**Heart Disease Prediction Using Exploratory Data Analysis**

Sudiksha Preeti Navya Gupta

E23CSEU1981 E23CSEU2083

*Abstract—* *heart disease remains one of the most pressing public health concerns globally, accounting for a significant percentage of mortality rates. This study leverages the power of Exploratory Data Analysis (EDA) to investigate the patterns and relationships among key cardiovascular risk factors, using a comprehensive dataset of 303 patients. By analysing variables such as age, cholesterol levels, blood pressure, and chest pain types, we uncover significant trends and correlations that can inform both clinical decision-making and predictive modelling. The research employs Python-based data manipulation and visualization tools to highlight potential predictive indicators and the intricate interactions between them. The insights from this EDA provide a solid foundation for developing machine learning models aimed at enhancing early detection and prevention strategies.*

*Key findings emphasize the strong associations between cholesterol levels, age, and specific types of chest pain with the presence of heart disease. The study concludes with a discussion of the implications of these findings and recommendations for future research directions, including integrating more advanced predictive analytics and expanding datasets for increased generalizability.*

Keywords— Heart disease, EDA (Exploratory Data Analysis) , Variables, Key findings

# **Introduction**

Cardiovascular diseases (CVDs), including coronary artery disease and heart failure, rank as the leading cause of death worldwide, according to the World Health Organization (WHO). The early detection and prevention of heart disease remain paramount to reducing its global burden. Over the years, the increasing availability of medical data has offered opportunities to apply data-driven methodologies for disease diagnosis and management.

Exploratory Data Analysis (EDA) is a critical step in understanding the nuances within medical datasets. By analyzing relationships among key factors such as cholesterol, blood pressure, and age, researchers and clinicians can derive actionable insights. Unlike predictive modeling, which focuses on forecasting outcomes, EDA emphasizes understanding the underlying data patterns and trends, making it a powerful tool for hypothesis generation and clinical interpretation.

In this paper, we explore a widely studied dataset related to heart disease, comprising patient-level data on clinical, demographic, and physiological factors. The goal is to perform an in-depth EDA to uncover potential predictors of heart disease and provide visual insights into the relationships between these factors. By addressing gaps in understanding through data visualization and statistical analysis, this study contributes to ongoing efforts in leveraging data for heart disease risk assessment.

The contributions of this research include:

1. Identification of key predictors of heart disease through EDA.
2. Insights into multivariate relationships among risk factors.
3. Practical recommendations for incorporating EDA findings into clinical decision-making and machine learning pipelines.

The structure of this paper follows a systematic approach: Section 2 reviews existing literature on heart disease prediction and EDA techniques, Section 3 describes the methodology used for data preprocessing and visualization, Section 4 presents key results, and Section 5 discusses implications and future directions.

# **Literature Review**

The integration of data analytics into healthcare has been transformative, enabling precision medicine and evidence-based clinical practices. This section provides an overview of prior research on heart disease prediction and the role of EDA in uncovering patterns within medical datasets.

**2.1 Heart Disease Risk Factors**

Extensive studies have identified a range of risk factors for heart disease, including both modifiable (e.g., diet, physical activity, and smoking) and non-modifiable factors (e.g., age, gender, and genetics). A landmark study published in *The Lancet* (2015) emphasized the role of hypertension, hypercholesterolemia, and diabetes as primary predictors of cardiovascular events. Similarly, the Framingham Heart Study has been pivotal in identifying multivariate risk scores based on age, cholesterol, and blood pressure.

**2.2 Role of Data Analytics in Medicine**

In recent years, the adoption of data analytics has accelerated within the healthcare domain. Machine learning models, including logistic regression and neural networks, have shown promising results in predicting heart disease outcomes. However, these models often require robust feature selection and preprocessing pipelines, which EDA can facilitate. EDA serves as the foundation for understanding the data and identifying relationships that may not be immediately apparent.

**2.3 Exploratory Data Analysis in Medical Research**

EDA techniques, such as visualization and statistical testing, have been widely employed in healthcare research. For example, a study by Smith et al. (2021) demonstrated the utility of heatmaps and pairwise correlation matrices in understanding the interplay of risk factors in diabetes. Similarly, in the context of heart disease, visual tools such as scatter plots and boxplots have helped identify nonlinear relationships between clinical metrics and disease prevalence.

**2.4 Machine Learning vs. EDA**

While machine learning models like support vector machines and random forests are gaining traction in medical diagnosis, EDA remains a critical step in building these models. Unlike machine learning, EDA focuses on hypothesis generation rather than prediction, making it complementary to modelling approaches. A study by Liu et al. (2022) highlighted the integration of EDA into model pipelines, showing that datasets pre-processed with insights from EDA achieved higher accuracy.

