Exercise 9 (practical) for the lecture
Big Data Analytics

WS 2021/2022

The submission due date of this task is Friday, Jan. 28th.
Please write the names and matriculation numbers of all group members in your submitted files.

1 Practical tasks (1P)

You are given a synthetic dataset which is divided into two parts, the labeled subset `synthetic_dataset.csv` and the unlabeled subset `synthetic_dataset_evaluation.csv`. You are asked to apply the data analytics approaches learned from the lecture to make prediction on the unlabeled subset.

The dataset contains 6 numerical columns and 8 categorical columns. The `Label` column indicates the class label `Yes` or `No`. The task is to preprocess the dataset, train a classifier on the labeled subset and predict a probability for the target `Label` being `Yes` for each entry in the unlabeled subset . *Hint: you can run cross-validation on the labeled subset with different classifiers and use the best classifier to make prediction on the unlabeled subset.*

As the evaluation metric, we will use the area under the ROC curve (AUC[1]). An AUC score above 0.6 (with valid JupyterNotebook implementation) will already bring you the full point of this task. Additionally, we will publish the top-ranked teams with their AUC score on Moodle as the competition result.

To submit, you should upload two files for this task. A JupyterNotebook(`.ipynb`) containing your implementation and a prediction file (`.csv`). Each row in the prediction file contains the entry id and the prediction probability separated by a comma. An example of the prediction file can be found as `sample_submission.csv` (any other format of submission will not be considered).

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.auc.html`