# report_executed

August 27, 2021

1 Statistics

In the statistics module we analyze data for different responses and at different spectral peak locations. We use Python package scipy in this module.

1.1 T-Test

T-test checks for difference in the mean between two sample from different responses. We assume the data is independent and follows the normality assumption. Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ be the two samples and we test whether the means are equal. The null hypothesis states means μ_1 and μ_2 are equal and the alternative hypothesis states they are not equal. If the p-value is lower than the chosen significance level, we can reject the null hypothesis, i.e. the samples do not have the same means.
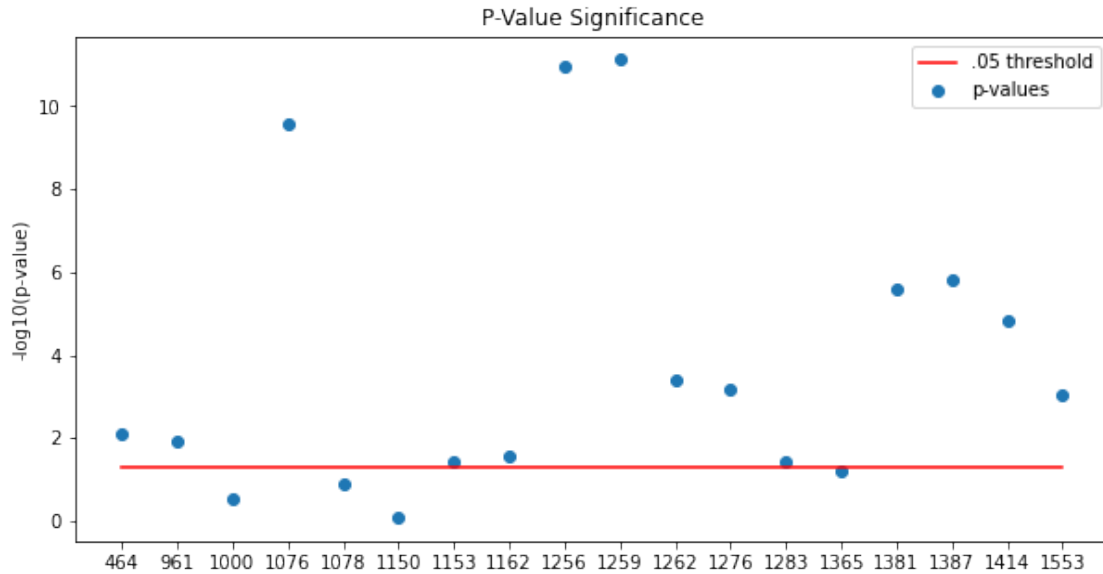
1.2 Bonferroni Correction

The family wise error (FWER) is defined as the probability of yielding one or more false positives out of all hypotheses tested. When the number of hypotheses tested incrases, so does the FWER,

if the significance level is kept constant. In multiple hypotheses testing, this is corrected by the Bonferroni correction.

```
The significance level after the Bonferroni correction with FWER=0.05 is 0.0025
The significance level after the Bonferroni correction with FWER=0.01 is 0.0005
```

2 Dimension-Reduction

Dimension-reduction methods are used to condense high dimensional data down to dimensions which provide the most information. We have implemented the principal component analysis (PCA). It performs a change of basis and the new basis is chosen, such that the i-th principal component is orthogonal to the first i-1 principal components and the direction maximizes the variance of the projected data. We use the Python library sklearn.

