

ASSIGNMENT: EXPLORATORY DATA ANALYSIS (EDA) - Level 1

Problem Statement

The objective of this assignment is to perform an in-depth Exploratory Data Analysis on the Heart Disease Prediction dataset, uncovering insights, patterns, and relationships within the data. Through this analysis, students will gain familiarity with various EDA techniques and derive meaningful conclusions from the dataset.

Guidelines

1. Foundational Knowledge

- Understand the Heart Disease Prediction dataset structure and available features.
- Grasp the importance of EDA in extracting insights from data.
- Familiarize yourself with statistical measures, visualization techniques, and data preprocessing.

2. Data Exploration

- Utilize techniques such as summary statistics, histograms, boxplots, and correlation matrices to understand feature distributions, identify outliers, and explore relationships among variables.

3. Basic Questions

- *Investigate basic statistics: patient counts, presence of heart disease, age distribution, etc.*
- *Analyze the presence of heart disease based on gender, age, and other relevant factors.*
- *Identify any missing or null values and devise strategies for handling them.*

4. Intermediate Questions

- *Explore correlations: Does cholesterol level correlate with the presence of heart disease?*
- *How does age relate to heart disease presence?*
- *Apply feature engineering techniques: create new features if relevant (e.g., BMI, risk factors based on medical history).*

5. Advanced Questions

- *Utilize advanced visualization methods like heatmaps, pair plots, or t-SNE to identify complex relationships.*
- *Perform multivariate analysis: How does the presence of heart disease vary concerning multiple variables simultaneously?*
- *Implement predictive modeling (optional): Prepare the data for a predictive model like logistic regression or random forests to predict the presence of heart disease.*

Step-by-Step Approach to EDA

1. Data Loading and Preprocessing

- Import necessary libraries: pandas, matplotlib, seaborn, etc.
- Load the Heart Disease Prediction dataset and handle missing values or inconsistencies.
- Perform data cleaning and preprocessing for analysis.

2. Basic Analysis

- Compute basic statistics, visualize distributions, and generate initial insights.

3. Intermediate Analysis

- Explore correlations, patterns, and relationships among features.
- Create visualizations to support findings and explore data interactions.

4. Advanced Analysis

- Utilize advanced visualization techniques and multivariate analysis to uncover deeper insights.
- Implement predictive modeling (optional) to predict the presence of heart disease.

5. Conclusion and Interpretation

- Summarize key findings and insights derived from the analysis.
- Reflect on the significance of different features in predicting the presence of heart disease.
- Suggest potential further steps or analyses that could be performed on the dataset.

Dataset Source:

- Heart Disease Prediction Dataset

<https://www.kaggle.com/datasets/utkarshx27/heart-disease-diagnosis-dataset/data>