DATA AND INFORMATION QUALITY PROJECT GUIDELINES

The project gives you the opportunity to obtain a maximum of **4** additional **points**.

PROJECT ASSIGNMENT

A) Compile the form (by **16/10**) by entering the group members
Please contact me (camilla.sancricca@polimi.it) if you have not been able to find a group

B) I will assign a different **dirty dataset** to each group (by 23/10)

C) You must execute a complete **pipeline** on the assigned dataset (deadline: first exam call on January 2026)

1/2-person projects: Design and implementation of a data preparation pipeline
3-person projects: Design and implementation of a data science pipeline (data preparation + data analysis)

Important Take advantage of the weekly exercises and apply what you have learn every week to your dirty dataset!

The **DATA PREPARATION PIPELINE** that you will perform on the assigned dataset will have the following steps:

1. Data profiling and data quality assessment
2. Data cleaning
   a. Data transformation/standardization (bringing everything to the same format, detecting and correcting typos, performing wrangling operations, etc.)
   b. Error detection and correction (dealing with missing values and the detection and correction of potential outliers)
   c. Data deduplication (detecting and handling non-exact duplicates)

Important After cleaning the data, verify the desired quality level has been achieved (additional data quality assessment — brief)

3. Data analysis [only for 3-people groups]
   a. Choose the type of analysis (classification-regression-clustering):
      i. Choose one column as the target column (categorical = **classification** OR numerical = **regression**)
      OR
      ii. Perform unsupervised **clustering** analysis
   b. Perform a data analysis pipeline on (1) the dirty dataset and (2) the cleaned dataset (model selection, training and testing)
   c. Compare the results using the right performance metrics (Precision, Recall, F1, etc. [Classification], MSE, RMSE, etc. [Regression], Silhouette, etc. [Clustering])

Important Some datasets that we will assign are not specifically made for machine-learning analysis! It is OK if the performance is very low. The important thing is that the dataset is cleaned properly and the pipeline is complete.

The **PROJECT REPORT** that you deliver must contain this information:

PROJECT ID
ASSIGNED DATASET
STUDENTS (NAME SURNAME ID)

1. SETUP CHOICES
Describe the setup choices made: libraries, data preparation techniques used, etc.

2. PIPELINE IMPLEMENTATION
Describe all the pipeline steps in detail: what did you find from the data exploration? How did you decide to use it in the data preparation phase? Why did you used specific that data preparation technique?

3. RESULTS [only for 3-people groups]
Discuss the main results obtained: verify the desired quality level has been achieved, compare the data analysis results

Very important Justify your choices! (for example, why you have chosen a specific data preparation technique for a specific column than all those seen in the lectures?) Justifications account for 1/3 of the final grade.

EVALUATION
You must deliver a **.zip** folder named with the project ID and your surnames
(example: 1_Sancricca_Sancricca_Sancricca.zip) containing:

1. A report of few pages — more details on writing the report at the end of the document
2. The code you made (.py, or .ipynb) — better if well-commented
3. The dataset you have cleaned

DEADLINE on the first exam call (**??/01/2026**)