**Pre-processing**

All ten batches will be combined into a single data set. An additional column ‘BATCH’ denoting the batch to which the observation belongs will be added. Next, all features will obtain an appropriate column name.

**Data exploratory analysis**

First, the general characteristics of the dataset itself will be investigated. More precisly, the shape of the dataset and the amount of observations for each gas will be requested. Next, the distribution of the target variable will be analysed. This by creating plots that demonstrate the targets variable global distribution as well as its distribution for each particular gas. Then, scatterplots representing the mean and standard deviation will be created. These plots will again be created for the full dataset and for each particular gas. At last an isolation forest model will be fitted to the data to obtain an estimate of the proportion of outliers according to the model. The observations that are identified as outliers will be further investigated. To be precise, it will be examined whether there is a particular gas or particular concentration that frequently is identified as outlier and what the potential reasons can be.

**Research**

In general, we want to apply 6 machine learning models for each particular task and epress their performance by using different metrics. First these models will be trained on data without feature selection. Next, the same machine learning models will be trained on the same data but after the applycation of different dimensionality reduction mechanisms. The same performance metrics will be computed, to compare performance with the baseline models.

The research starts with the creation of 8 different datasets, 1 dataset without the gas feature, 1 dataset that includes a categorical gas feature and 6 datasets whereby each contains only the observations of a particular gas. Then, each of these datasets will be divided into a training and test set. Next, a MLPRegressor, kNeighborsRegressor, ElasticNet, HuberRegressor, BayesianRidge and AdaBoost model will be created and fitted to each training set. Hyperparameters will be tuned by applying a grid search with 5-fold-crossvalidation. However, if the amount of hyperparameter combinations is too large to evaluate separately, a random search will be used instead. The resulting fitted models will be used to predict the target variable for each test set. Consequently, the MSE, MAE and R^2 computed.

**STUK DIMENSION REDUCTION**

**Note: If a algorithm makes use of distance metric, normalisation will be applied to the data.**