PROJECT 1

Data Preprocessing is the process of transforming aw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.

Data preprocessing is important because it helps with
1. Accuracy
2. Completeness
3. Consistency
4. Timeliness
5. Believability
6. Interpretability

Data Preprocessing Consists of:
- Data cleaning
- Data integration
- Data reduction
- Data Transformation and Data Discretization

We will be using the Stroke Prediction Dataset
https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Context:
According to the World Health Organization stroke is the 2nd leading e 2nd leading cause of death globally, responsible for approximately 11% of total deaths.
This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

Within a python notebook you must do the following steps:

1. Generate the descriptive statistics (what are the observations) (BMI has 201 values missing)

2. Stroke frequency on different parameters. This is a barplot (E.g # of Males who had a stroke) (**You do not need to do this for BMI or age)** No need to describe the observations for this step.

3. Create a Distribution plot to understand how age impacts having a stroke. Describe your results.

4. Create a violin plot to understand the patients likelihood of getting a stroke.

5. Is this dataset imbalanced? The definition of an imbalanced dataset is refers to those types of datasets where the target class has an uneven distribution of observations. (Please write this answer in your code.)

6. Generate a heat map to understand the correlation among the variables. Describe your results. (Which variables have the strongest correlation)

7. Check for outliers in the BMI column and Average_Glucose Column and remove them.

8. Handle the null values for BMI . (DO NOT DROP the null values.) (Use one of the imputation methods we talked about in class)

9. Transform the variables that are an object datatype. (Use the Label Encoder library)