2.1 Principal Component Analysis

The principal component analysis (PCA) is one of the methods for dimension-reduction. It performs a change of basis and the new basis is chosen, such that the i-th principal component is orthogonal to the first i-1 principal components and the direction maximizes the variance of the projected data. Instead of considering all the dimensions, we pick the necessary number of principal components.

```
PCA Projections
Projections of data into latent space.
Data is colored by response
```
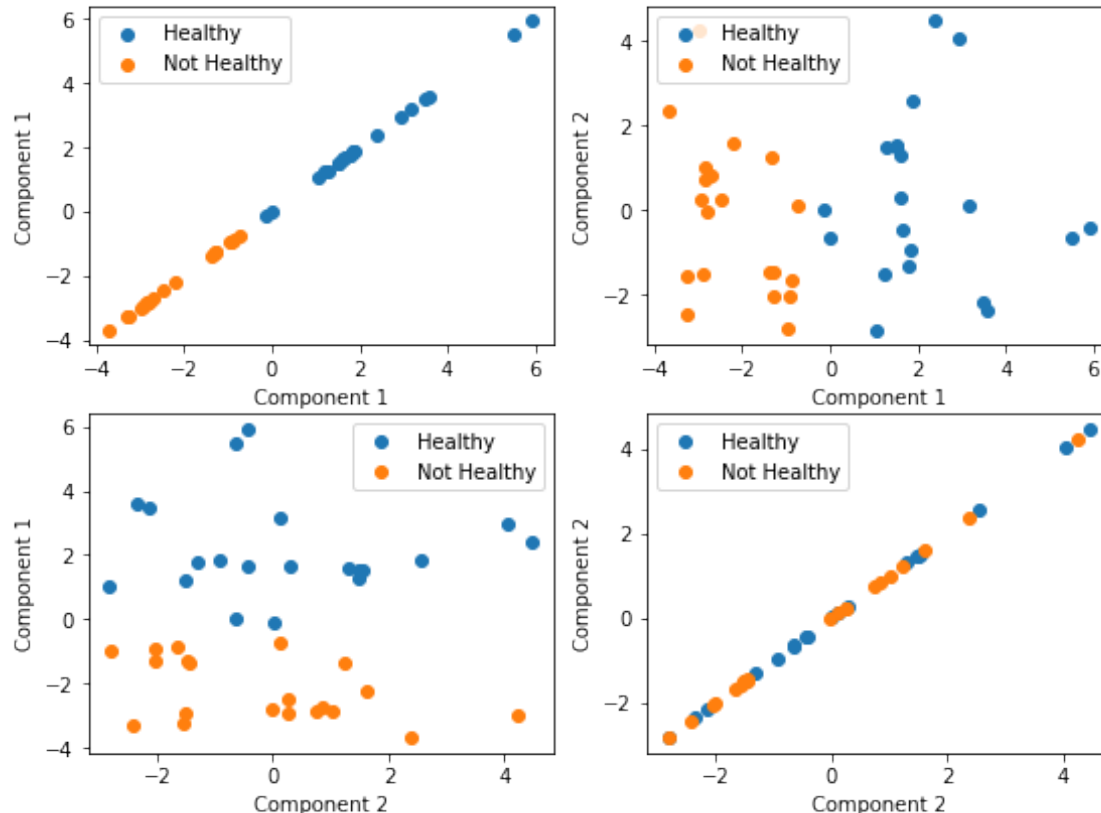
2.2 Linear Discriminant Analysis

Linear discriminant analysis is a classifier with a linear decision boundary. We assume normality and fit conditional densities $p(x \mid y = 0)$ and $p(x \mid y = 1)$ with mean and covariance parameters (μ_0, σ_0) and (μ_1, σ_1), where x, μ_0 and μ_1 are vectors. Dimensionality-reduction is done by projecting the input to the most discriminative directions.

```
LDA Projections
Projections of data into latent space.
Data is colored by response
```
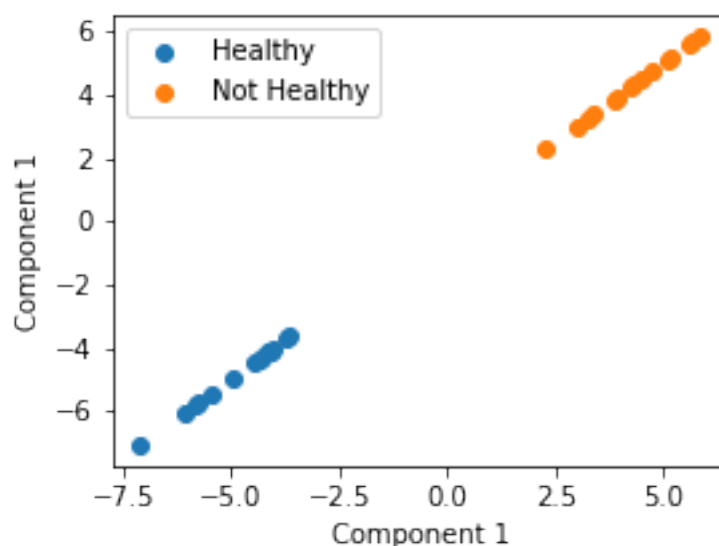