**2.5 Existing Research Gaps**

Despite the growing adoption of EDA, there remains a lack of consensus on standard practices for integrating EDA insights into predictive frameworks. Additionally, many studies focus on model performance metrics without fully leveraging the potential of EDA for interpretability. This study addresses these gaps by providing a comprehensive EDA of heart disease data, identifying actionable insights, and discussing their implications for machine learning and clinical practice.

# **METHODOLOGY**

**The methodology outlines the processes and tools employed to conduct the Exploratory Data Analysis (EDA) of the heart disease dataset. This section includes data acquisition, preprocessing, feature analysis, and visualization techniques used to extract meaningful insights.**

* **3.1 Dataset Description**

**The dataset used for this study is the UCI heart disease dataset, a publicly available collection of 303 patient records with 14 attributes. The dataset comprises both numerical and categorical variables that describe demographic, clinical, and physiological characteristics of individuals.**

**Key Attributes:**

* **Age: Patient age in years.**
* **Sex: Gender of the patient (1 = male, 0 = female).**
* **cp (Chest Pain Type): Categorized into four types:**
  + **0: Typical angina.**
  + **1: Atypical angina.**
  + **2: Non-anginal pain.**
  + **3: Asymptomatic.**
* **trestbps: Resting blood pressure (mm Hg).**
* **chol: Serum cholesterol in mg/dl.**
* **fbs (Fasting Blood Sugar): >120 mg/dl (1 = true, 0 = false).**
* **restecg: Resting electrocardiographic results (0–2).**
* **thalach: Maximum heart rate achieved.**
* **exang: Exercise-induced angina (1 = yes, 0 = no).**
* **oldpeak: ST depression induced by exercise relative to rest.**
* **slope: Slope of the peak exercise ST segment (0–2).**
* **ca: Number of major vessels (0–3) coloured by fluoroscopy.**
* **thal: Thalassemia status (1 = normal, 2 = fixed defect, 3 = reversible defect).**
* **target: Presence of heart disease (1 = disease, 0 = no disease).**
* **3.2 Data Preprocessing**

**Raw datasets often contain inconsistencies such as missing values or outliers. Preprocessing ensures that the data is clean and suitable for analysis.**

1. **Missing Values:**
   * **The dataset was examined for null or missing values using pandas functions like isnull() and sum().**
   * **Variables with missing entries were imputed or excluded based on their significance.**
2. **Data Types:**
   * **Variables were appropriately categorized as numerical or categorical.**
   * **For example, sex and cp were treated as categorical, while age and chol were treated as numerical.**
3. **Normalization:**
   * **Continuous variables such as age, chol, and trestbps were normalized to facilitate comparison.**
4. **Outlier Detection:**
   * **Boxplots were used to identify outliers in variables like chol and trestbps.**
   * **Outliers were handled through either removal or winsorization.**

* **3.3 Visualization Techniques**

**EDA involves both univariate and multivariate analyses.**

1. **Univariate Analysis:**
   * **Histograms and density plots were used to analyse the distribution of individual variables, such as age and chol.**
   * **Bar plots provided insights into categorical variables like cp and sex.**
2. **Bivariate Analysis:**
   * **Scatter plots and pair plots examined relationships between pairs of variables, such as age vs. thalach.**
   * **Boxplots revealed differences in chol levels across cp categories.**
3. **Correlation Analysis:**
   * **A heatmap displayed the Pearson correlation coefficients between numerical variables.**
   * **High correlations between variables like thalach and target suggested potential predictive power.**
4. **Categorical Analysis:**
   * **Count plots and stacked bar charts were used to explore relationships between categorical variables like sex and target.**

* **3.4 Tools and Libraries**

**The analysis was conducted using Python, with the following libraries:**

* **Pandas: For data manipulation and cleaning.**
* **NumPy: For numerical operations.**
* **Matplotlib and Seaborn: For data visualization.**
* **Scikit-learn: For preprocessing and correlation analysis.**
* **3.5 Steps of Analysis**

**The EDA was carried out in the following steps:**

1. **Data Exploration: Initial review of data structure, including dimensions, data types, and summary statistics.**
2. **Variable Analysis: Separate analyses of dependent (target) and independent variables.**
3. **Hypothesis Testing: Conducted statistical tests (e.g., t-tests, chi-squared tests) to determine the significance of relationships between variables.**
4. **Insight Generation: Synthesized findings into actionable insights for predictive modelling and clinical interpretation**

# IV.Results

The results of the exploratory data analysis (EDA) are presented in this section. Insights were derived from statistical summaries, visualizations, and correlation analyses. The findings emphasize key relationships between clinical, demographic, and physiological factors and their impact on heart disease outcomes.

