U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO	**Master Degree in Artificial Intelligence** **Artificial Intelligence and Society**	**2024/2025** **1st Year** **1st Semester**
INSTRUCTOR: Miriam Seoane Santos		

Individual Assignment 1: Data-Centric AI and Data Profiling

Submission Deadline: Oct 20

Objective

The purpose of this assignment is to apply data profiling techniques to identify and discuss key data quality issues and data characteristics in a dataset of your choice. **The deliverables of this assignment are a 2-page report with your insights and a jupyter notebook with your experiments. The assignment might be subjected to defense.**

Instructions

Dataset Selection

- Start by selecting a dataset that you find interesting. Focus on strutured datasets, although you can select either tabular or time-series domains.
- Try to find a novel data rather than exploring well-known datasets such as the titanic, iris, wine, etc. These have already been extensively discussed.
- Be mindful of the number of observations and features to ensure there is enough complexity for analysis.

Data Profiling and Key Analysis

- Conduct a detailed data profiling process. You should focus on the components explored in the in-class tutorial and perform an in-depth analysis of some particular aspects of your dataset that you find relevant.
- Identify three key data quality issues or charcateristics in your dataset that you believe are worth investigating. For each issue, aim to explain:
    - Why is this an issue for data analysis and/or modeling? What could be possible consequences?
    - How did you detect it? Can you further characterize or quantify it?
    - Does the issue have social, moral, or societal implications?
    - What potential solutions or actions could be taken to address the issue?
- Include tables and visualizations to support your discussion and add any other elements that may help you communicate your insights and conclusions.

Deliverables

For this assignment, the following deliverables are required:

- A report (up to 2 pages -- using Template 1 or Template 2) summarizing your main findings. The report should include the following sections:
    - **Introduction:** Brief description of the dataset and its context;
    - **Data Profiling Summary:** Overall findings from the data profiling process;
    - **Key Issues and Analysis:** In-depth discussion of the identified data quality issues;
    - **Conclusions and Recommendations:** Final thoughts and suggestions for improving data quality and/or further exploration of the identified issues.
    - **References:** Cite all resources used in your work (e.g., books, research papers, websites). Use a consistent citation format.
- A Jupyter Notebook (or .py project) containing the complete code used for analysis, inline comments explaining important aspects and visualizations/data included in the report.

Evaluation Criteria

Your submission will be evaluated based on the following criteria:

- **Dataset Selection and Depth of Analysis**: Ability to select an interesting problem/domain, identify key data issues and demonstrate a deep understanding of their consequences, and suggest relevant strategies for further investigation or mitigation of the identified issues.
- **Clarity and Organization of Deliverables**: Writing and logical flow and structure of the report and the Jupyter Notebook (or project).
- **Bibliography**: Quality of the sources and their relevance to the analysis.

Bibliography

- **[Example]** Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., & Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(3), e1452.
- **Public Data Repositories:**
  - Kaggle Datasets
  - UCI Machine Learning Repository
  - data.world
  - Open ML (Bischl, B., Casalicchio, G., Feurer, M., Gijsbers, P., Hutter, F., Lang, M., ... & Vanschoren, J. (2017). Openml benchmarking suites. *arXiv preprint arXiv:1708.03731*.)