# Report on Heart Attack Dataset

Byte Brigade

14th July, 2024

## Objective

The goal of this project is to create an AI-driven solution that can efficiently process and analyze structured heart attack datasets to represent knowledge and generate valuable insights. This solution should identify patterns within the data and produce meaningful information to support decision-making processes.

## Problem Description

In today's era of big data, organizations across various sectors generate enormous amounts of data daily. Properly processing and analyzing this data can uncover valuable insights that greatly enhance decision-making processes. The challenge lies in effectively representing this knowledge and extracting useful insights. Your task is to develop an AI-based solution capable of addressing this challenge by processing a provided structured dataset, representing the knowledge within it, and generating meaningful insights.

## Dataset Source

`www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset`

## Key Features

The key feature of the dataset is that it contains reports from 304 patients on the likelihood of experiencing a heart attack, based on the following factors:

- Age: Age of the patient
- Sex: Gender of the patient (1 = male; 0 = female)
- CP: Chest pain type (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic)
- TRTBPS: Resting blood pressure (in mm Hg)
- CHOL: Serum cholesterol in mg/dl
- FBS: Fasting blood sugar (¿ 120 mg/dl, 1 = true; 0 = false)
- RESTECG: Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy)
- THALACHH: Maximum heart rate achieved
- EXNG: Exercise induced angina (1 = yes; 0 = no)
- OLDPEAK: ST depression induced by exercise relative to rest
- SLP: Slope of the peak exercise ST segment (0 = upsloping, 1 = flat, 2 = downsloping)
- CAA: Number of major vessels (0-3) colored by fluoroscopy
- THALL: Thalassemia (1 = normal; 2 = fixed defect; 3 = reversible defect)

## Methods Used

- Clustering Algorithm: K-Means
- Classification Algorithms: Decision Tree Classifier, Logistic Regression
- Regression Algorithms: Linear Regression, Decision Tree Regressor

## Tools

- Python Libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn

# Results & Findings

# Key Findings from Heart Attack Risk Analysis

### 1. Distribution of Heart Attack Risk

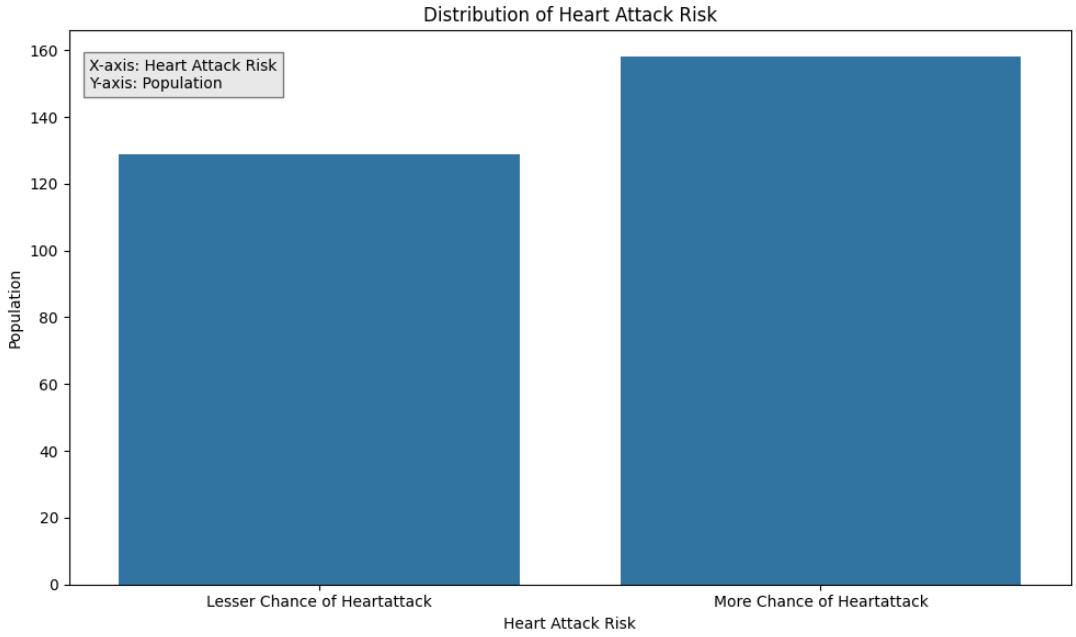


Figure 1: Distribution of Heart Attack Risk

- Population Split:
  - Lesser Chance of Heart Attack: 129 patients
  - More Chance of Heart Attack: 158 patients
- Implications: Balanced dataset, important for training predictive models and interpreting performance metrics.

