**ML Assignment 2 – Data Pre-Processing**

**Data Pre-Processing of Diabetes Dataset**

The objective is to perform a thorough pre-processing of the diabetes dataset. The goal is to understand the dataset's structure, clean and prepare the data for further analysis, and address any issues related to outliers and missing values.

**Dataset Link:**

<https://raw.githubusercontent.com/ArchanaInsights/Datasets/main/diabetes.csv>

**Pre-Processing Steps:-**

1. **Data Cleaning:**
   1. **Rename Columns:** Update column names such as 'ID' to 'Visit\_ID' and 'No\_Pation' to 'Patient\_ID' to make them more descriptive and clear.
   2. **Check Unique Values in Categorical Columns:** Identify the unique values in categorical columns:

**Gender**: F (Female) & M (Male),

**CLASS**: N (No diabetes), P (Pre-diabetes), Y (Yes, Diabetes).

Review and correct any unexpected or incorrect values if necessary.

* 1. **Statistical Summary of Numerical Columns:** Generate a statistical summary (mean, median, min, max, standard deviation) for all numerical columns to gain insights into the data's distribution.
  2. **Box Plot of Numerical Columns:** Create a box plot for all numerical columns in a single graph to visualize their distribution, range, and outliers. Explain any visible trends or outliers.
  3. **Identify Missing (Null) Values:** Determine the number of missing values in each column and report which columns have null values.
  4. **Impute Missing Values:** Select an appropriate strategy (mean, median, or mode) to impute missing values in the dataset. Justify the imputation technique for each column. After imputation, verify that no missing values remain.
  5. **Outlier Handling:** Outliers in medical data may represent critical conditions and, depending on the context, should be retained or removed with care.

1. **Retain Outliers:** Retain outliers in the AGE, HbA1c, and BMI columns, explaining why they should be preserved.
2. **Filter Outliers:** In the Creatinine Ratio (Cr) column, exclude values that exceed the 99.5th percentile threshold. Similarly, in the Urea column, exclude values that exceed the 99.9th percentile threshold. Use these specific percentile thresholds for filtering, and do not apply IQR or other methods.
3. **Remove Extreme Outliers:** For lipid-related columns (LDL, VLDL, HDL, TG, Chol), remove extreme outliers using an appropriate method such as IQR or Z-score.
   1. **Check for Duplicates:** Identify any duplicate rows and remove them to avoid redundant data in analysis.
4. **Data Transformation:**
   1. **Feature Engineering:** Apply Label Encoding or One-Hot Encoding to the categorical column Gender to convert it into numeric format for model-building purposes.
   2. **Feature Scaling:** Standardize or normalize the numerical columns that have varying scales (e.g., Patients, AGE, BMI, Cr) to ensure consistent contribution in subsequent analyses. Justify the scaling method used (standardization vs. normalization).