



University of M'sila, Algeria

Computer science department

1st year undergraduate (2025/2026)

Introduction to Artificial intelligence

Lab4 (Data cleaning)

In **Lab 3**, you learned that intelligence is no longer trapped in the code but lives within the **data**. Since the data itself now carries the "knowledge," this lab focuses on **Data Preparation**, which is the most critical stage of the AI lifecycle. Just as you modified rows in Lab 3 to change how a program "learned" to predict labels, we will now learn to clean and structure raw files to ensure the machine finds the most accurate patterns possible.

Scenario:

Your brother has owned a gym for 5 years. He notices many members sign up but stop coming after a few months. He wants your help to identify "**Loyal**" vs. "**Leaving**" members so he can offer special discounts to those at risk of quitting.

Tasks:

1. **Create two CSV files** from the data provided in the tables below (data_part1.csv and data_part2.csv).
2. **Load and Inspect the Data:**
 - Load both files into Pandas DataFrames.
 - Check for missing values, inconsistent column names, or incorrect data types.
3. **Convert "Date of Birth" to "Age"** in data_part1.csv:
 - Calculate Age using the Date of Birth (assume the current year is **2026**).
 - Drop the original Date of Birth column.
4. **Combine Both Files** into a single DataFrame.
5. **Clean the Data:**
 - **Drop irrelevant columns:** Number, Full name, and Membership ID.

- **Handle missing values:**
 - Impute missing **Age** and **Monthly Visits** with the **median**.
 - Drop rows where the target (**Status**) is missing.
- **Encode the target variable:**
 - Convert Status from text (Loyal/Leaving) to numerical values (\$1\$ / \$0\$).

6. **Export the Cleaned Data:** Save the result to a new file named cleaned_data.csv.

Table1: data_part1

Number	Full name	Date of Birth	Membership ID	Monthly Visits	Status
1	Alice Smith	15/05/1990	M101	12	Loyal
2	Bob Jones	20/11/1985	M102		Leaving
3	Charlie Brown	05/02/2000	M103	15	Loyal
4	Daisy Miller	12/08/1998	M104	2	Leaving
5	Edward Norton	30/01/1982	M105	14	Loyal

Table2: data_part2

Number	Full name	Age	Membership ID	Monthly Visits	Status
6	Fiona Gallagher	28.0	M106	1	Leaving
7	George Costanza		M107	0	Leaving
8	Hannah Abbott	22.0	M108	18	Loyal
9	Ian Wright	45.0	M109	5	
10	Julia Roberts	35.0	M110		Loyal

Note: you may use AI tools to work on this Lab (chatgpt, claude AI,...)