

# EAS 503 Final Project

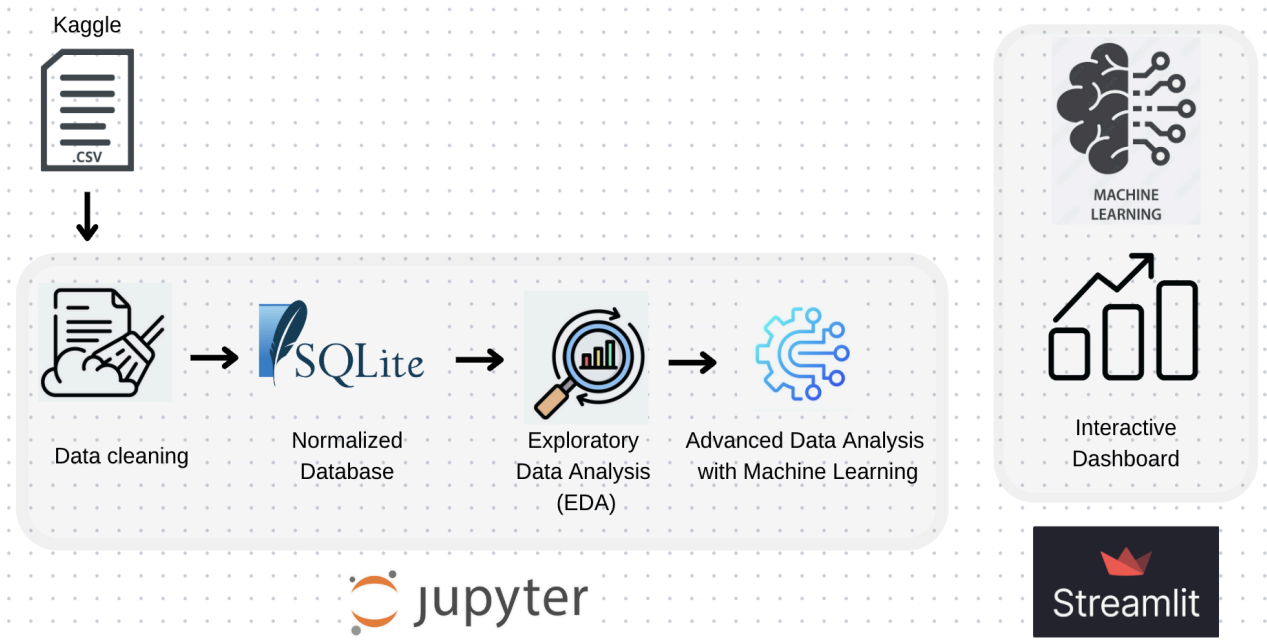
Title : Heart Attack Analysis and Prediction

Course : EAS 503 Python for Data Scientists (Fall 2024)

Team 13

Member Name	UB ID
Jeevaharan Magudanchavadi Jayasankar	50604147
Bharathraaj Nagarajan	50556348
Harish Suresh Babu	50604190

## Project Architecture



## 1.Dataset Collection and Justification

### Why is this dataset relevant to your chosen topic?

- Cardiovascular diseases are among the leading causes of death globally, making heart attack prediction a critical area of research.
- The dataset addresses a significant health issue by providing information that can be used to predict the likelihood of a heart attack. It contains patient-level clinical and demographic data, which are essential for analyzing risk factors and predicting outcomes.

Source of Data (Kaggle) : Heart Attack Analysis and Prediction

### Key variables and their importance to your analysis.

The dataset includes both numerical and categorical features that are critical in understanding the factors contributing to heart health:

