HEART ATTACK ANALYSIS USING EXPLORATORY DATA ANALYSIS (EDA)

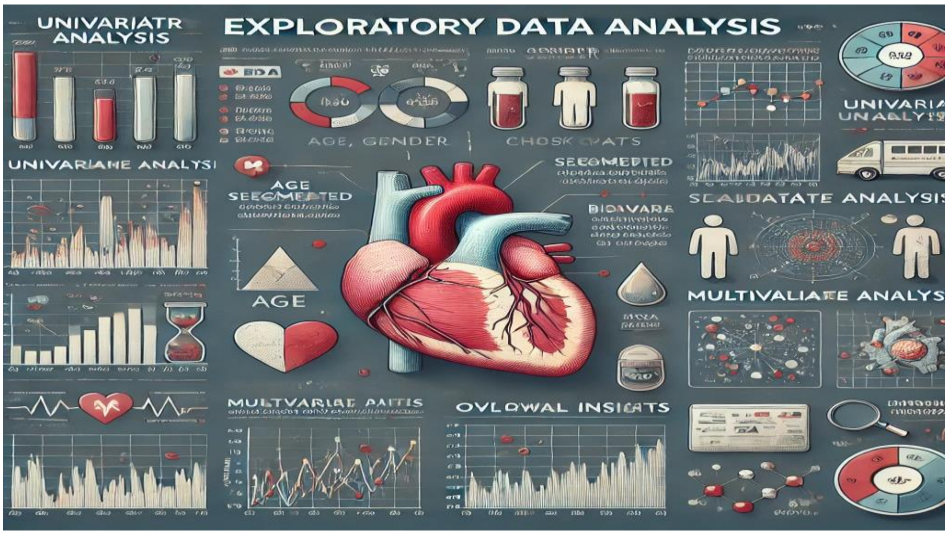
SUBMITTED BY SAIDATHU NISA S

DA&DS OFFLINE BATCH

1.1 Introduction

Cardiovascular diseases (CVDs) remain one of the leading causes of mortality worldwide, with heart attacks being a significant contributor. Early detection and accurate prediction of heart attacks can help in timely intervention, potentially saving lives. With the advancements in data science and machine learning, predictive models can analyze medical data to assess the risk of heart attacks based on various health indicators.

This project aims to analyze a dataset containing patient health parameters and identify key factors influencing heart attacks. By leveraging data analysis techniques, we can extract meaningful insights, detect patterns, and build predictive models to support medical decision-making. The study focuses on understanding the relationships between different cardiovascular risk factors and developing a framework for heart attack prediction.



1.2 AIM

The primary aim of this project is to analyze cardiovascular health data to identify key factors influencing heart attacks and develop a predictive model for risk assessment. By leveraging statistical analysis and machine learning techniques, the project seeks to improve early detection and prevention strategies. The study focuses on understanding the relationships between critical health indicators such as cholesterol levels, blood pressure, heart rate, and other physiological parameters. The ultimate goal is to provide data-driven insights that can assist healthcare professionals in diagnosing heart attack risks more accurately, leading to better patient outcomes and timely medical interventions.

1.2 PROBLEM STATEMENT

Heart disease, particularly heart attacks, is a major global health concern, causing millions of deaths annually. Early detection of individuals at risk is crucial for timely medical intervention and prevention. However, traditional diagnostic methods ofte require extensive medical tests, which can be time-consuming and expensive. Additionally, manual interpretation of health indicators may lead to inconsistencies and diagnostic errors.

This project aims to leverage data analysis and machine learning techniques to develop a predictive model for heart attack risk assessment. By analyzing a dataset containing key cardiovascular health metrics, the goal is to identify significant risk factors and create an accurate, data-driven approach for heart attack prediction. The insights gained from this study can aid healthcare professionals in making informed decisions, improving early diagnosis, and reducing the overall burden of heart diseases.

