**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**

Three tasks have to be fulfilled. A unique approach will be performed for each task. However, each of these approaches will start with the creation of a baseline-model. Such that the basic performance of different models can be compared.

**1. Single regression model to estimate the concentration of any gases**

The gas feature is dropped from the dataset. A linear regression model will be created as baseline model to compare the performance. Next the dataset will be normalized and PCA will be applied with a minimum variance of 95%. A tree-boosting algorithm will be applied to the PCA output. To tune the hyperparameter ‘max\_depth’ a grid search with cross-validation (k=3). Next, the hyperparameter ‘n\_estimators’ will be tuned again with a grid\_search and cross-validation (k=10).

**2. Model with all the data and gas feature**

Dummy coding will be used to transform the categorical “gas” to k numerical columns. A linear regression model will be created as baseline model to compare the performance. Next, a random forest algorithm will be applied to indentify the feature importance of all the non-gas columns. Features that are indentified as not important are removed. A grid search with k=10 cross-validation will be used to tune the hyperparameter.

**3. Model for each gas**

The sub dataset for each gas will be created. Feature agglomeration will be used as dimension reduction technique to combine similar features. Next an elastic net regression model will be created for each dataset. The hyperparameter will be tuned by the use of cross-validation.

**No exact dimension reduction technique is used.**