- **Age:** Age is a significant risk factor for heart disease. Older individuals are generally at higher risk.
- **Sex:** Gender-specific differences in heart disease risk provide valuable insights.
- **Chest Pain Type (cp):** Different types of chest pain may indicate varying severity and causes, which are critical for diagnosis.
- **Resting Blood Pressure (trestbps):** High blood pressure is a well-known risk factor for cardiovascular disease.
- **Cholesterol (chol):** Elevated cholesterol levels are strongly linked to the development of atherosclerosis and heart attacks.
- **Fasting Blood Sugar (fbs):** Diabetes is a known contributor to cardiovascular disease.
- **Maximum Heart Rate Achieved (thalachh):** A key indicator of cardiovascular fitness and heart function.
- **Exercise-Induced Angina (exng):** Helps identify symptoms triggered by physical exertion.
- **ST Depression (oldpeak):** Indicates stress-induced abnormalities in heart function.
- **Target Variable:** A binary variable indicating the presence or absence of a heart attack, making the dataset suitable for classification tasks.

### Anticipated challenges (e.g., missing data, merging multiple datasets)

- **Missing Data:** The dataset may have missing values, requiring careful imputation strategies to avoid biased analysis.
- **Outliers:** Extreme values in clinical data such as blood pressure and cholesterol levels could skew results and must be detected and handled appropriately.

- **Feature Engineering:** Creating new variables, such as age groups or risk scores, could enhance the analysis and predictive modeling.
  - **Data Cleaning:** Ensuring data type consistency, removing duplicates, and addressing anomalies are essential for preparing the dataset for analysis.
- 

## 2. Data Wrangling and Cleaning

### Handle Missing Data:

- Checked for missing values in all columns. (No missing values)
- Imputed missing values in numerical columns using the median. (For future purposes, if data has missing values).

### Outlier Detection and Handling:

- Used the Interquartile Range (IQR) method to detect outliers.
- Handled outliers by capping (Winsorization), replacing values beyond IQR bounds with the nearest acceptable value.

### Data Type Consistency:

- Ensured numerical columns are integers or floats.
- Converted categorical columns and the target variable (output) to category type.

### Feature Engineering:

- Added a new feature age\_group to categorize patients into age groups for better analysis.

### Documentation:

- The cleaned dataset is saved for step 3 in csv format.
- 

## 3. SQL Database Design and Querying

### Create a relational Database

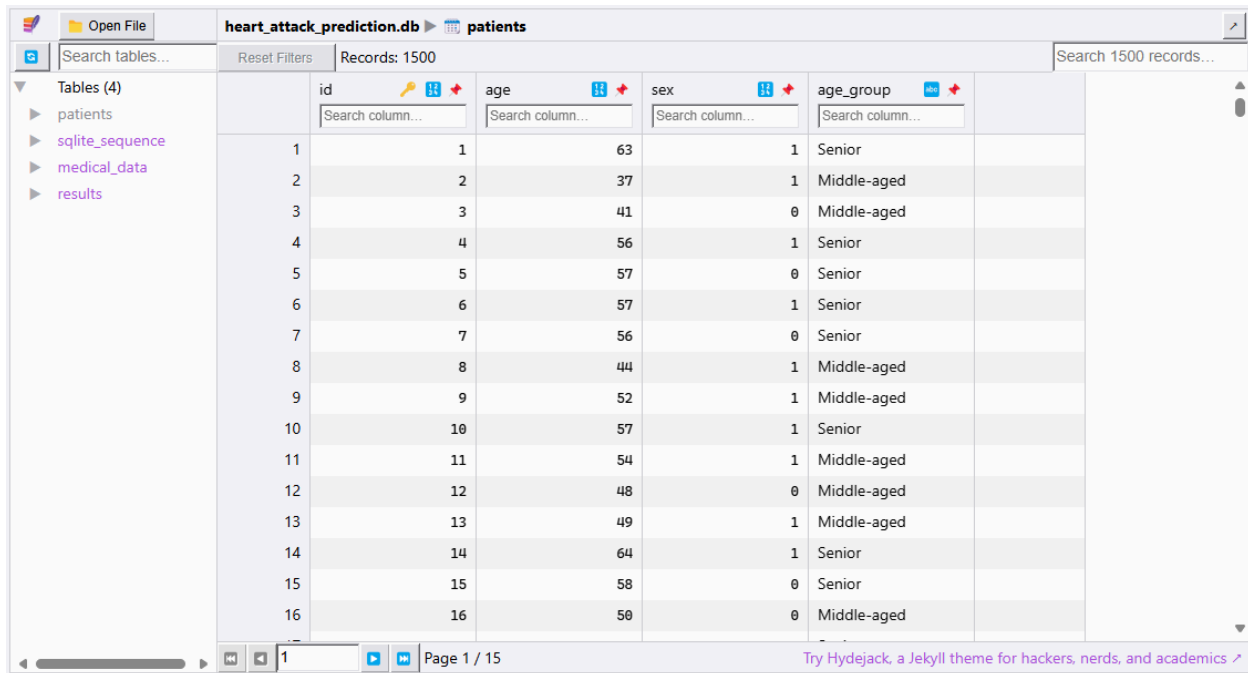
The database utilized here is SQLite.

