Laboratory 5 - Data Preprocessing

Prof. Dr. Laura E. Raileanu, Cédric Campos Carvalho, Elliot Ganty

27th November 2025

1 Introduction

This laboratory will help you go over the main steps of data preprocessing. The goal is to provide practical examples to help you understand the key concepts and make them easier to visualize.

1.1 Organisation

This laboratory must be carried out in groups of no more than 2 students.

1.2 Submission

Submit all four completed notebooks. Ensure you do not clear the outputs, as some answers may include plots in the output cells.

Please write the names of the members of the group **in each** Notebook filename (i.e. `X_Labo5_DataX_Name1_Name2.ipynb`) as they will be reviewed separately.

No report is needed as all questions can be answered directly in the notebook files.

Then, upload **all notebook files** in Moodle, the deadline is **7th January, 2026 at 23:59:59**.

2 Setup

To complete the required tasks within this laboratory, it is preferably to use **Python version 3.9** or above. You may find it also useful to create a virtual environment. The required dependencies are described in the given `requirements.txt` file.

3 Objective

In this laboratory, you receive multiple files that you need to complete. Each of them is a Jupyter Notebook that will guide you through all the exercises that you need to complete. Please read **all** the cells as they are giving you instructions that should be followed at each step.

There's a total of four different exercises :

- "*Data Integration*", an introduction on how to handle multiple data sources.
- "*Data Transformation*", an introduction to data Discretization and concept hierarchy generation.
- "*Data Cleaning*", where you explore how to handle noisy data, missing values, inconsistencies and outliers.

- "*Data Reduction*", where you explore how to reduce the volume of a dataset while trying to keep the same analytical results.