## 2. Age vs. Cholesterol

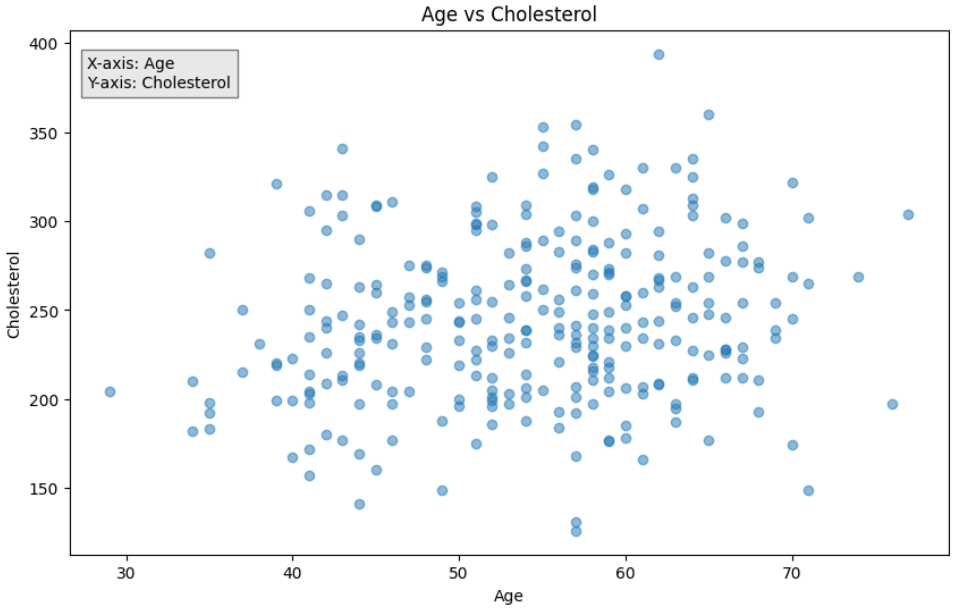


Figure 2: Age vs. Cholesterol

- Correlation: Weak positive correlation (0.21), indicating slight increase in cholesterol with age.
- Notable Cases: Oldest person (77 years) with cholesterol 304 mg/dl. Highest cholesterol (564 mg/dl) at age 67.
- Health Implications: Monitoring cholesterol levels across age groups is crucial for assessing cardiovascular health risks.

## 3. Correlation Matrix Insights

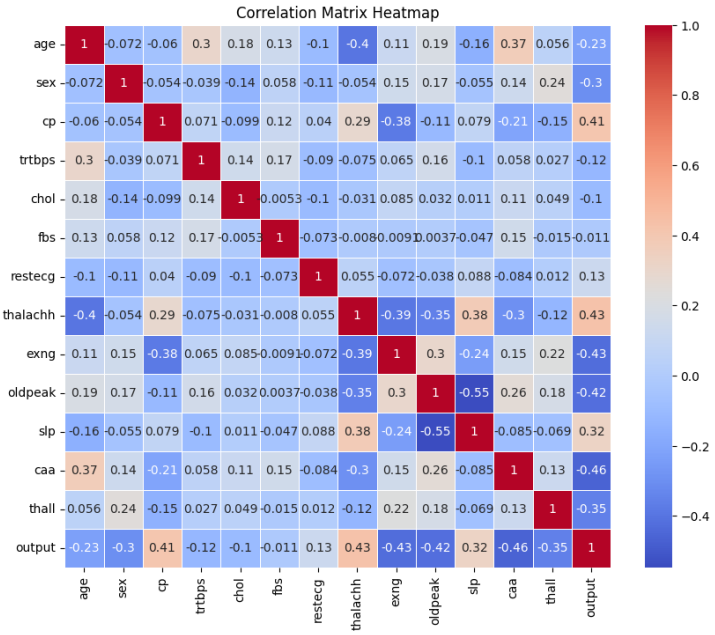


Figure 3: Correlation Matrix Insights

- Age and Health Factors: Moderate positive correlation with resting blood pressure, moderate negative correlation with maximum heart rate achieved.
- Sex and Heart Health: Males might have a slightly higher likelihood of experiencing a heart attack.
- Heart Health Indicators: Moderate correlations with maximum heart rate, chest pain type, exercise-induced angina, and number of major vessels.

