R Notebook

What was done:

- split Data (80/20) in Train (data_train) and Test (data_test)

Used RandomForest,Stepwise and Lasso-regression for featureselection on Train set.
- builded 5 train, test splits (train_rows) out of Train (data_train)
- used step-algorithm to select 5 formulas with 1 to 5 features on every train/test split (train_rows) –> 25 formulas - selected features are based on ascending advanced R^2 - used RandomForest-algorithm to select 5 formulas with 1 to 5 features on every train/test split (train_rows) –> 25 formulas - selected features are based on ascending importance measure (https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/importance) - used lasso-regression with cross validation to get formulas from 1 to 5 features. The number of features that are used in a lasso regressin are based on the penalty parameter lambda. The cv.lmnet Lasso function uses 100 different lambda to fit 100 different lasso-regressions and to calculate the MSE with a 5- fold cross validation (same trainrows)

The found formulas and percentage occurence of the features were combined in the following tables: Formulas with only 1 feature and their occurence in percentage.:

```
occurence_feature_1_analysis[,1:4]
```

```
##       q_mean_mean_12 q_mean_mean_123 n_formulas
## step             0.8             0.2          2
## rf               0.8             0.2          2
## lasso            1.0              NA          1
##                                            found_formulas
## step    formula_1_: q_mean_mean_123  formula_2_: q_mean_mean_12
## rf      formula_1_: q_mean_mean_12  formula_2_: q_mean_mean_123
## lasso                                formula_1_: q_mean_mean_12
```

Lasso only found the feature : q_mean_mean_12 step and RandomForest both found the features : q_mean_mean_12 (in 80% of the train/test_splits from the train_rows) and q_mean_mean_123 (in 20% of the train/test_splits from the train_rows)

Formulas with only 2 feature and their occurence in percentage.:

```
occurence_feature_2_analysis[,1:6]
```

```
##       q_mean_mean_12 q_mean_mean_123 r_mean_mean_12 r_mean_mean_34 n_formulas
## step             0.8             0.2            0.8            0.2          3
## rf               1.0             1.0             NA             NA          2
## lasso            1.0             1.0             NA             NA          1
##
## step    formula_1_: q_mean_mean_123 + r_mean_mean_12  formula_2_: q_mean_mean_12 + r_mean_mean_12  fo
## rf                                                       formula_1_: q_mean_mean_12 + q_mean_mean_123  fo
## lasso                                                                                                  fo
```

Formulas with only 3 feature and their occurence in percentage.:

```
occurence_feature_3_analysis[,1:9]
```

```
##       q_mean_mean_12 q_mean_mean_123 r_mean_mean_12 r_mean_mean_1234
## step             0.8             0.2            0.8              0.2
```

```
## rf               1.0             1.0             NA               NA
## lasso            1.0             1.0            1.0               NA
##       r_mean_mean_234 r_mean_mean_34 q_mean_mean_123:r_mean_mean_12 n_formulas
## step              0.2            0.4                           0.2          5
## rf                 NA             NA                            NA          4
## lasso              NA             NA                            NA          1
##
## step    formula_1_: q_mean_mean_123 + r_mean_mean_12 + q_mean_mean_123:r_mean_mean_12  formula_2_: q
## rf                            formula_1_: q_mean_mean_12 + q_mean_mean_123 + q_mean_mean_23  formula
## lasso
```

Formulas with only 4 feature and their occurence in percentage.:

```
occurence_feature_4_analysis[,1:14]
```

```
##       i_mean i_mean_mean_34 q_mean_mean_12 q_mean_mean_123 r_mean_mean_12
## step     0.2            0.2            0.8             0.2            0.8
## rf        NA             NA            1.0             1.0             NA
## lasso     NA             NA            1.0             1.0            1.0
##       r_mean_mean_1234 r_mean_mean_234 r_mean_mean_34
## step               0.2             0.4            0.4
## rf                  NA              NA             NA
## lasso               NA              NA             NA
##       q_mean_mean_12:r_mean_mean_1234 q_mean_mean_123:r_mean_mean_12
## step                              0.2                            0.2
## rf                                 NA                             NA
## lasso                              NA                             NA
##       i_mean_abs_5:i_mean_mean_123 n_formulas
## step                            NA          5
## rf                              NA          5
## lasso                            1          1
##
## step                                                                                        forr
## rf      formula_1_: q_mean_mean_12 + q_mean_mean_123 + q_mean_mean_23 + i_mean_abs_3:r_mean_mean_34
## lasso
##       found_best_formula
## step               FALSE
## rf                  TRUE
## lasso              FALSE
```

Formulas with only feature and their occurence in percentage.:

```
occurence_feature_5_analysis[,1:19]
```

```
##       i_mean i_mean_mean_34 q_mean_mean_12 q_mean_mean_123 q_mean_mean_2345
## step     0.2            0.2            0.8             0.2              0.2
## rf        NA             NA            1.0             1.0               NA
## lasso     NA             NA            1.0             1.0               NA
##       r_mean_mean_12 r_mean_mean_1234 r_mean_mean_45 r_mean_mean_234
## step             0.8              0.2            0.2             0.4
## rf                NA               NA             NA              NA
## lasso            1.0               NA             NA              NA
##       r_mean_mean_34 i_mean_mean_34:r_mean_mean_34
## step             0.4                          0.2
## rf                NA                           NA
## lasso             NA                           NA
```

```
##       q_mean_mean_12:r_mean_mean_12 q_mean_mean_12:r_mean_mean_1234
## step                           0.2                              0.2
## rf                              NA                               NA
## lasso                           NA                               NA
##       q_mean_mean_123:r_mean_mean_12 i_mean_abs_1:i_mean_abs_5
## step                            0.2                        NA
## rf                               NA                        NA
## lasso                            NA                         1
##       i_mean_abs_5:i_mean_mean_123 n_formulas
## step                            NA          5
## rf                              NA          5
## lasso                            1          1
##
## step
## rf      formula_1_: q_mean_mean_12 + q_mean_mean_123 + q_mean_mean_23 + i_mean_abs_3:r_mean_mean_34 +
## lasso
##       found_best_formula
## step               FALSE
## rf                 FALSE
## lasso               TRUE
```

