Clinical Data Analysis Project

Objective

I developed this project to analyze a clinical healthcare dataset and derive actionable insights into patient demographics and health metrics related to stroke occurrence. My goal was to create meaningful visualizations and statistical summaries to support clinical decision-making.

Dataset

- **Source**: Stroke Prediction Dataset, sourced from Kaggle.
- **Description**: The dataset includes patient records with key features:
    - `age`: Patient age (years).
    - `gender`: Male or Female.
    - `avg_glucose_level`: Average blood glucose level (mg/dL).
    - `bmi`: Body Mass Index.
    - `stroke`: Stroke status (0 = No, 1 = Yes).
- **Data Preparation**: I cleaned the dataset by removing missing values to ensure accurate analysis.

Tools and Environment

- **Programming Language**: Python.
- **Libraries**: Pandas for data manipulation and Matplotlib for visualization.
- **Platform**: Google Colab, a cloud-based environment I chose for its accessibility and pre-installed libraries.

Analysis Questions

I designed the project to answer the following clinical questions:

1. **What is the age distribution of patients?**
    - Visualized with a histogram to display the age range and distribution.
2. **How does average glucose level vary between patients with and without stroke?**
    - Analyzed using a boxplot, with mean and standard deviation statistics.
3. **Is there a correlation between BMI and stroke status?**
    - Evaluated with a scatter plot and correlation coefficient.
4. **What is the gender distribution among stroke patients?**
    - Presented with a pie chart showing the percentage of male and female patients.

Methodology

- **Data Loading**: I uploaded the dataset (CSV file) to Google Colab and loaded it using Pandas.
- **Data Cleaning**: I removed rows with missing values to streamline the analysis.
- **Analysis**: I computed descriptive statistics (e.g., mean, mode, standard deviation) and created visualizations (histogram, boxplot, scatter plot, pie chart) to answer each question.
- **Execution**: The analysis runs efficiently in Colab, completing in approximately 7 minutes.

Deliverables

- **Google Colab Notebook**: A `.ipynb` file I created, containing:
  - Python code for data loading, cleaning, and analysis.
  - Visualizations and written summaries addressing the four questions.
- **GitHub Repository**: I hosted the notebook in a GitHub repository (`clinical-data-analysis`) for version control and sharing.
- **Outputs**: Four visualizations and statistical summaries providing clear clinical insights.