Data Mining

ENSIA 2025-2026

Lab sheet N°2: Data Preprocessing and Cleaning

Objectives

- Introduction to Data Preprocessing and Cleaning
- Load datasets and perform initial exploration to understand the data structure and statistics
- Handle duplicates and missing values, identify/remove outliers, smooth noisy data, etc.
- Transform data into a format suitable for analysis: sampling, encoding, normalization, discretization
- Gain practical experience using Python libraries for data preprocessing and cleaning

Required tools

Programming language: Python 3

Platforms: Anaconda, Jupyter Notebook, JupyterLab, Google Colab (cloud-based environment)

Python libraries:

- **Numpy:** A library for efficient numerical operations and multidimensional arrays, widely used in scientific computing and data analysis
- **Pandas:** A data manipulation and analysis library, providing data structures and functions to easily handle and process structured data.
- **Matplotlib:** A plotting library used for creating static, animated, and interactive visualizations.
- **Seaborn:** A data visualization library based on Matplotlib, providing high-level functions for creating attractive statistical graphics.

Resources

- Python for Data Analysis, 3E - 7 Data Cleaning and Preparation

Install required libraries

1. Activate the environment DM_ENV (created in Lab 1):

- conda activate DM_ENV

2. Install the required libraries if not already installed:

- conda install -c conda-forge matplotlib
- conda install -c anaconda seaborn

Part 1: Notebook 1: Lab2-pre-processing - part1 (90 minutes)

Part 2: Notebook 2: Lab2-pre-processing - part2 (90 minutes)