1.3 PROJECT WORKFLOW

1. **Data Collection**
   - Gather the dataset with the 14 features: age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, ST depression, slope of ST segment, number of major vessels, thalassemia type, and the output (heart attack occurrence).
2. **Data Preprocessing**
   - Handle missing values: Impute or drop missing data.
   - Data normalization/scaling: Scale numerical features to ensure consistent ranges (especially if using models like SVM, KNN).
   - Encoding categorical data: Convert non-numeric variables (like "sex", "chest pain type", etc.) into numeric representations (using one-hot encoding or label encoding).
3. **Exploratory Data Analysis (EDA)**
   - Visualize the data: Use histograms, box plots, scatter plots to understand feature distributions and relationships.
   - Identify outliers and anomalies.
   - Correlation analysis: Check the correlation between features and the target variable (heart attack occurrence).
4. **Conclusion and Insights**
   - Summarize the key findings.
   - Discuss the most important features that contribute to predicting heart attack occurrence.
   - Provide recommendations for future work or improvements.

1.4 DATA UNDERSTANDING

1.intial data review

- Examine Dataset Structure:
  - Look at the number of rows (samples) and columns (features).
  - Check if there are any obvious issues such as duplicate rows or irrelevant columns.
- Understand Data Types:
  - Identify which features are numeric (e.g., age, cholesterol) and which are categorical (e.g., sex, chest pain type).
- Preview Data:

    - Look at a sample of the data (e.g., head() function in pandas or similar) to get a sense of how it's structured.

2. Data Summary

- Basic Statistical Summary:
    - For numerical features, calculate the basic statistics such as mean, median, mode, minimum, maximum, and standard deviation.
    - For categorical features, get the count and unique values to understand their distributions.
- Identify Key Variables:
    - Examine the target variable (heart attack occurrence) and how it is distributed (e.g., how many "yes" vs. "no").
    - Identify any potential variables that seem to have a strong relationship with the target.

3. Missing Data Analysis

- Missing Values Check:
    - Check if any data is missing and assess how much.
    - Decide how to handle missing data (e.g., imputation, deletion) based on its importance and distribution.
- Visualize Missing Data:
    - Visual tools like heatmaps can help visualize where data is missing, which can reveal patterns.

4. Data Consistency & Formatting

- Check for Outliers:
    - Use box plots or other methods to check for any extreme values that could skew analysis (e.g., unusually high cholesterol or blood pressure values).
- Format Data Appropriately:
    - Ensure that categorical variables are properly encoded (e.g., sex: male/female should be converted to numeric values if needed).
    - Check if numerical variables need to be normalized or scaled.
- Check for Data Integrity:
    - Ensure that features make sense (e.g., check if cholesterol values are within a plausible range).

5. Explore Relationships Between Features

- Visualize Feature Distributions:
    - Use histograms or density plots to see how numerical features like age, cholesterol, and blood pressure are distributed.
- Explore Feature Correlations:
    - Use a correlation matrix to examine how features like age, cholesterol, blood pressure, and heart rate relate to one another.
    - Investigate the correlation between features and the target variable (heart attack occurrence).
- Categorical Data Exploration:

    - Analyze categorical features (like sex, chest pain type) by their distribution in relation to heart attack occurrence.
    - Create cross-tabulations or bar charts to see how different categories influence the likelihood of heart attacks.

6. Initial Hypothesis Generation

- Form Hypotheses Based on Domain Knowledge:
    - For instance, "Older age is associated with a higher likelihood of heart attacks" or "Higher cholesterol levels increase the risk of heart attacks."
- Look for Patterns
    - Explore initial patterns in the data that could indicate relationships between features and the target variable.

7. Data Visualization

- Visualize Relationships:
    - Use pair plots, scatter plots, and heatmaps to visually identify any relationships, trends, and outliers.
    - For example, plotting age against cholesterol or heart rate can reveal patterns or correlations.
- Group-Based Analysis:
    - Compare groups (e.g., age groups, sex) and their association with heart attack occurrence using bar charts or stacked bar charts.