Database name : heart\_attack\_prediction

## Normalization the tables to 1 NF, 2 NF 3 NF

### Normalised table names :

1. **Patients** : General patient information (e.g., id, age, sex, age\_group).
2. **Medical\_data** : Key Medical indicators (e.g., id, cp, trestbps, chol, thalach, oldpeak, etc.)
3. **Results** : Target and additional derived outcomes (e.g., id, output).



	id	age	sex	age_group
1	1	63	1	Senior
2	2	37	1	Middle-aged
3	3	41	0	Middle-aged
4	4	56	1	Senior
5	5	57	0	Senior
6	6	57	1	Senior
7	7	56	0	Senior
8	8	44	1	Middle-aged
9	9	52	1	Middle-aged
10	10	57	1	Senior
11	11	54	1	Middle-aged
12	12	48	0	Middle-aged
13	13	49	1	Middle-aged
14	14	64	1	Senior
15	15	58	0	Senior
16	16	50	0	Middle-aged

Data stored in SQLite Database (Viewed fromSQLite Viewer Web App)

## Import data from cleaned CSV file

Cleaned data (heart\_data\_cleaned.csv) is read using pandas and the data is inserted into the corresponding tables.

## 4. Exploratory Data Analysis

### Data Retrieval from the Database

All data used for EDA was retrieved efficiently using SQL queries from the relational database. Key subsets extracted include:

- **Complete Dataset:** Combined patient demographics, medical measurements, and outcomes.
- **Summary Statistics:** Calculated averages for trtbps (blood pressure), chol (cholesterol), thalachh (max heart rate), and the total number of heart disease cases.
- **Group-Level Insights:** Patient distribution by age group and gender, along with average and maximum cholesterol levels.

## Univariate Analysis

- **Age Distribution:**
- Patients are concentrated in the 40-70 age range (middle-aged and senior groups), which highlights a high-risk demographic for heart disease.
- **Cholesterol Levels:**
- Cholesterol values exhibit a wide spread with outliers on the higher end, indicating that several individuals have abnormally high cholesterol, a well-known risk factor for heart disease.
- **Heart Attack Outcome:**
- The outcome variable is reasonably balanced, with 42.5% of patients having heart disease and 57.5% without.

## Bivariate Analysis

### Age vs. Cholesterol:

- Older individuals tend to have higher cholesterol levels. Among these, patients with heart disease (output=1) often exhibit elevated cholesterol compared to those without.

### Average Blood Pressure by Age Group:

- Blood pressure increases with age, with the 50-60 and 60+ age groups showing the highest average values. This trend confirms age as a key factor for hypertension and heart disease risk.

## Multivariate Analysis

### Correlation Analysis:

#### Positive Correlations:

- oldpeak (ST depression) shows a strong positive correlation with heart disease outcomes, making it a significant diagnostic indicator.
- cp (chest pain type) has a moderate positive correlation with heart disease.

### **Negative Correlations:**

- thalachh (max heart rate achieved) shows a strong negative correlation with heart disease, suggesting that lower heart rates are associated with higher risk.

### **Pairplot Insights:**

- Visual analysis of selected features (age, chol, trtbps, thalachh, output) reveals patterns where patients with lower thalachh or higher cholesterol are more likely to experience heart attacks.