## 4. Age vs. Maximum Heart Rate Achieved

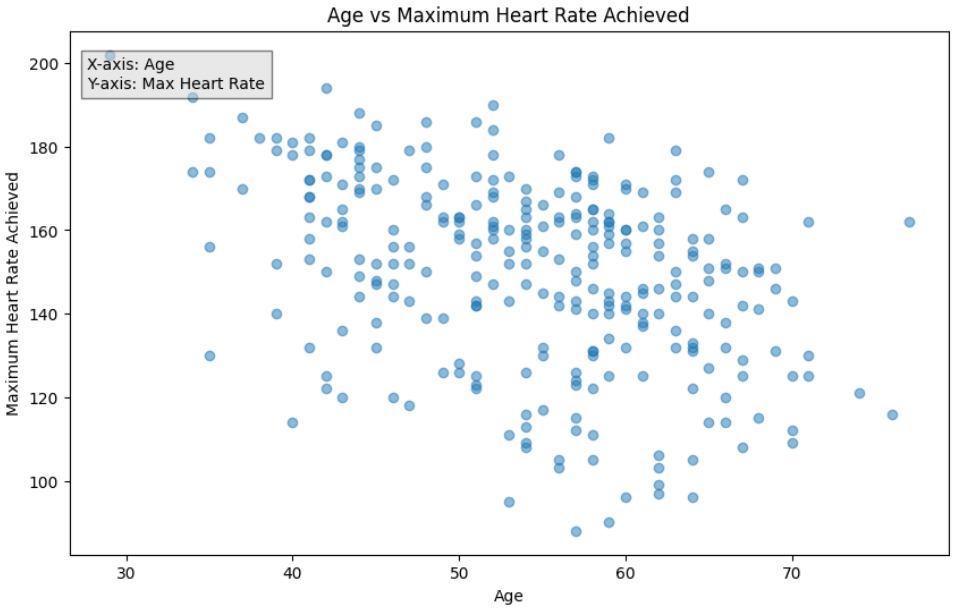


Figure 4: Age vs. Maximum Heart Rate Achieved

- Mean Values: Average age: 54.37 years, Average maximum heart rate: 149.65 bpm.
- Correlation: Moderate negative correlation (-0.40), suggesting that as age increases, maximum heart rate tends to decrease.