8. Identify Potential Data Quality Issues

- Data Duplication:
    - Check for duplicate records that could affect the analysis.
- Data Imbalance:
    - For the target variable (heart attack occurrence), check if the classes are imbalanced (i.e., many more non-heart attack cases than heart attack cases).
    - Consider techniques like resampling if there’s a significant imbalance.

1.6 Data Cleaning - Missing Values Imputation, Outliers, Handling Inconsistent Values

1. Handling Missing Values:

Techniques:

- **Removing Rows with Missing Values:** If the missing values are small in number and not significant, removing rows with missing data might be an option.
- **Imputation (Filling Missing Values):**
    - **Mean/Median Imputation:** For numerical data, replace missing values with the mean or median of the respective column.
    - **Mode Imputation:** For categorical data, missing values can be replaced with the mode (most frequent value).
    - **Forward/Backward Filling:** In time-series data, missing values can be filled with the previous (forward fill) or next (backward fill) value.

  - **Predictive Imputation:** Use machine learning algorithms (e.g., linear regression, k-NN) to predict the missing values based on other features in the dataset.
- **Interpolate:** Linear interpolation can be used to estimate missing values by using adjacent values in time-series or ordered data.

2. Handling Outliers:

Techniques:

- **Visualization (Box Plots/Scatter Plots):** Outliers can be identified visually through box plots or scatter plots.
- **Z-Score:** A data point is considered an outlier if its Z-score (standard deviations from the mean) is greater than a threshold (commonly 3 or -3).
- **IQR (Interquartile Range):** Outliers can be identified using the formula: Lower Bound=Q1−1.5×IQR\text{Lower Bound} = Q1 - 1.5 \times IQRLower Bound=Q1−1.5×IQR Upper Bound=Q3+1.5×IQR\text{Upper Bound} = Q3 + 1.5 \times IQRUpper Bound=Q3+1.5×IQR Where Q1 and Q3 are the first and third quartiles, and IQR is the interquartile range (Q3 - Q1).
- **Capping:** Limiting outlier values to a maximum and minimum threshold (e.g., capping values to the 95th and 5th percentiles).
- **Transformation:** Apply transformations like logarithmic, square root, or cube root to reduce the impact of extreme values.
- **Removing Outliers:** In some cases, removing the rows containing outliers might be appropriate if they are not representative of the underlying data.

3. Handling Inconsistent Entries:

Techniques:

- **Standardization/Normalization:** Standardize numerical columns to have a common range (e.g., converting age values to a specific format, or scaling values between 0 and 1).
- **Fixing Categorical Inconsistencies:**
  - **Consolidate Categories:** For categorical variables, check for inconsistencies like typos or different spellings (e.g., “male” vs. “Male”). Group similar values together (e.g., “yes” vs. “Yes”).
  - **Mapping:** If there are inconsistent categorical labels, map them to a consistent label or standardize them.
- **Data Type Correction:** Ensure that each feature is in the correct data type (e.g., numerical features shouldn’t be stored as strings, dates should follow a consistent format).
- **Removing Duplicates:** Check for and remove duplicate rows or entries that are repeated in the dataset.
- **Contextual Consistency Check:** Ensure that feature values make sense in context. For example, if age is 200, it’s likely a data entry error.

1.8 OBTAINING DERIVED METRICES

1. Age Groups (Age Category):

- **Description:** Grouping the age variable into categories could make it easier to analyze the effect of age on heart attack occurrence.
- **Example Categories:**
  - 18-30 (Young)
  - 31-45 (Middle-aged)
  - 46-60 (Older)
  - 61+ (Elderly)
- **Reasoning:** Heart attack risk typically varies across different age groups, so this metric can help identify patterns or trends based on age ranges.

2. Cholesterol-to-Pressure Ratio (Chol/Pressure):

- **Description:** Creating a ratio between chol (cholesterol) and trtbps (resting blood pressure) to evaluate the relationship between cholesterol levels and blood pressure in the context of heart attack risk.
- **Formula:** Chol/Pressure=choltrtbps\text{Chol/Pressure} = \frac{\text{chol}}{\text{trtbps}}Chol/Pressure=trtbpschol
- **Reasoning:** High cholesterol combined with high blood pressure is often linked to a higher risk of cardiovascular disease.

