Lab 4: Comparison of Classifiers (2%)

This exercise introduces the main components of the predictive data-analytics workflow. These are *model training* and *model evaluation*. You are expected to train and evaluate a few alternative classification models (i.e., classifiers) with the same dataset, and select the best among them by evaluating their performance with a few different metrics.

Please note that the focus of this exercise is NOT on understanding how the classifiers get trained. The classification algorithms used in the example notebook below are Support-Vector Machines (SVM) and Random Forest. **You are NOT expected to fully understand how they work.** For this exercise, it is sufficient to understand that a classification algorithm takes a training data set as an input and somehow trains a model that can be used to predict the value of a particular binary attribute.

The focus of this exercise is on understanding the correct workflow for training and comparing the performance of a few alternative classifiers.

Preparation

- Watch the video playlist Exercise 4 (ca. 40 min).
- Read the article How and When to Use ROC Curves and Precision-Recall Curves for Classification in Python.

Task 1

- Download the following files:
  - Lab 4 - Comparison of Binary Classifiers.ipynb - this is the example notebook to follow in this exercise. It contains code for training and evaluating SVM and Random Forest classifiers.
  - bcwd.csv - this is the dataset used in the example notebook.
  - seeds.csv - this is the dataset to work with in Tasks 2 and 3. It is taken from https://archive.ics.uci.edu/ml/datasets/seeds.

Task 2

- Replicate the classifier training and evaluation demonstrated in the example notebook **Lab 4 - Comparison of Binary Classifiers.ipynb** but with the dataset **seeds.csv**. You will need to formulate a binary classification problem and transform the type column accordingly. That is, replace the **type** column with a binary column called **class** in which one of the original three types (it doesn't matter which one you choose) is **class 0**, and the other two types are **class 1**.

Task 3

- Train a third probabilistic classifier (e.g., MLPClassifier, Naive Bayes, kNN) with **seeds.csv** add compare it to both SVM and Random Forest. You may encounter warnings for 0 values of some of the metrics. Ignore them and aim at having at least one classifier that has acceptable results.

Submission

Save your Jupyter notebook and name it *lab4.ipynb*.

Submit your notebook in the Course Tools > Assignment section by 15-Mar-2024 23:59. Late submissions will not be accepted.