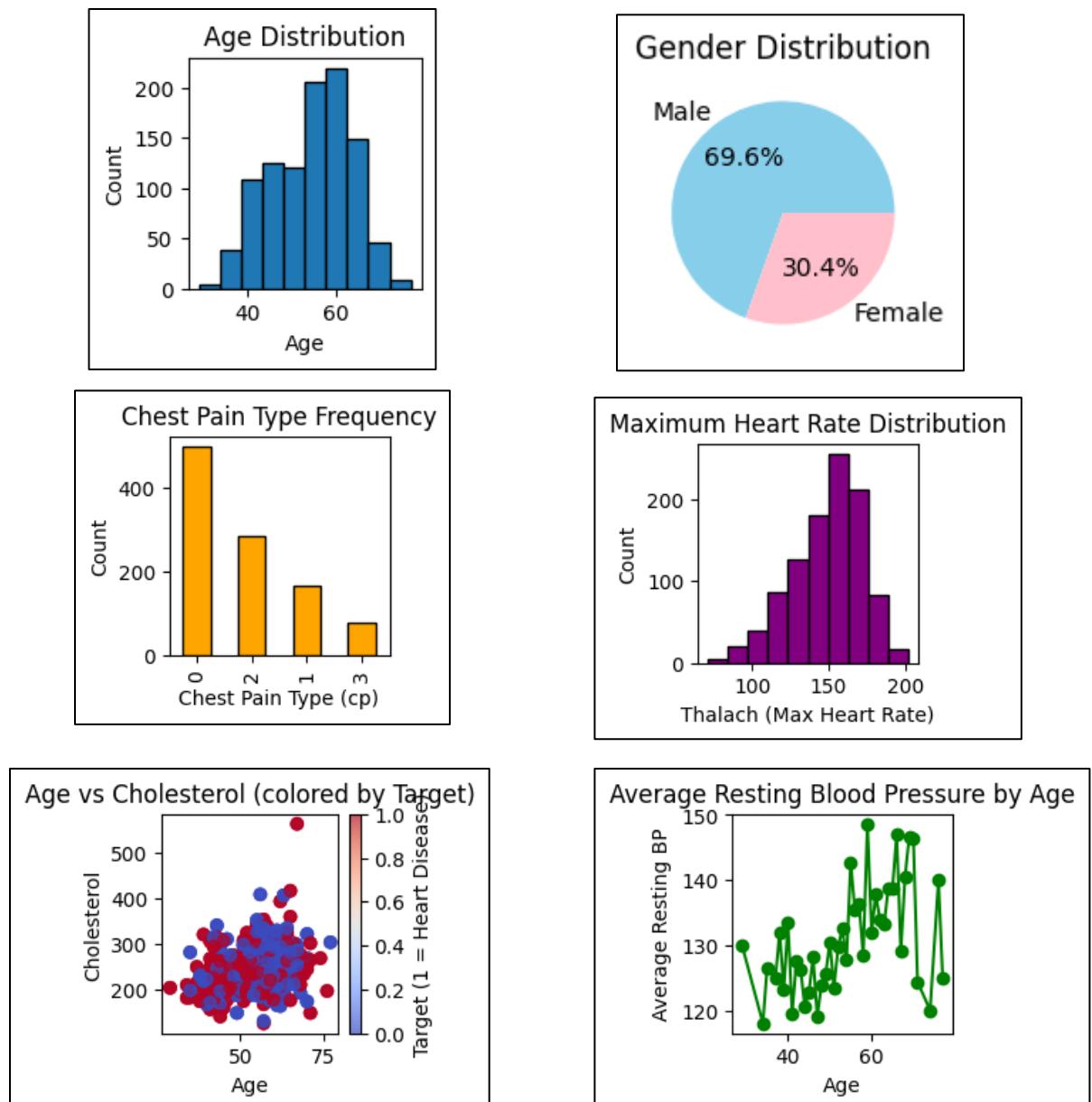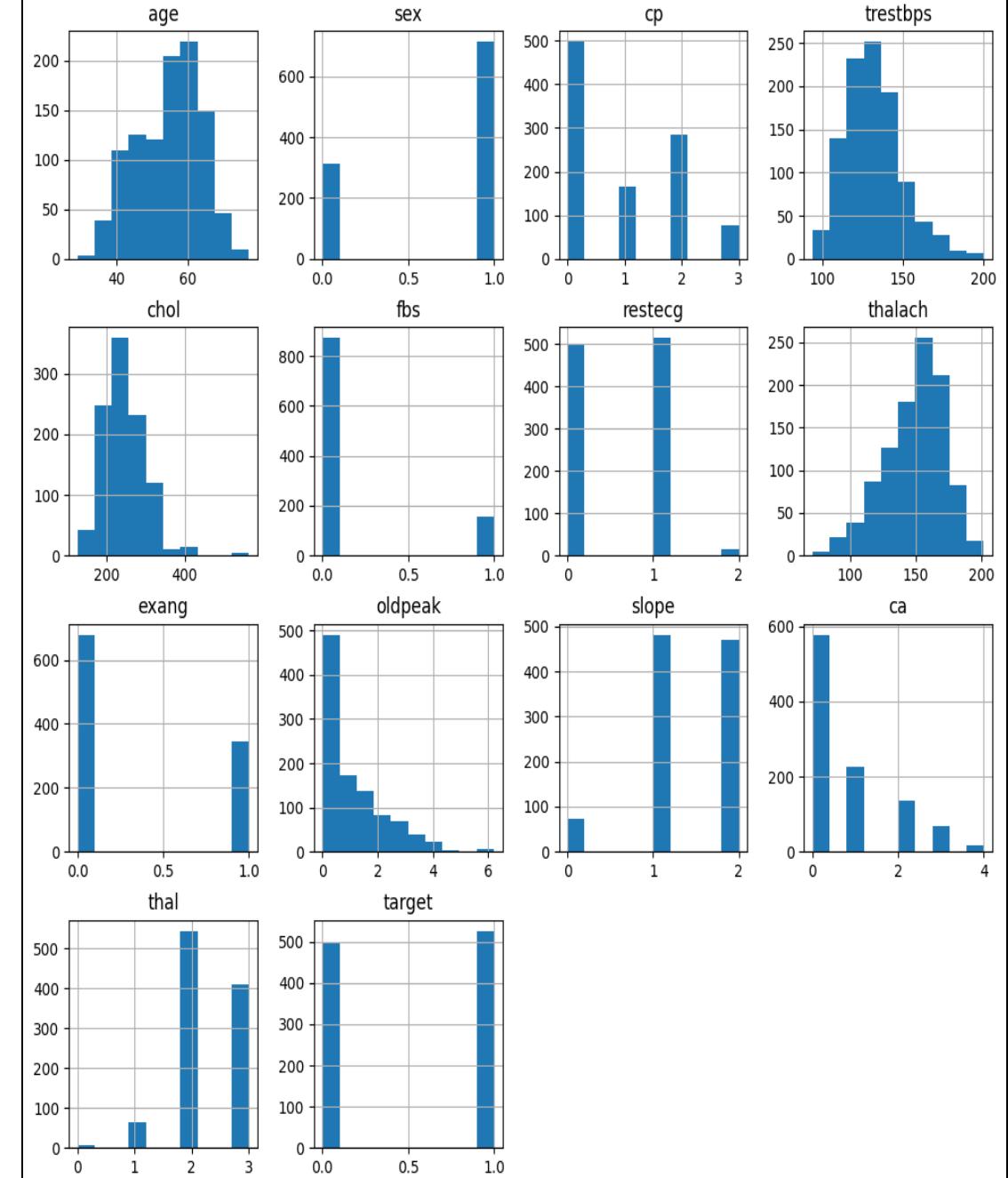
Key Reasons:

