U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO	L.EIC – BSc/First Degree in Informatics and Computing Engineering **Artificial Intelligence**	2023/2024 (3rd Year) 2nd Semester

Assignment No. 2

Supervised Learning

Theme

IART's second practical assignment consists in the application of machine learning models and algorithms related to supervised learning.

In supervised learning problems, the idea is to learn how to classify examples in terms of the concept under analysis. An initial exploratory data analysis should be carried out (class distribution, values per attribute, and so on). Different learning algorithms should be employed and compared using appropriate evaluation metrics (performance during learning, confusion matrix, precision, recall, accuracy, F1 measure) and the time spent to train/test the models.

Supervised learning includes the following steps: dataset analysis to check for the need for data pre-processing, identification of the target concept, definition of the training and test sets, selection and parameterization of the learning algorithms to employ, and evaluation of the learning process (in particular on the test set). At least 3 supervised learning (classification) algorithms should be employed (Decision Trees, Neural Networks, K-NN, SVM, …), but more may be employed and compared using the Scikit-Learn Python library and considering the characteristics of the dataset. Results should be compared using tables or plots (e.g., using Seaborn or Matplotlib libraries).

Programming Language/Libraries

Any programming language and development system can be used, including, at the language level, Python, C++, Java, C#, among others. However, it is strongly advised to use Python due to the availability of very strong machine learning libraries for this language. Although you may use any library or tool specific for developing supervised machine learning models (after validating it with the course teachers), it is highly advisable that the libraries used are the ones lectured on the course, such as Pandas, NumPy/SciPy, Scikit-learn and Matplotlib/Seaborn.

Groups

Groups must be composed of 3 students (exceptionally 2). Groups should be composed of students attending the same practical class. All students in a group must be present in the checkpoint sessions and presentation/demonstration of the work. Groups composed of students from different classes are discouraged, given the logistic difficulties of performing work that this can cause.

Checkpoint

Each group must submit in Moodle a brief presentation (max. 5 slides), in PDF format, which will be used in the class to analyse, together with the teacher, the progress of the work. The presentation should contain (1) a specification of the work to be performed (definition of the machine learning problem to address), (2) related work with references to works found in a bibliographic search (articles, web pages and/or source code), (3) a description of the tools and algorithms to use in the assignment, and (4) implementation work already carried out.

Final Delivery

Each group must submit in Moodle two files: a presentation (max. 10 slides), in PDF format, and the implemented code, properly commented, including a "readme" file with instructions on how to compile, run and use the program. The code and comments may be submitted as a complete Jupyter Notebook. Based on the submitted presentation, students must carry out a demonstration (about 10 minutes) of the work, in the practical class, or in another period to be designated by the teachers of the course.

The file with the final presentation should include, in addition to the aforementioned for the checkpoint, details on data pre-processing, the developed models and their evaluation and comparison, using appropriate graphical elements (tables, plots, etc.).

Datasets

A) Autism Dataset for Toddlers

B) Bank Client Attributes and Marketing Outcomes

C) Banking Customer Churn Prediction Dataset

D) BCCC-CIRA-CIC-DoHBrw-2020

E) Breast Cancer

F) Drug Consumption Classification

G) Evasive PDF Samples

H) Glioma Grading Clinical and Mutation Features

I) League of Legends SoloQ matches at 15 minutes 2024

J) Metaverse Financial Transactions Dataset

K) NASA Asteroids Classification

L) SDSS Galaxy Classification DR18

M) Spruce tree type Detection

N) Steel Plate Defect Extended Dataset

O) Titanic Dataset