**Pre-processing**

The data is saved into ten files. These files will be combined together to obtain a dataframe with all observations. Next, a missing value analysis will be performed. Each observation with a missing value for the target variable will be removed. Columns with more than 30% missing values will be removed. The proportion of rows with at least one missing value will be determined. If this proportion is more than 30%, then the missing values will be replaced by mean imputations. If the proportion is less than 30%, each of these rows will be removed. An isolation forest will be used to detect the proportion of outliers. …

The different tasks depent on whether gas type is used as feature. To obtain an idea of the influence of gas as an unique feature, the difference of the mean …

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. 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.**