- **Binary Outcome:** Heart disease prediction is a two-class problem (disease present or not), which fits perfectly with Logistic Regression's design.

- **Interpretability:** The model provides clear insights into how each health factor contributes to the risk, making it ideal for healthcare applications where explainability is crucial.

- **Efficiency:** Logistic Regression is computationally lightweight and performs well even with moderate datasets, which is ideal for quick backend implementation.

- **Baseline Model:** It serves as a strong baseline before moving to more complex models like Random Forest or Neural Networks.

Tech stack used

Layer	Technology	Purpose
Programming Language	🐍 **Python**	Core language for ML model and backend
Machine Learning Library	🤖 **Scikit-learn (sklearn)**	Model training, prediction, scaling, and tuning
Data Handling	📃 **Pandas, NumPy**	Loading and manipulating data (`heart.csv`)
Data Visualization	📊 **Matplotlib, Seaborn**	EDA, correlation plots, and histograms
Version Control (Optional)	🧩 **Git / GitHub**	Managing and sharing project code

Various charts

Learning outcome. ⧗

- The Heart Disease Prediction project successfully demonstrates how data analysis and visualization can help in understanding key factors that influence heart health. By using Python libraries such as **NumPy**, **Pandas**, and **Matplotlib**, we were able to clean, analyze, and visualize important health parameters like age, cholesterol level, blood pressure, and heart rate.

- Through various plots—such as histograms, pie charts, bar graphs, and scatter plots—we gained valuable insights into patterns within the dataset. For example, we observed how age and cholesterol correlate with heart disease, and how factors like chest pain type and resting blood pressure vary across individuals.

- This backend project provides a strong foundation for future integration with a predictive model that can help in early detection of heart disease using machine learning techniques.