## 5. Chest Pain Types by Sex

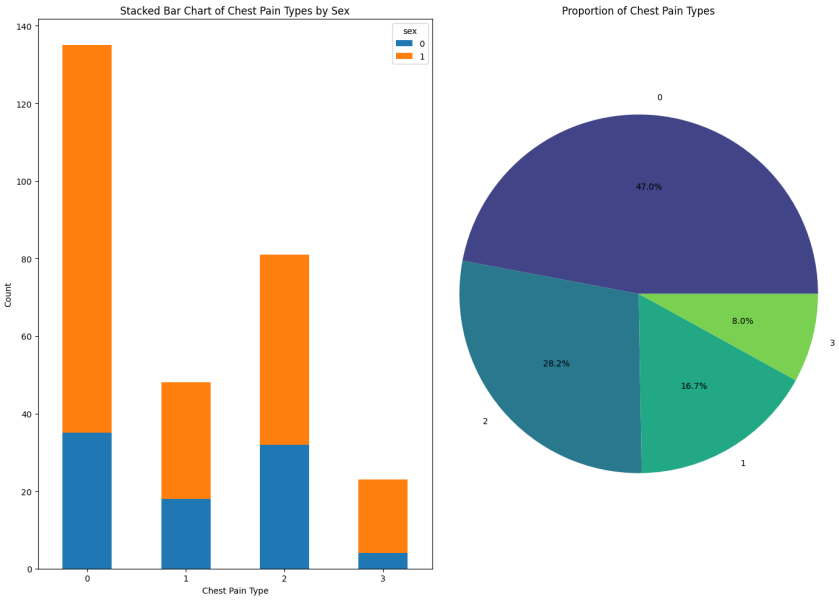


Figure 5: Chest Pain Types by Sex

- Distribution:
  - Typical Angina: Higher prevalence among males.
  - Atypical Angina: Similar distribution between males and females.
  - Non-anginal Pain: Slightly higher prevalence among females.
  - Asymptomatic: Higher prevalence among males.

## 6. Cholesterol Levels

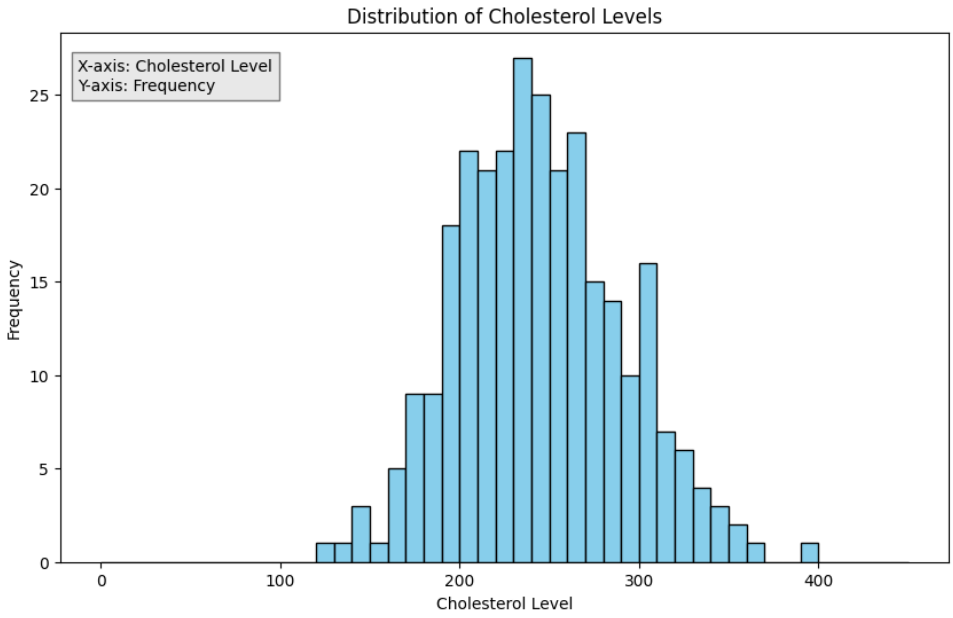


Figure 6: Cholesterol Levels

- Peak Frequencies: Most common levels: 200-250 mg/dl.
- Distribution: Clustering between 170-310 mg/dl.
- Health Implications: High frequencies in certain ranges indicate a significant portion of patients with borderline high or high cholesterol, a concern for heart disease risk.