3 Clustering

In this module we use various different clustering methods on spectra. We use the elbow method to find the optimal number of clusters. Clustering is done with scipy and sklearn libraries.

3.1 K-Means Clustering

K-means clustering aims to partition the data into k sets and to minimize the Euclidian within-cluster sum of squares (WCSS). It is solved by either Lloyd's or Elkan's algorithm and we use sklearn module in Python.

```
        Cluster 1          Cluster 2
0     SCLC_86M1_2        NSCLC_A549_1
1     SCLC_86M1_1       NSCLC_H1703_2
2     SCLC_16HV_1       NSCLC_H1703_1
3     SCLC_16HV_2        NSCLC_A549_2
```

```
4       SCLC_DMS79_1     NSCLC_H1437_1
5       SCLC_DMS79_2     NSCLC_H2228_1
6        SCLC_H187_2     NSCLC_H2228_2
7        SCLC_H187_1     NSCLC_H1437_2
8        SCLC_H209_1     NSCLC_H3122_1
9        SCLC_H524_1      NSCLC_H322_2
10       SCLC_H209_2      NSCLC_H322_1
11       SCLC_H524_2      NSCLC_H358_2
12        SCLC_H69_1     NSCLC_H3122_2
13        SCLC_H82_1      NSCLC_H522_1
14        SCLC_H82_2      NSCLC_H522_2
15        SCLC_H69_2  NSCLC_HCC4006_1
16       SCLC_N417_2      NSCLC_H358_1
17       SCLC_N417_1       NSCLC_PC9_1
18     SCLC_SW210-5_1       NSCLC_PC9_2
19     SCLC_SW210_5_2  NSCLC_HCC4006_2
```

3.2 BIRCH Clustering

BIRCH (balance iterative reducing and clustering using hierarchies) is a hierarchical clustering method. The hierarchy is created based on the linear sum and the square sum of data points.

```
          Cluster 1         Cluster 2
0        NSCLC_A549_1       SCLC_86M1_2
1       NSCLC_H1703_2       SCLC_86M1_1
2       NSCLC_H1703_1       SCLC_16HV_1
3        NSCLC_A549_2       SCLC_16HV_2
4       NSCLC_H1437_1      SCLC_DMS79_1
5       NSCLC_H2228_1      SCLC_DMS79_2
6       NSCLC_H2228_2       SCLC_H187_2
7       NSCLC_H1437_2       SCLC_H187_1
8       NSCLC_H3122_1       SCLC_H209_1
9        NSCLC_H322_2       SCLC_H524_1
10       NSCLC_H322_1       SCLC_H209_2
11       NSCLC_H358_2       SCLC_H524_2
12      NSCLC_H3122_2        SCLC_H69_1
13       NSCLC_H522_1        SCLC_H82_1
14       NSCLC_H522_2        SCLC_H82_2
15    NSCLC_HCC4006_1        SCLC_H69_2
16       NSCLC_H358_1       SCLC_N417_2
17        NSCLC_PC9_1       SCLC_N417_1
18        NSCLC_PC9_2    SCLC_SW210-5_1
19    NSCLC_HCC4006_2    SCLC_SW210_5_2
```

