U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO	First Degree in Artificial Intelligence and Data Science **Elements of Artificial Intelligence and Data Science**	2024/2025 1st Year 2nd Semester
TEACHERS: Luís Paulo Reis, Miriam Santos, Rita Ribeiro, Moisés Santos		

Exercise Sheet 4
Data Science Tools and Data Exploration

4.1 Software/Packages Installation

Python offers the full stack of software packages for Scientific and Machine Learning data analysis, from the NumPy to the Scikit-learn libraries. Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression, and clustering algorithms including neural networks, support vector machines, random forests, gradient boosting, k-means and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

In this exercise, we will install all the necessary libraries to work with Python Scientific Environment. **Please note, that some of these procedures will only need to be done in your personal computer as the computers from the LabCC already offer a pre-installation of these packages**. Moreover, given the permission system you may not be able to install additional packages at the LabCC computers, rather than those already pre-installed.

Start by installing Python, Anaconda, Jupyter Labs, NumPy, SciPy, Pandas, Scikit-Learn, MatPlotLib and Seaborn. In fact, it is only needed to install Anaconda that contain all the others following the link: https://www.anaconda.com/products/individual

Information about the rest of the packages/libraries may be found at:

- Python: https://www.python.org/
- Anaconda: https://www.anaconda.com/
- Project Jupyter: https://jupyter.org/
- NumPy: https://numpy.org/
- SciPy: https://www.scipy.org/
- Pandas: https://pandas.pydata.org/
- Scikit-Learn: https://scikit-learn.org/
- MatPlotLib: https://matplotlib.org/
- Seaborn: https://seaborn.pydata.org/

Consider using a new virtual environment for this part of the course. After installing all the packages, please open the example Notebook available at moodle that contains an example code containing several exercises: **Auto.ipynb**.

4.2 Test installation and open Jupyter Notebook

Once your packages are installed, open a new Jupyter Notebook and test the installation by importing the modules:

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
…
```

Explore the **Auto.ipynb** and how these packages are used. Execute step-by-step the instructions in those notebooks and analyze the results. Follow along with their official documentation, available online, and the additional material available on Moodle.

4.3 Data Exploration and Visualization

To complete these exercises we recommend that you follow the notebooks and additional material provide. The exercises proposed here follow pretty much the same ideas. Also, always consult the packages documentation and examples.

1. Create a notebook file name Default.ipynb
2. Load the modules pandas, seaborn, and numpy.
3. Open the provided file Default.tab (read_csv from pandas). Load into a variable called *df*.
4. Visualize the first rows (.head()) and the columns names (.columns). Get the type of attributes. Use .info().
5. Calculate the mean of *balance* and *income* attributes.
6. How many are students? Use *value_counts().*
7. Produce a scatter plot between income and balance. Use

   from pandas.plotting import scatter_matrix
8. Select the sub dataset with only individuals that are students.
9. Select a subset of the dataset with only the balance and the income attributes.
10. Plot the distribution of balance and the income attributes with boxplot (.boxplot()).
11. Calculate the difference of the means in the income of students and non-students.
12. Check if there are null or NAs.
13. Use *seaborn* and plot the boxplot of income with relation to student attribute.
14. Use *seaborn* and plot the boxplot of balances with relation to default attribute.