Results: - Features found by step and RandomForest are strongly dependend on the given training data. (Here Isar) - Lasso finds always the same formula with the same features

Second step:

- use the unique found formulas by the three algorithms to build linear models by training them on the whole data_train.
- The builded models were used to predict the never seen FIB concentrations from data_test
- the model with the lowest mean-squared error on data_test is the best_model. The used formula is the "best formula"

Best Formula with 1 coef:

```
best_formula_with_coef_1
```

```
##   formula_with_lowest_mse_on_test       mse
## 1                 q_mean_mean_12 0.7506375
```

Best Formula with 2 coef:

```
best_formula_with_coef_2
```

```
##    formula_with_lowest_mse_on_test     mse
## 1 q_mean_mean_12 + q_mean_mean_123 0.75341
```

Best Formula with 3 coef:

```
best_formula_with_coef_3
```

```
##                                    formula_with_lowest_mse_on_test       mse
## 1 q_mean_mean_123 + q_mean_mean_12 + i_mean_mean_123:i_mean_mean_23 0.6897531
```

Best Formula with 4 coef:

```
best_formula_with_coef_4
```

```
##                                                    formula_with_lowest_mse_on_test
## 1 q_mean_mean_123 + q_mean_mean_12 + i_mean_mean_123:i_mean_mean_23 + q_mean_mean_23
##         mse
## 1 0.6798171
```

Best Formula with 5 coef:

```
best_formula_with_coef_5
```

```
##                                                                                  formula_with_lowest_ms
## 1 q_mean_mean_12 + q_mean_mean_123 + r_mean_mean_12 + i_mean_abs_1:i_mean_abs_5 + i_mean_abs_5:i_mean
##         mse
## 1 0.6527487
```

Last step:

Check if different algorithm types found the "best-formula"

All three Algorithms found best formula with n_coef = 1

```
occurence_feature_1_analysis[,-c(1:4)]
```

```
##       found_best_formula mse_best_formula
## step                TRUE   q_mean_mean_12
## rf                  TRUE   q_mean_mean_12
## lasso               TRUE   q_mean_mean_12
```

only lasso and rf found best formula with n_coef = 2

```
occurence_feature_2_analysis[,-c(1:6)]
```

```
##       found_best_formula                 mse_best_formula
## step               FALSE q_mean_mean_12 + q_mean_mean_123
## rf                  TRUE q_mean_mean_12 + q_mean_mean_123
## lasso               TRUE q_mean_mean_12 + q_mean_mean_123
```

only rf found best formula with n_coef = 3

```
occurence_feature_3_analysis[,-c(1:9)]
```

```
##       found_best_formula
## step               FALSE
## rf                  TRUE
## lasso              FALSE
##                                                           mse_best_formula
## step  q_mean_mean_123 + q_mean_mean_12 + i_mean_mean_123:i_mean_mean_23
## rf    q_mean_mean_123 + q_mean_mean_12 + i_mean_mean_123:i_mean_mean_23
## lasso q_mean_mean_123 + q_mean_mean_12 + i_mean_mean_123:i_mean_mean_23
```

only rf found best formula with n_coef = 4

```
occurence_feature_4_analysis[,-c(1:13)]
```

```
##       found_best_formula
## step               FALSE
## rf                  TRUE
## lasso              FALSE
##                                                                            mse_best_formula
## step  q_mean_mean_123 + q_mean_mean_12 + i_mean_mean_123:i_mean_mean_23 + q_mean_mean_23
## rf    q_mean_mean_123 + q_mean_mean_12 + i_mean_mean_123:i_mean_mean_23 + q_mean_mean_23
## lasso q_mean_mean_123 + q_mean_mean_12 + i_mean_mean_123:i_mean_mean_23 + q_mean_mean_23
```

only lasso found best formula with n_coef = 3

```
occurence_feature_5_analysis[,-c(1:18)]
```

```
##       found_best_formula
```

```
## step               FALSE
## rf                 FALSE
## lasso               TRUE
##                                                                                                                                ms
## step  q_mean_mean_12 + q_mean_mean_123 + r_mean_mean_12 + i_mean_abs_1:i_mean_abs_5 + i_mean_abs_5:i
## rf    q_mean_mean_12 + q_mean_mean_123 + r_mean_mean_12 + i_mean_abs_1:i_mean_abs_5 + i_mean_abs_5:i
## lasso q_mean_mean_12 + q_mean_mean_123 + r_mean_mean_12 + i_mean_abs_1:i_mean_abs_5 + i_mean_abs_5:i
```

Step-algorithm only found in 1 out of 5 cases (20%) the best formula. Only in the one with only 1 variable

RandomForest found the best formula in 4 out of 5 cases (80%).

Lasso found the best formula in 3 out of 5 cases (60%).