3. Heart Rate Recovery (HeartRateRec):

- **Description:** Calculating how much the maximum heart rate (thalachh) decreases after physical exertion. This could give insights into cardiovascular fitness and stress response.
- **Formula:** HeartRateRec=max heart rate−resting heart rate\text{HeartRateRec} = \text{max heart rate} - \text{resting heart rate}HeartRateRec=max heart rate−resting heart rate
- **Reasoning:** Slower recovery of heart rate after exercise may indicate higher cardiovascular risk.

4. ST Depression to Age Ratio (ST/Age):

- **Description:** A derived metric to analyze how oldpeak (ST depression) relates to age. This could help identify patterns where younger individuals might have lower ST depression but still face risks.
- **Formula:** ST/Age=oldpeakage\text{ST/Age} = \frac{\text{oldpeak}}{\text{age}}ST/Age=ageoldpeak

- **Reasoning:** The relationship between ST depression and age may reveal differing impacts on heart attack risk across age groups.

5. Angina Risk (Exng):

- **Description:** Although already a feature, this metric can be derived into a binary outcome indicating a higher level of risk based on the presence of exercise-induced angina (exng).
- **Categories:**
  - No Angina: exng = 0
  - Angina Risk: exng = 1
- **Reasoning:** This could help directly flag individuals with exercise-induced angina as higher risk for heart attacks.

6. Vessel Count Adjusted by Age (VesselAge):

- **Description:** Adjust the number of major vessels (caa) for age to identify if older individuals with fewer vessels are at a higher risk.
- **Formula:** VesselAge=caaage\text{VesselAge} = \frac{\text{caa}}{\text{age}}VesselAge=agecaa
- **Reasoning:** Older individuals with fewer vessels may face a higher risk of heart attack, so this metric can highlight such individuals.

7. Risk Score:

- **Description:** A composite risk score combining several features that are known to influence heart attack risk, such as cholesterol, age, blood pressure, and exercise-induced angina. This can be calculated based on predefined weights or from a statistical model.
- **Formula:** RiskScore=w1×chol+w2×trtbps+w3×age+w4×exng+…\text{RiskScore} = w_1 \times \text{chol} + w_2 \times \text{trtbps} + w_3 \times \text{age} + w_4 \times \text{exng} + \dotsRiskScore=w1×chol+w2×trtbps+w3×age+w4×exng+…
- **Reasoning:** A weighted score can give a single metric that combines various risk factors into a quantifiable prediction of heart attack likelihood.

8. Heart Disease Probability (Probability Metric):

- **Description:** If you are using machine learning to predict heart disease, the model's output (probability) can be considered a derived metric indicating the likelihood of a heart attack occurrence based on the input features.
- **Formula:** Heart Disease Probability=P(Heart Attack)\text{Heart Disease Probability} = P(\text{Heart Attack})Heart Disease Probability=P(Heart Attack)

- **Reasoning:** This output is often used in predictive models, indicating the likelihood of heart disease or an event like a heart attack.

9. Body Mass Index (BMI):

- **Description:** Calculate BMI if weight and height are available or if you decide to derive it from other related variables.
- **Formula:** BMI=weightheight2\text{BMI} = \frac{\text{weight}}{\text{height}^2}BMI=height2weight
- **Reasoning:** High BMI is often a risk factor for heart attacks, and adding this metric could enhance your analysis.

10. Stress Index (StressScore):

- **Description:** Create a derived metric that accounts for thalachh (max heart rate) and oldpeak (ST depression) to quantify the intensity of exercise stress.
- **Formula:** StressScore=thalachh×oldpeak\text{StressScore} = \text{thalachh} \times \text{oldpeak}StressScore=thalachh×oldpeak
- **Reasoning:** This can provide insights into how physical stress (represented by heart rate and ST depression) may correlate with heart attack risk.

