Lab 2: Data Mining using scikit-learn

The task is to create a notebook (ipynb file) in which the student will attempt to participate in the EMVIC 2012 (Eye Movement Verification and Identification Competition).

Information about the competition is available here:

http://kasprowski.pl/emvic2012/

https://www.kaggle.com/c/emvic

The participants' task was to train a model using the signed training data (train_a.emd) and apply the model to the unsigned test data (test_a.emd). The prediction file was sent to the organizers and the participant was informed about the score (but only the aggregate score, e.g. accuracy: 82%). Multiple trials were allowed.

The dataset (which is the part of the dataset used for the actual competition) should be downloaded from platforma.polsl.pl. The first column is the class (user identifier as aXX), the meaning of the other columns is described on the EMVIC 2021 web page.

You should divide the given dataset into training and testing sets, build a model using the training set, and report results for the testing set. You are allowed to use any algorithms and models from the scikit-learn package. Use several machine learning algorithms and different conversions and compare the results! Tips on how to process the data can be found in the lecture notebooks, among other things.

HINT: When opening the dataset using Pandas remember that the column separator is not a comma but a tab (\t) and the file doesn’t have a header.

Rules for creating a notebook:

1. All operations on data must be documented in the notebook - so that they can be repeated by running the notebook again!
2. The saved notebook should contain all output messages, so that it will be possible to check how it works even without running it.
3. Comments and descriptions should be in the markdown fields.

The assessment depends on the scope of the work performed:

For grade 3, you need to run 2 classifiers and compare the results.

For grade 4.5:
Prepare a dataset - some columns are unnecessary.
Divide the dataset into training and testing sets.
Perform experiments on at least 7 classifiers.
Perform cross-validation experiments on the 3 best classifiers
Perform feature selection and run the experiments on the top 3 classifiers.

For grade 5 - additionally:
Illustrate an example of eye movement for a given person (left, right, and displayed point).
Optimize 1 classifier by selecting the appropriate parameters or add additional columns with processed data (e.g. the difference between the right and left eye, speed, acceleration of eye movement,) or normalize the data or use another interesting idea.
Perform experiments.
Save the best dataset to a file.