# Exercise 1 - Anonymisation

Group 7 - Oliver Stritzel - Dejana Stevanovic

## Dataset

As a dataset we used the adults dataset from the UCI ML Repository. The dataset consits of a target variable, the income, which is denoted as higher or lower than 50.000$ per year income. The dataset includes 14 variables, consisting of categorical and numerical values, with over 32000 rows.

The whole code used to achieve the results can be found in the public repository

## Pre-Processing and General Classifier

We pre-processed the dataset, one-hot encoding categorical features, dropping some redundant ones (f.i. *education-num*). Rows with missing values were dropped. After pre-processing, we split the data into test and train (80/20) and trained three baseline classifiers on it. This was done without any anonymization. As metrics we used *f1-score* and *roc-auc score*.

The results shown below.

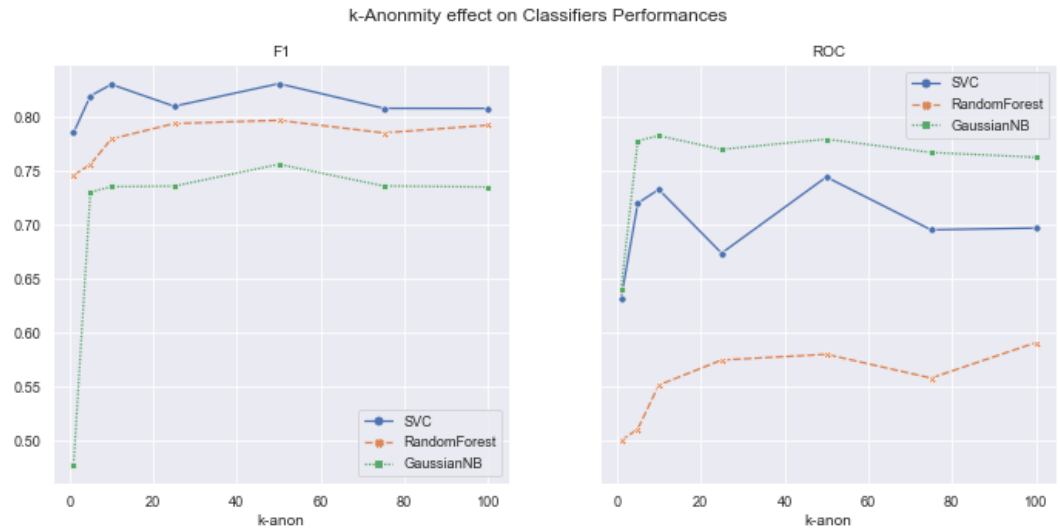
| | SVC | RandomForest | GaussianNB |
|---|---|---|---|
| f1 | 0.818 | 0.746 | 0.534 |
| roc_auc | 0.722 | 0.5 | 0.673 |

## k-Anonymity

For k-Anonymity, we used the java library ARX. Using the raw data as input we created the hierarchies to anonymize categorical variables. For the following variables hierarchies were created and they were used in different levels of k-Anonymity:

| age | capital-gain | capital-loss | education | fnlwgt | hours-per-week | marital-status | native-country | occupation | race | relationship | sex | workclass |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| x | | | x | | | x | x | x | x | x | x | x |

We created datasetes for different *k*'s, namely which were used as input for the same code that was used above. We ended up with the following results:




Comparing this with the results from above, one can see that some k-anonymity had a positive effect on some of the classifiers. With a higher number of *k* though, the performance decreased a little again. Especially the Random Forest did increase his performance. This leads to the assumption that indeed some level of generalization through anonymization was good for the classifiers performance, which might have overfitted on the non-generalized data, but too much generalization might worsen the effect again.

## Microaggregation