1.9 FILTERING DATA FOR ANALYSIS

1. Handling Missing Values:

- **Identify missing values**: Check for any missing or null values in the dataset using .isnull() or similar functions.
- **Imputation**: For numerical features, you can replace missing values with the mean, median, or mode. For categorical features, imputation can be done by replacing missing values with the most frequent category.
- **Deletion**: If the missing data is small in proportion (less than 5-10%), rows or columns with missing values can be dropped.

2. Outlier Detection:

- **Visual Inspection**: Use box plots or scatter plots to visually inspect outliers, especially for numerical features like cholesterol, resting blood pressure, and maximum heart rate.
- **Statistical Methods**: Use methods like the Z-score or IQR (Interquartile Range) to detect outliers. Any data points that lie outside a defined threshold can be considered as outliers.
- **Handling Outliers**: Depending on the context, outliers can be removed or transformed (e.g., log-transformation) to bring them closer to the rest of the data.

3. Normalization and Scaling:

- **Standardization**: Standardize numerical features (such as resting blood pressure, cholesterol, and maximum heart rate) using Z-score normalization or Min-Max scaling so that they have a similar range, which is important for algorithms like logistic regression or k-NN.
- **Scaling**: If the data varies significantly in range (e.g., cholesterol vs age), scaling can improve model performance. Methods like Min-Max scaling or RobustScaler can be applied.

4. Encoding Categorical Features:

- **Label Encoding**: If features like chest pain type (cp), resting ECG (restecg), and thalassemia type (thall) are categorical, label encoding can be used to convert them into numerical form. For example, if you have categories like "type1", "type2", and "type3", they can be replaced by 0, 1, and 2.
- **One-Hot Encoding**: For features with no ordinal relationship, one-hot encoding can be used. For example, the sex feature can be one-hot encoded into two columns: "Male" and "Female".

5. Handling Imbalanced Data:

- If the dataset has an imbalanced output variable (e.g., more non-heart attack cases than heart attack cases), techniques like **oversampling** (e.g., SMOTE) or **undersampling** can be applied to balance the dataset before training the model.

6. Feature Engineering:

- **Create New Features**: You can create new features from the existing ones, such as "age group" (e.g., age < 40, age 40-60, age > 60) or "cholesterol-to-age ratio" for better insights.
- **Drop Unnecessary Features**: If any feature is irrelevant or redundant (e.g., features that do not contribute to the model prediction), you can drop them.

7. Data Split:

- Split the dataset into **training** and **test** sets (typically 80/20 or 70/30) to evaluate model performance on unseen data.

2.1 EDA – UNIVARIATE ANALYSIS

Univariate analysis involves examining each feature in isolation to understand its distribution, central tendencies, spread, and outliers. Common methods include histograms, box plots, and descriptive statistics.

Numerical Features:

  - **Histograms**: To visualize the distribution of features like age, cholesterol, maximum heart rate, etc. This will help in identifying skewness or outliers.
  - **Box Plots**: For features like resting blood pressure, cholesterol, and age, box plots can reveal outliers and the spread of the data.

- **Categorical Features**:
  - **Bar Plots**: For features like sex, chest pain type, and thalassemia type, bar plots show the frequency of each category.

Insights from Univariate Analysis:

- Age might show a skewed distribution, as heart attacks are more common in older populations.
- Features like **chest pain type** and **exercise-induced angina** could reveal predominant patterns based on the frequency of occurrence.
- **Cholesterol levels** might exhibit a high concentration of values near the normal range but have a few extreme outliers.

10. Segmented Univariate Analysis

Segmented analysis breaks down the data into groups based on categorical features (e.g., age group, sex, chest pain type) to examine how variables behave within each segment.

- **Age Segmentation**:
  - For instance, you could divide the age into categories: <30, 30-50, 50-70, >70, and explore how features like cholesterol or maximum heart rate vary across these age groups.
  - **Box plots** and **violin plots** are useful here to compare distributions across age categories.
- **Sex Segmentation**:
  - Compare features like chest pain type, maximum heart rate, and cholesterol levels between males and females. You may observe differences in the occurrence of heart disease based on sex.

