Comprehensive Data Cleaning and Exploratory Analysis of:

Heart Attack Dataset

Submitted by: SMRITI PRIYA

Enrollment no: 23116091

Branch: ECE (EC4)

---

1. Overview

Understanding heart attack risks requires a detailed analysis of contributing factors. This study focuses on refining raw data, identifying trends, and ensuring high-quality insights by employing systematic data cleaning and in-depth exploratory data analysis (EDA).

The workflow includes:

- Pre-processing the dataset
- Identifying missing data patterns and resolving them
- Spotting outliers and inconsistencies
- Conducting thorough statistical and graphical analysis
- Drawing meaningful insights to support predictive modeling

Heart disease remains one of the leading causes of mortality worldwide. By analyzing patient data, we can gain insights into critical factors influencing heart attack risks and contribute to better healthcare decision-making.

---

2. Data Pre-Processing

2.1 Data Examination

- The dataset is loaded and reviewed using Python's Pandas library.
- Column names, data types, and summary statistics are checked.
- The first few records are inspected to get an overview of structure and patterns.
- Visual inspection using .head() and .describe() ensures that variables align with expected values.
- Understanding variable distributions helps in determining data-cleaning strategies.

2.2 Managing Missing Information

- **Numerical Attributes:** Missing values are handled using mean/median imputation based on data distribution.
- **Categorical Attributes:** Mode imputation or "unknown" category assignment is used.
- A missing value heatmap is generated to visualize the distribution of absent data.
- If missing values exceed 20% in a feature, consideration is given to removing the column or imputing based on domain knowledge.

2.3 Removing Redundant Data

- Identical rows that do not provide additional information are eliminated.
- Unique patient entries are ensured using .duplicated().sum() and .drop_duplicates().
- Duplicate patient IDs, if any, are cross-checked to ensure data integrity.

2.4 Addressing Anomalies

- **Outlier Detection:**
  - The interquartile range (IQR) method flags extreme values.
  - Box plots help visualize inconsistencies in heart rate, cholesterol, and blood pressure.
  - Z-score detection is used for numerical features with normal distributions.
- **Handling Outliers:**
  - Capping extreme values using percentile-based thresholds.
  - Log transformations for highly skewed variables to normalize data.
  - Winsorization is applied where necessary to keep data within reasonable bounds.

2.5 Standardizing Variables

- Text inconsistencies in categorical columns are corrected.
- Case uniformity is maintained for labels.
- Units of measurement are aligned for consistency.
- Encoding categorical variables using one-hot encoding or label encoding for machine learning compatibility.

---

3. Exploratory Data Analysis (EDA)

3.1 Individual Feature Investigation

- **Numerical Features:**
  - Summary statistics such as mean, variance, and percentiles are computed.
  - Histograms illustrate the spread and skewness of data.
  - Kernel density plots estimate probability distributions.
  - Outlier detection through box plots helps assess the need for transformation.
- **Categorical Features:**
  - Frequency distribution of values is plotted using bar charts.
  - Pie charts depict proportion comparisons.
  - Cross-tabulations identify relationships between categorical variables.

3.2 Two-Variable Relationships

- **Correlation Matrix:**
  - Pearson's coefficient measures strength between numerical variables.
  - Heatmaps highlight highly correlated factors.
  - Spearman correlation is used for non-linear relationships.
- **Comparative Analysis:**
  - Box plots explore how numerical features vary across categories.
  - Scatter plots showcase associations between key variables.
  - Swarm plots help visualize distributions in smaller datasets.
- **ANOVA (Analysis of Variance):**
  - Determines whether there are statistically significant differences between groups.

3.3 Multi-Feature Interactions

- **Pairwise Plot Analysis:**
  - Displays connections between age, cholesterol, and heart attack risk.
- **Multicollinearity Check:**
  - Variance Inflation Factor (VIF) identifies redundant predictors.
  - Variables with high VIF are reconsidered or combined to improve model efficiency.
- **Clustering Patterns:**
  - Principal Component Analysis (PCA) helps in dimensionality reduction.
  - K-means clustering groups similar patient profiles.
  - Hierarchical clustering examines relationships among variables.

---

4. Major Insights

4.1 Data Cleaning Takeaways

- Addressing missing values prevented biased interpretations and ensured robust analysis.
- Removing duplicates ensured uniqueness in patient data, leading to more accurate conclusions.
- Outlier handling refined the dataset for better predictive modeling.

4.2 Key Observations from Analysis

- **Age and heart attack risk:** A positive correlation exists between age and heart attack probability.
- **Cholesterol and high blood pressure:** Strongly influence heart attack occurrences, confirming existing medical studies.
- **Physical activity:** An inverse correlation is found between exercise levels and heart disease risks.

- **BMI impact:** Higher body mass index values are linked to increased risk, but other lifestyle factors contribute.
- **Gender-based trends:** Males show a higher prevalence of heart attack risks compared to females in the dataset.
- **Diabetes and smoking habits:** Both contribute significantly to elevated heart disease risks.
- **Seasonal variations:** Certain months record higher heart attack cases, indicating possible environmental or lifestyle triggers.
- **Medical intervention history:** Patients with previous medical treatments showed a lower risk factor in predictive modeling.
- **Machine learning considerations:** Feature selection can improve model accuracy for predicting heart attack risks.

---

5. Conclusion & Strategic Direction

- **Summary:**
    - Data refinement significantly improved dataset reliability.
    - Exploratory analysis highlighted critical risk indicators.
    - Findings align with medical research on heart disease risk factors.
- **Next Steps:**
    - Feature engineering can further enhance predictive analysis.
    - Inclusion of lifestyle-related attributes (diet, stress levels) would provide a broader risk assessment.
    - More sophisticated modeling techniques, such as ensemble learning, can improve predictions.
    - Medical professionals should validate findings before integrating them into clinical practice.
    - Collaboration with healthcare providers to collect additional patient data could improve future research.

This document serves as a foundation for advanced modeling and targeted research in heart disease prediction, offering a structured approach to understanding risk factors and improving patient outcomes.