3.3 DBSCAN Clustering

DBSCAN is a non-parametric density-based clustering algorithm. It clusters together nearby neighbors, marking further away points as outliers, as they are in the low density area.

```
       Cluster 1       Cluster 2
0        NSCLC_A549_1  NSCLC_H1703_2
1        NSCLC_A549_2  NSCLC_H1703_1
2       NSCLC_H1437_1            NaN
3       NSCLC_H2228_1            NaN
4       NSCLC_H2228_2            NaN
5       NSCLC_H1437_2            NaN
6       NSCLC_H3122_1            NaN
7        NSCLC_H322_2            NaN
8        NSCLC_H322_1            NaN
9        NSCLC_H358_2            NaN
10      NSCLC_H3122_2            NaN
11       NSCLC_H522_1            NaN
12       NSCLC_H522_2            NaN
13     NSCLC_HCC4006_1           NaN
14       NSCLC_H358_1            NaN
15        NSCLC_PC9_1            NaN
16        NSCLC_PC9_2            NaN
17     NSCLC_HCC4006_2           NaN
18         SCLC_86M1_2           NaN
19         SCLC_86M1_1           NaN
20         SCLC_16HV_1           NaN
21         SCLC_16HV_2           NaN
22        SCLC_DMS79_1           NaN
23        SCLC_DMS79_2           NaN
24         SCLC_H187_2           NaN
25         SCLC_H187_1           NaN
26         SCLC_H209_1           NaN
27         SCLC_H524_1           NaN
28         SCLC_H209_2           NaN
29         SCLC_H524_2           NaN
30          SCLC_H69_1           NaN
31          SCLC_H82_1           NaN
32          SCLC_H82_2           NaN
33          SCLC_H69_2           NaN
34         SCLC_N417_2           NaN
35         SCLC_N417_1           NaN
36      SCLC_SW210-5_1           NaN
37      SCLC_SW210_5_2           NaN
```

3.4 Mean Shift Clustering

The mean shift algorithm is a nonparametric clustering technique which does not require prior knowledge of the number of clusters, and does not constrain the shape of the clusters. It works by starting at data points and iteratevely finding the convergence points for kernel estimate gradient.

```
       Cluster 1       Cluster 2
0        NSCLC_A549_1  NSCLC_H1703_2
```

```
1      NSCLC_A549_2   NSCLC_H1703_1
2      NSCLC_H1437_1            NaN
3      NSCLC_H2228_1            NaN
4      NSCLC_H2228_2            NaN
5      NSCLC_H1437_2            NaN
6      NSCLC_H3122_1            NaN
7       NSCLC_H322_2            NaN
8       NSCLC_H322_1            NaN
9       NSCLC_H358_2            NaN
10     NSCLC_H3122_2            NaN
11      NSCLC_H522_1            NaN
12      NSCLC_H522_2            NaN
13    NSCLC_HCC4006_1           NaN
14      NSCLC_H358_1            NaN
15       NSCLC_PC9_1            NaN
16       NSCLC_PC9_2            NaN
17    NSCLC_HCC4006_2           NaN
18       SCLC_86M1_2            NaN
19       SCLC_86M1_1            NaN
20       SCLC_16HV_1            NaN
21       SCLC_16HV_2            NaN
22      SCLC_DMS79_1            NaN
23      SCLC_DMS79_2            NaN
24       SCLC_H187_2            NaN
25       SCLC_H187_1            NaN
26       SCLC_H209_1            NaN
27       SCLC_H524_1            NaN
28       SCLC_H209_2            NaN
29       SCLC_H524_2            NaN
30        SCLC_H69_1            NaN
31        SCLC_H82_1            NaN
32        SCLC_H82_2            NaN
33        SCLC_H69_2            NaN
34       SCLC_N417_2            NaN
35       SCLC_N417_1            NaN
36     SCLC_SW210-5_1           NaN
37     SCLC_SW210_5_2           NaN
```