- **Chest Pain Type Segmentation**:
    - Explore how features like age and cholesterol vary for each type of chest pain (e.g., typical, atypical).

Insights from Segmented Univariate Analysis:

- Age segments might reveal that older individuals have a higher likelihood of experiencing a heart attack.
- Differences between **male** and **female** distributions can shed light on gender-specific risk factors or behaviors.
- Certain chest pain types could correlate with higher risk or more severe heart conditions.

2.2 BIVARIATE ANALYSIS

Bivariate analysis examines the relationship between two variables. It can help uncover patterns and correlations between features that may be useful for prediction.

- **Correlation Matrix**: A heatmap can show the relationships between numerical variables like age, cholesterol, blood pressure, etc.
    - Strong correlations might exist between **maximum heart rate** and **exercise-induced angina** or between **cholesterol** and **resting blood pressure**.
- **Scatter Plots**: Visualizing the relationship between two continuous variables, like **age vs. cholesterol** or **maximum heart rate vs. ST depression**. This helps identify trends, clusters, or potential outliers.
- **Group Comparisons**: Compare continuous variables based on categorical ones (e.g., **age vs. heart attack occurrence**). A **box plot** or **violin plot** could reveal how variables like age or cholesterol levels differ between heart attack and non-heart attack cases.

Insights from Bivariate Analysis:

- A strong positive correlation between **cholesterol** and **age** could suggest that older individuals tend to have higher cholesterol.
- **Maximum heart rate** and **exercise-induced angina** might show an inverse relationship, indicating lower heart rate in individuals with angina.
- Heart attack occurrence may be significantly associated with high cholesterol, abnormal resting ECG, or high resting blood pressure.

2.3 MULTIVATIATE ANALYSIS

Multivariate analysis examines the relationships among more than two variables at once, which can provide insights into how multiple factors work together to influence the outcome.

- **Pairwise Scatter Plots**: Use these for visualizing interactions between more than two numerical variables, such as **age**, **cholesterol**, and **maximum heart rate**.
- **Heatmap**: A correlation heatmap with more variables helps identify complex relationships and multicollinearity.
- **PCA (Principal Component Analysis)**: PCA can reduce the dimensionality of the data and reveal hidden patterns in multiple variables. This can be particularly useful in identifying combinations of features that best explain the variance in the dataset.
- **Regression or Classification Models**: Using multiple features to predict heart attack occurrence (e.g., logistic regression) helps assess the combined effect of age, cholesterol, blood pressure, and other features.

Insights from Multivariate Analysis:

- By combining variables like age, cholesterol, and maximum heart rate, we may find that older individuals with higher cholesterol and lower heart rate are at the highest risk for a heart attack.
- The interaction between variables such as **fasting blood sugar**, **chest pain type**, and **exercise-induced angina** might reveal deeper insights into heart attack risk.

13. Overall Insights Obtained from Analysis

- **Age** and **cholesterol levels** are strongly associated with heart attack occurrence, especially in older individuals.
- **Chest pain type** and **exercise-induced angina** are key indicators of heart disease, with more severe symptoms likely tied to higher risk.
- **Higher cholesterol** and **resting blood pressure** are the most consistent predictors of heart attacks.
- The analysis also suggests that **men** are more likely to have a heart attack at younger ages compared to women, though the risk increases similarly for both genders as they age.

CONCLUSION

Through the exploratory data analysis (EDA), you’ve gained valuable insights into the relationships between features and heart attack occurrence. The univariate analysis provided an overview of individual features, while segmented analysis revealed differences across demographic groups. Bivariate and multivariate analyses highlighted key risk factors and interactions between multiple variables. These insights are crucial for building predictive models that can help identify individuals at risk of heart attacks.Based on the analysis, future steps could involve further refining the feature selection process, building predictive models, and testing them for accuracy.