### **Statistical Metrics**

#### **1. Correlation Coefficients:**

Key relationships include:

- thalachh negatively correlates with output (-0.42), highlighting its importance.
- oldpeak positively correlates with output (0.42), supporting its diagnostic role.

#### **2. Variance and Standard Deviation:**

- Features like chol (cholesterol) have high variance, indicating significant variability in cholesterol levels among patients.

### **Conclusion**

#### **The EDA highlights critical insights:**

- Age, cholesterol, and blood pressure increase with heart disease risk.
- Features like thalachh and oldpeak are statistically significant indicators for predicting heart disease outcomes.
- These findings underscore the importance of monitoring key risk factors to identify high-risk patients and prevent heart disease.

---

## **5. Advanced Data Analysis and Machine Learning**

### **Hypothesis Testing (T-Test)**

- The T-test revealed a statistically significant difference in cholesterol levels between patients with and without a heart attack (p-value = 0.00000, T-statistic = -5.48).

Insight:

Cholesterol levels are significantly higher in patients who experienced heart attacks. This emphasizes the need for cholesterol monitoring and management to reduce cardiovascular risks.

**Regression Analysis (Multiple Linear Regression)**

- The regression model demonstrated that age and chest pain type (cp) are significant predictors of cholesterol levels.
- Age Coefficient: 0.9109 (p-value = 0.000) → Cholesterol increases as age increases.
- Chest Pain Coefficient: -4.9947 (p-value = 0.000) → Chest pain type influences cholesterol inversely.
- The model had an R-squared value of 0.05, indicating that age and chest pain explain 5% of the variation in cholesterol levels.

Insight:

Older patients and those with specific chest pain types are at higher risk of elevated cholesterol, highlighting the need for age-based cholesterol screenings.

**Machine Learning Model Performance**

To predict heart attack outcomes (output), multiple models were evaluated based on accuracy, precision, recall, and F1-score.

Model	Accuracy	Precision (0/1)	Recall (0/1)	F1-Score (0/1)
Logistic Regression	<b>78.67%</b>	0.77 / 0.80	0.73 / 0.83	0.75 / 0.81
Decision Tree	<b>87.00%</b>	0.82 / 0.92	0.90 / 0.85	0.86 / 0.88
K-Nearest Neighbors	<b>93.67%</b>	0.95 / 0.92	0.89 / 0.96	0.92 / 0.94

## Key Observations

### 1. Best Performing Models:

- K-Nearest Neighbors (KNN) achieved 93.67% accuracy with good performance across precision and recall.

### 2. Logistic Regression:

- Performed reasonably well with 78.67% accuracy but showed lower recall for class 0 (heart attack absent).

### 3. Decision Tree:

- Achieved 87.00% accuracy, with strong recall for class 0 (0.90), making it suitable for balanced datasets.

## Actionable Insights

### 1. Model Recommendation:

- Decision Tree can be an alternative model with 87.00% accuracy for its robustness and simplicity.

### 2. Importance of Features:

- Key predictors like age, chest pain type (cp), cholesterol (chol), ST depression (oldpeak), and maximum heart rate (thalach) play a significant role in model performance.
- These features should be monitored closely in clinical settings.

### 3. Addressing Cholesterol Management:

- Elevated cholesterol levels were confirmed as a significant risk factor through hypothesis testing and regression analysis. Programs to monitor and control cholesterol can reduce heart attack risks.

### 4. Early Screening for At-Risk Patients:

- Patients with high cholesterol, ST depression, and chest pain should undergo regular screenings.
- Implement machine learning tools (Gradient Boosting) in hospitals to screen and identify high-risk patients efficiently.

## Conclusion

In conclusion, we recommend KNN as the preferred model for clinical workflows due to its strong performance. For faster interpretations, Decision Tree can be an alternative. Key features like age, chest pain type, and cholesterol must be prioritized in assessments, and cholesterol management should remain a focus to reduce cardiovascular risks.