3.5 Gaussian Mixture Clustering

Gaussian mixture models (GMMs) cluster the data by fitting a mixture of Gaussian models to the data and clustering together data points with similar parameter estimates. It's closely related to k-means clustering but allows for less restrictive cluster shapes. K-means fits a multi-dimensional ball as the perimeter, but GMMs can also fit ellipsoidal shapes and other shapes.
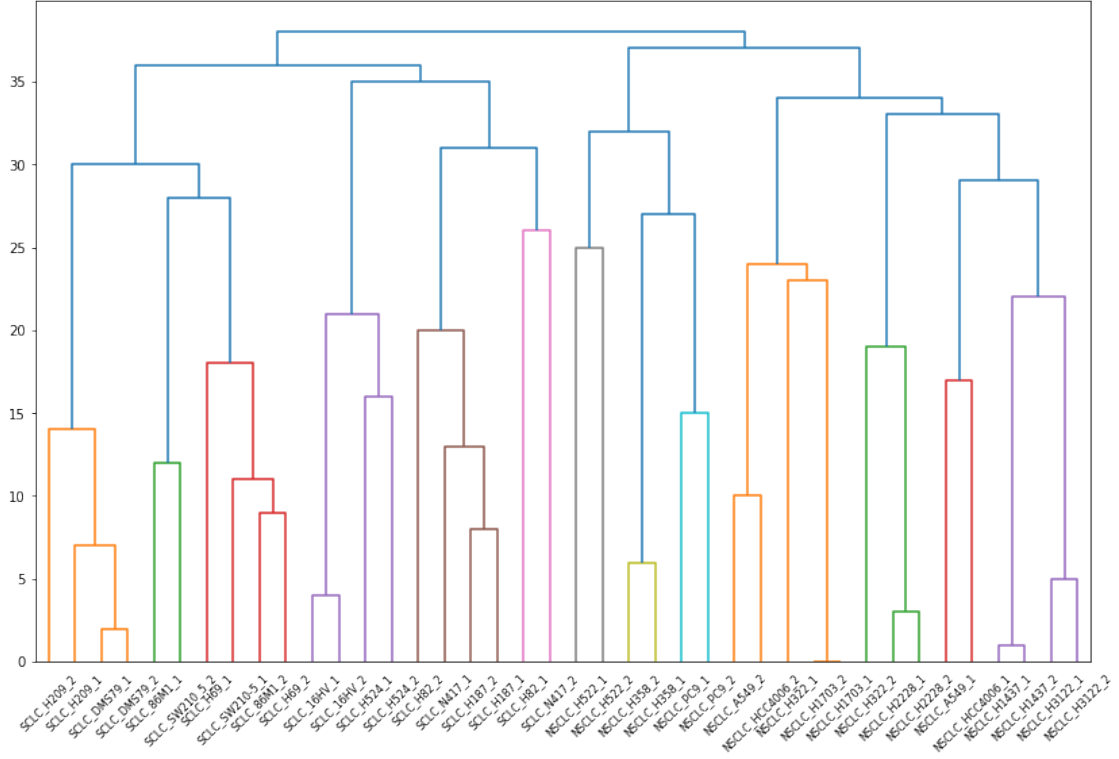
```
        Cluster 1          Cluster 2
0     NSCLC_H522_1       NSCLC_A549_1
1     NSCLC_H522_2      NSCLC_H1703_2
```

```
2        SCLC_86M1_2      NSCLC_H1703_1
3        SCLC_86M1_1       NSCLC_A549_2
4        SCLC_16HV_1      NSCLC_H1437_1
5        SCLC_16HV_2      NSCLC_H2228_1
6       SCLC_DMS79_1      NSCLC_H2228_2
7       SCLC_DMS79_2      NSCLC_H1437_2
8        SCLC_H187_2      NSCLC_H3122_1
9        SCLC_H187_1       NSCLC_H322_2
10       SCLC_H209_1       NSCLC_H322_1
11       SCLC_H524_1       NSCLC_H358_2
12       SCLC_H209_2      NSCLC_H3122_2
13       SCLC_H524_2    NSCLC_HCC4006_1
14        SCLC_H69_1       NSCLC_H358_1
15        SCLC_H82_1        NSCLC_PC9_1
16        SCLC_H82_2        NSCLC_PC9_2
17        SCLC_H69_2    NSCLC_HCC4006_2
18       SCLC_N417_2                NaN
19       SCLC_N417_1                NaN
20    SCLC_SW210-5_1                NaN
21    SCLC_SW210_5_2                NaN
```

