**Task 1 - Data Cleaning and Preprocessing**

# Objective:

To clean and prepare the 'Medical Appointment No Shows' dataset by handling missing values, duplicates, inconsistent formats, and invalid data to make it ready for analysis.

# Tools Used:

Python (Pandas library)

# Step-by-Step Process:

## 1. Loading the Dataset

Used pandas to load the dataset into a DataFrame and viewed the first few rows to understand its structure.

## 2. Understanding the Data

Checked the data types and basic statistics using .info() and .describe(). This helped in identifying inconsistencies and unusual values (like negative ages).

## 3. Identifying Missing and Duplicate Data

Used .isnull().sum() to check for missing values and .duplicated().sum() to count duplicates. Duplicates were removed using .drop\_duplicates().

## 4. Handling Invalid Entries

Filtered out rows with invalid age values (e.g., negative numbers). This ensures data integrity.

## 5. Date Format Conversion

Converted the 'ScheduledDay' and 'AppointmentDay' columns to datetime format using pd.to\_datetime(). This standardizes the date format and allows further time-based analysis.

## 6. Text Standardization

Standardized text fields like 'Gender' and 'No-show' using .str.upper().str.strip() for consistency. This helps in avoiding misclassification due to inconsistent casing or spacing.

## 7. Column Name Cleaning

Renamed all column headers to lowercase and replaced spaces with underscores using string methods. This improves readability and compatibility in coding.

## 8. Final Data Verification

Performed final checks to ensure no missing values or duplicates remained. Verified data types and unique values to confirm consistency and cleanliness.

# Summary:

The dataset was successfully cleaned and standardized using Python and Pandas. This included removing duplicates, fixing date formats, eliminating invalid data entries, standardizing categorical values and cleaning column names. The final cleaned dataset is saved as 'cleaned\_medical\_appointments.csv' and is now ready for further analysis or modeling.