Milestone 1: Data Collection, Exploration, and Preprocessing

Objectives

- Acquire a relevant healthcare dataset.
- Perform exploratory data analysis (EDA) to understand the structure and quality of the data.
- Preprocess the data to prepare it for predictive modeling.

Tasks & Activities

1. Data Collection

- **Source**: A structured dataset containing clinical and demographic information was obtained. The dataset includes:
    - Age, gender, blood pressure, cholesterol levels, glucose, smoking/alcohol status, physical activity, and cardiovascular disease outcome.
- **Format**: The data was loaded into a Jupyter notebook environment as a .csv file for analysis.

2. Data Exploration

- Conducted EDA using Pandas, Matplotlib, and Seaborn.
- Analyzed the distribution of numerical and categorical features.
- Key findings included:
    - Some skewness in age and cholesterol levels.
    - Imbalances in target class (cardio).
    - Potential outliers in height and weight values.
- Visualizations included:
    - Histograms for continuous variables.
    - Boxplots to detect outliers.
    - Heatmaps to explore feature correlations.

3. Data Preprocessing

- Missing values were handled using:
    - Row-wise removal where applicable.
    - Imputation based on median/mode for clinical measurements.
- Feature scaling applied:

    - StandardScaler was used to normalize continuous features.
- Categorical encoding:
    - Binary variables retained as-is.
    - One-hot encoding or label encoding used for other categorical features where needed.

Deliverables

- **Dataset Exploration Report** summarizing data structure, quality, and trends.
- **EDA Notebook** with visualizations such as histograms, boxplots, and a correlation heatmap.
- **Cleaned Dataset** ready for machine learning model training and evaluation.