3.6 Hierarchical Clustering

Hierarchical clustering builds hierarchies of clusters based on a chosen metric and a linkage scheme. We used cosine distance and average linkage scheme.

```
          Cluster 1         Cluster 2
0       NSCLC_A549_1       SCLC_86M1_2
1      NSCLC_H1703_2       SCLC_86M1_1
2      NSCLC_H1703_1       SCLC_16HV_1
3       NSCLC_A549_2       SCLC_16HV_2
4      NSCLC_H1437_1      SCLC_DMS79_1
5      NSCLC_H2228_1      SCLC_DMS79_2
6      NSCLC_H2228_2       SCLC_H187_2
7      NSCLC_H1437_2       SCLC_H187_1
8      NSCLC_H3122_1       SCLC_H209_1
9       NSCLC_H322_2       SCLC_H524_1
10      NSCLC_H322_1       SCLC_H209_2
11      NSCLC_H358_2       SCLC_H524_2
12      NSCLC_H3122_2        SCLC_H69_1
13      NSCLC_H522_1        SCLC_H82_1
14      NSCLC_H522_2        SCLC_H82_2
15    NSCLC_HCC4006_1        SCLC_H69_2
16      NSCLC_H358_1       SCLC_N417_2
17       NSCLC_PC9_1       SCLC_N417_1
18       NSCLC_PC9_2    SCLC_SW210-5_1
19    NSCLC_HCC4006_2    SCLC_SW210_5_2
```
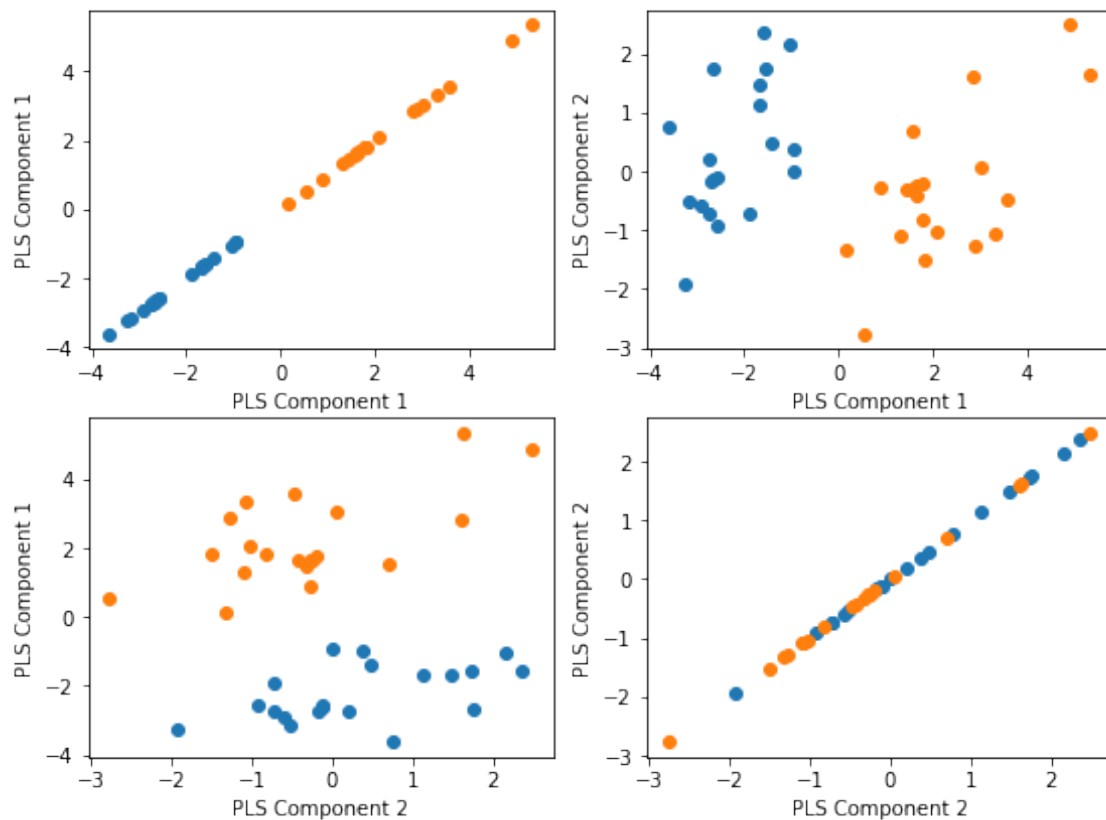
4 Classification

