Assignment 2

Stroke Prediction Dataset

Prepare a clean, well-documented Kaggle notebook that demonstrates full data analysis, cleaning, exploration, and preprocessing of the Stroke Prediction Dataset, including handling missing values using KNN imputation.

Dataset:

Please download the dataset from the following link: Dataset File

Requirements:

1. **Dataset Description and Problem Definition**
    - Brief description of the dataset
    - Define the problem you're addressing.
2. **Exploratory Data Analysis (EDA)**
    - Descriptive statistics for all relevant features.
    - Visualizations to understand distributions and relationships.
    - Identify potential issues like skewness.
3. **Data Cleaning**
    - Detect and fix any inconsistencies or errors.
    - Identify and treat missing values.
4. **Preprocessing**
    - Encode categorical variables (if needed).
    - Scale numerical features (if appropriate).
    - Use **KNN imputation** to fill in missing values.
5. **Final Dataset Preparation**
    - Return **a clean, well-structured DataFrame** ready for modeling.

Expected Deliverables:

- Your Public Kaggle notebook Link.

Instructions:

- Add **inline comments** to explain your code logic.
- Add **Markdown cells** to describe the purpose of each section.
- Do not use any Ai tool.

Submission Deadline: Wednesday 7/5/2025 before 11:59 PM.