## 7. Key Relationships (Graph Representation)

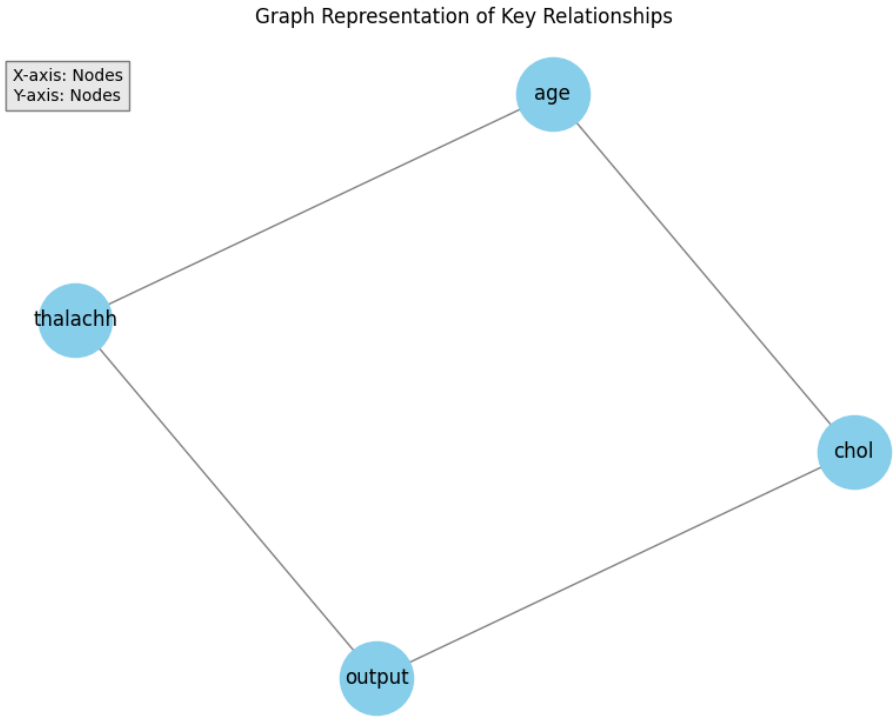


Figure 7: Key Relationships (Graph Representation)

- Highlighted Relationships:
  - Age vs. Maximum Heart Rate Achieved
  - Age vs. Cholesterol Level
  - Maximum Heart Rate Achieved vs. Heart Attack Risk
  - Cholesterol Level vs. Heart Attack Risk

## 8. Visualization

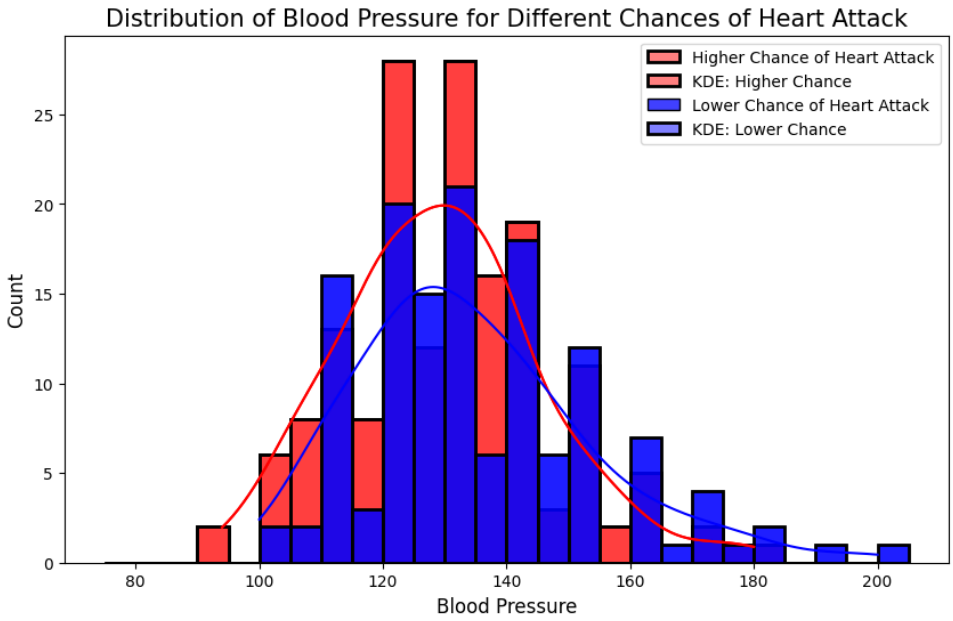


Figure 8: Visualization of Blood Pressure and Heart Rate for Heart Attack Risk

- Helps understand variable relationships for further analysis or predictive modeling. Blood Pressure and Heart Rate for Heart Attack Risk.
- Blood Pressure: Certain intervals (120-125 and 130-135 mmHg) indicate critical ranges for cardiovascular risk assessment.
- Heart Rate: Higher heart rates show a higher incidence of heart attack, with notable risk shifts in certain intervals.

## 9. Pair Plot Insights

- Age: Positive correlation with resting blood pressure, negative correlation with maximum heart rate achieved.
- Effect of Heart Attack Risk: Clear differences in variable distributions between lower and higher risk groups.
- Analytical Use: Valuable for exploratory data analysis and understanding interactions among health metrics.

