Data Mining

ENSIA 2025-2026

Lab sheet N°4: Dimensionality Reduction

Objectives

- Introduction to **Dimensionality Reduction** and its importance in reducing the number of variables
- Introduction to the Python library **Scikit-learn** for data analysis and modeling
- Use common dimensionality reduction techniques in **Scikit-learn**, including **PCA**
- Visualize the reduced data to gain insights into its characteristics
- Understand how to evaluate the effectiveness of dimensionality reduction

Required tools

Programming language: Python 3

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

Python libraries:

- **Scikit-learn:** A data analysis and modeling library, including Data mining and Machine learning algorithms for various tasks: classification, regression, clustering, dimensionality reduction, …

Resources

Scikit-learn:

- **Reference card:** Python For Data Science Cheat Sheet Scikit-Learn
- **Official documentation:** https://scikit-learn.org/stable/
- **User guide:** User Guide — scikit-learn 1.7.2 documentation
- **Examples:** Examples — scikit-learn 1.7.2 documentation
- **Dataset loading utilities:** 8. Dataset loading utilities — scikit-learn 1.7.2 documentation

PCA:

- **Unsupervised reduction:** https://scikit-learn.org/stable/modules/unsupervised_reduction.html
- **Official documentation:** PCA — scikit-learn 1.7.2 documentation
- **Examples:** Gallery examples: PCA — scikit-learn 1.7.2 documentation

Notebooks

- **Exercise:** PCA from scratch.ipynb
- **Guided Notebook:** Principal_Component_Analysis_(Guided).ipynb
- **Non-guided Notebook:** Principal_Component_Analysis_(Non_Guided).ipynb

Part 1: Dimensionality reduction with Scikit-learn and PCA (150 minutes)

- Activate **DM_ENV** and install the required Python libraries: **Scikit-learn**
- Alternatively, students who find problems setting up the environment can use Google Colab
- Execute and understand the **guided** Jupyter Notebook file, in local or on Google Colab
- Fill in the gaps and write the missing code in the **non-guided** Jupyter Notebook file
- Take a look at the provided resources (documentation, tutorial, reference card) for more info

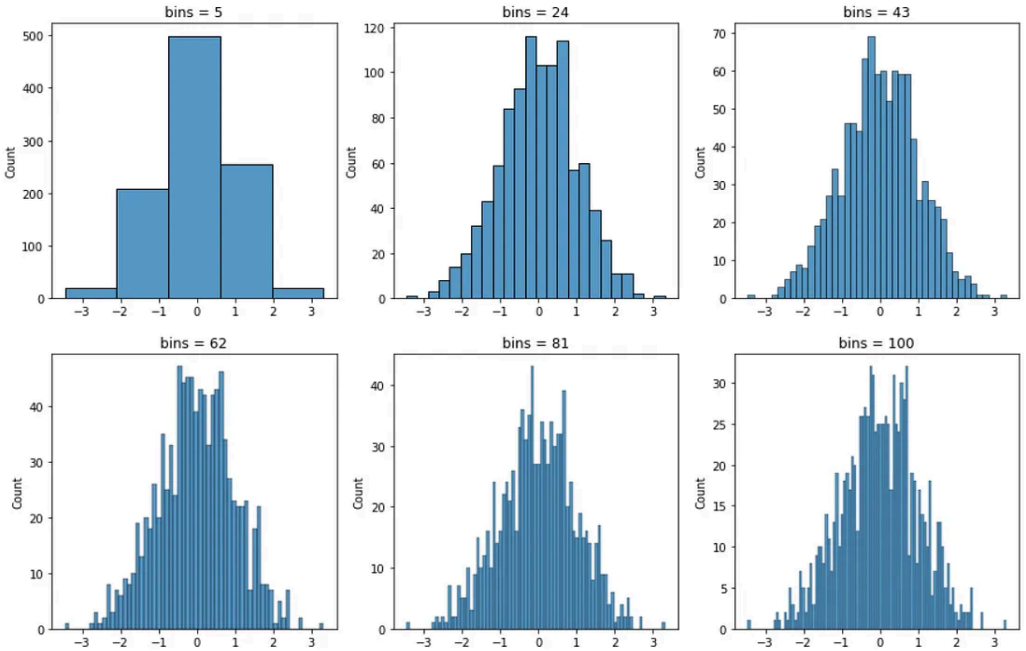
Part 2: Exercises on the Chapter Data - Part 2 (30 minutes)

1. How can a box plot be used to explore a data set containing four attributes: age, weight, height, and income?

2. Explain how a box plot can indicate whether the values of an attribute are symmetrically distributed.

3. What are the advantages of using a box plot over a histogram when exploring data distributions, and in what scenarios would a histogram provide more insight?

4. How does the choice of the number and location of bins affect the appearance of a histogram, and what methods can be used to reduce this dependency?



5. How can histograms, box plots, and density plots reveal information about skewness? Additionally, what strategies can be employed to handle skewed data?

6. When should you use equal-width bins versus equal-frequency bins? What considerations should be taken into account when making this choice?

7. How do you interpret the results of a quantile plot? What do straight lines and deviations from the line indicate about the distribution of the data?

8. When might stratified sampling be more beneficial than simple random sampling, and how does it influence the representation of subgroups within a data set?