---



## 6. Interactive Dashboard

In this project, we utilized Streamlit to build an interactive and user-friendly dashboard for Heart Attack Analysis and Prediction. Streamlit was chosen due to its simplicity, rapid development capabilities, and seamless integration with Python-based workflows.

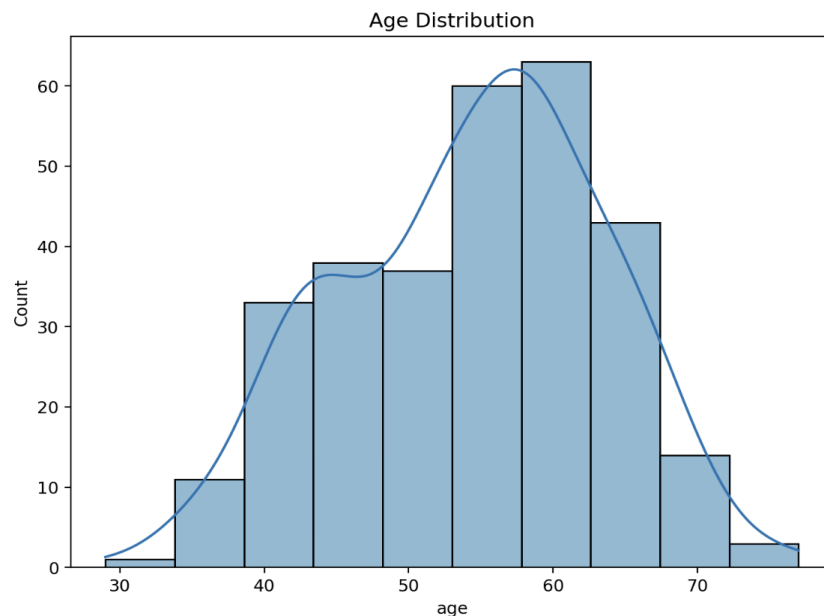
The dashboard serves two primary purposes:

### 1. Exploratory Data Analysis (EDA):

Users can interactively explore the dataset through various visualizations such as histograms, scatter plots, and correlation heatmaps. These visualizations help identify critical features, such as cholesterol levels, chest pain types, and ST depression, that influence heart attack outcomes.

- ☒ Age Distribution
- ☐ Cholesterol vs Age
- ☐ Correlation Heatmap
- ☐ Pairplot of Selected Features
- ☐ Average Blood Pressure by Age Group

#### Age Distribution



Age Distribution: The dataset includes individuals across a broad age range, with a concentration in middle-aged and senior groups (40–70 years). This suggests heart disease risks are evaluated for a high-risk demographic.

### 2. Machine Learning-Based Prediction:

The dashboard integrates a trained **K-Nearest Neighbor model** to predict the likelihood of a heart attack based on user-provided inputs (e.g., age, chest pain type, cholesterol

levels, and ST depression). Users can enter clinical parameters through an intuitive sidebar, and the dashboard dynamically outputs the prediction.

### Heart Attack Prediction

Provide the following details to predict the likelihood of a heart attack:

Age

50

-

+

Sex

Female

▼

Chest Pain Type (cp)

2

▼

Resting Blood Pressure (trtbps)

120

-

+

Cholesterol (chol)

200

-

+

Max Heart Rate Achieved (thalachh)

150

-

+

ST Depression (oldpeak)

1.00

-

+

## Prediction Result

The model predicts a HIGH likelihood of a heart attack.

By combining advanced statistical analysis, interactive visualizations, and machine learning, this dashboard provides a comprehensive platform for analyzing heart attack risk factors and predicting outcomes. It is designed to be accessible for both healthcare practitioners and researchers, offering actionable insights to aid in cardiovascular risk assessment.

## 7. References:

SQLite: <https://www.sqlite.org/docs.html>

Pandas: [https://pandas.pydata.org/docs/user\\_guide/missing\\_data.html](https://pandas.pydata.org/docs/user_guide/missing_data.html)

Scikit-learn: <https://scikit-learn.org/stable/index.html>

Streamlit: <https://docs.streamlit.io/>