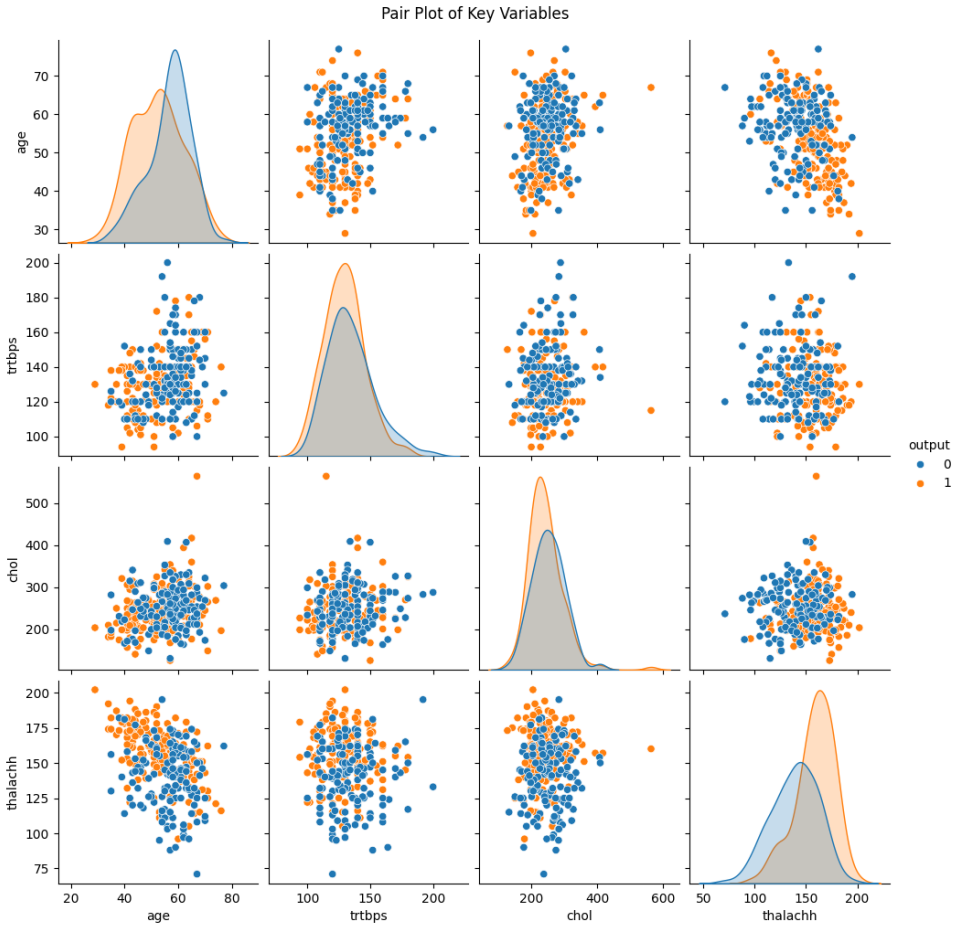


Figure 9: Pair Plot Insights

# Future Work

Requires the Linear Regression and Decision Tree Regressor to be more accurate and also requires a massive dataset for better machine learning algorithm's accuracy.

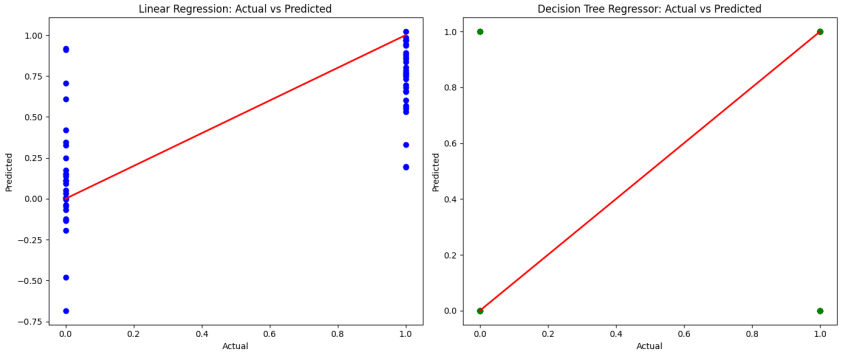


Figure 10: Regression Models