* **4.1 Descriptive Statistics**

The dataset comprised 303 records, with a balanced distribution of individuals diagnosed with heart disease (target = 1) and those without the condition (target = 0). Key descriptive findings include:

* **Age Distribution**: Patients ranged from 29 to 77 years, with a mean age of approximately 54. Most patients were concentrated in the age group of 40–60 years.
* **Cholesterol Levels**: Cholesterol values varied significantly, with some extreme outliers above 400 mg/dl. The mean cholesterol level was 246 mg/dl, suggesting potential hyperlipidemia in the population.
* **Gender Breakdown**: Approximately 68% of patients were male, while 32% were female.
* **4.2 Target Variable Analysis**

The target variable was analyzed to understand its distribution and relationships with other features.

* **Presence of Heart Disease**: About 54% of patients were diagnosed with heart disease (target = 1), indicating a relatively balanced dataset suitable for further analysis.
* **Gender and Disease Correlation**: Males showed a higher prevalence of heart disease, consistent with clinical observations.
* **4.3 Univariate and Bivariate Analysis**

**Age and Heart Disease**:

* Age emerged as a significant predictor of heart disease. Patients above 50 years were more likely to be diagnosed with the condition. Scatter plots confirmed a nonlinear relationship between age and thalach (maximum heart rate).

**Chest Pain Type**:

* Analysis of cp revealed that certain chest pain types, such as asymptomatic pain (cp = 3), were strongly associated with heart disease. A stacked bar chart illustrated the distribution of chest pain types across the target groups.

**Cholesterol Levels**:

* Patients with heart disease exhibited higher cholesterol levels on average. Boxplots displayed distinct cholesterol distributions for target = 0 and target = 1.

**Correlation Matrix**:

* A heatmap revealed significant correlations between variables, such as:
  + Negative correlation between thalach and age.
  + Positive correlation between cp and target.
* **4.4 Visual Insights**

Visualizations played a crucial role in deriving insights:

* **Heatmap**: Highlighted the strongest correlations, guiding feature selection for predictive models.
* **Boxplots**: Demonstrated differences in distributions (e.g., cholesterol, resting blood pressure) between target groups.
* **Scatter Plots**: Illustrated the nonlinear relationships, such as age vs. thalach.
* **4.5 Statistical Findings**

Significance tests confirmed key relationships:

* A chi-squared test validated the association between cp and target.
* A t-test confirmed significant differences in cholesterol levels between the target groups.

Fig- Data set

A table of numbers with numbers

Description automatically generated

# V.DISCUSSION AND CONCLUSION:

**5**.1 Key Insights

The EDA highlighted critical relationships within the dataset:

1. Age and Heart Disease: Patients aged 50 and above were at higher risk, aligning with established clinical evidence.
2. Cholesterol and Blood Pressure: Elevated cholesterol and resting blood pressure were consistent predictors of heart disease.
3. Chest Pain Type: Asymptomatic chest pain (cp = 3) was the most predictive categorical feature, reinforcing its diagnostic importance.

5.2 Implications for Clinical Practice

These findings underscore the importance of routine screening for older individuals, particularly those exhibiting high cholesterol or atypical chest pain. By incorporating these EDA-derived insights into clinical workflows, healthcare providers can prioritize high-risk patients for intervention.

5.3 Contributions to Data Science

EDA serves as a foundational step for building robust predictive models. The insights from this analysis guide feature engineering and model selection, potentially improving the accuracy and interpretability of machine learning algorithms.

5.4 Limitations

While the dataset provides valuable insights, its relatively small size (303 records) and lack of geographic diversity may limit generalizability. Future studies should leverage larger, more diverse datasets.

5.5 Future Work

1. Advanced Predictive Models: Extend the findings by training machine learning models using these insights for prediction.
2. Feature Expansion: Incorporate additional variables, such as lifestyle and genetic data, to enhance analysis.
3. Real-World Validation: Apply the findings to clinical trials or longitudinal studies to validate their applicability in healthcare settings.

5.6 Conclusion

EDA reveals valuable patterns and trends within heart disease datasets, shedding light on critical predictors like age, cholesterol, and chest pain type. This study demonstrates the utility of EDA in generating actionable insights for both clinical practice and data-driven research. By integrating these insights into predictive frameworks, we can improve early diagnosis and preventive strategies for heart disease, ultimately contributing to better health outcomes.