Classification methods aim to classify the response of samples. The given data is separated into a training set and a testing set. The model parameters are found from the training set and the testing set is used to quantify the model accuracy. The methods are from sklearn package.

4.1 Partial Least Squares-Discriminant Analysis

4.2 Support Vector Machines

Classification via SVM is done by fitting a linear plane to the latent space but only considering a subset of inputs in the fitting process. The quantity R^2 measures what percentage of variation was explained by the model in the training set. The quantity Q^2 shows the same measurement but for the test data set.

```
SVM Validated Parameters:  {'kernel': 'linear', 'shrinking': True}
SVM: R^2=1.0 Q^2=1.0
```

4.3 Random Forest

Random forests is an ensemble classification method. It works by constructing multiple decision trees based on the training data and then choosing the class, chosen by the most number of decision trees. The quantity R^2 measures what percentage of variation was explained by the model in the training set. The quantity Q^2 shows the same measurement but for the test data set.

```
Random Forest Validated Parameters:  {'criterion': 'gini', 'n_estimators': 50}
RF: R^2=1.0 Q^2=1.0
```

4.4 Logistic Regression

Logistic regression uses a logistic function to model a binary dependent variable. The confusion matrix displays the accuracy of the model for the test data set. We use the packages sklearn for the logistic regression and seaborn for the confusion matrix.

```
Accuracy: 1.0
<modules.adapml_classification.Classification object at 0x7f8b48046d60>
```
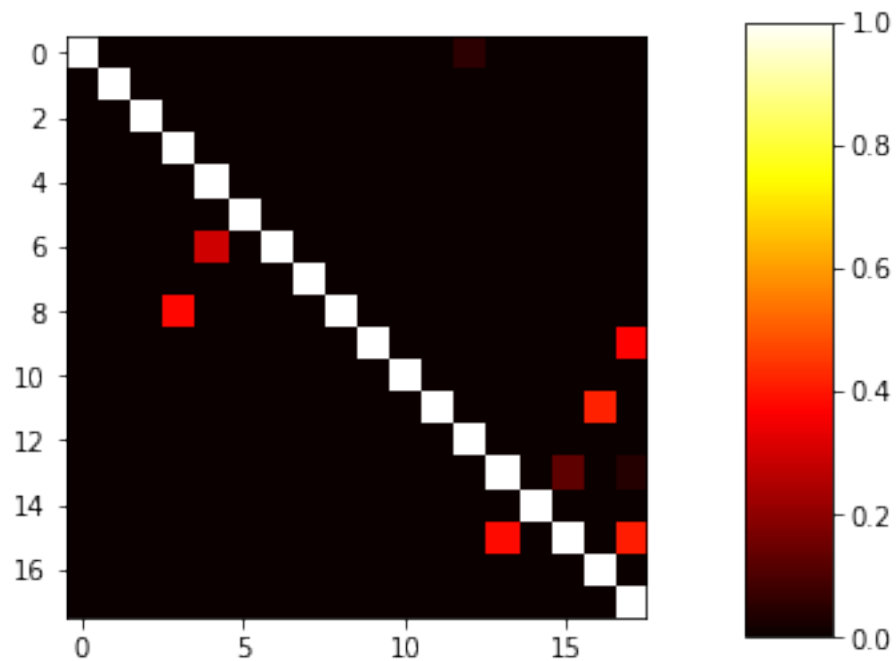

Confusion matrix
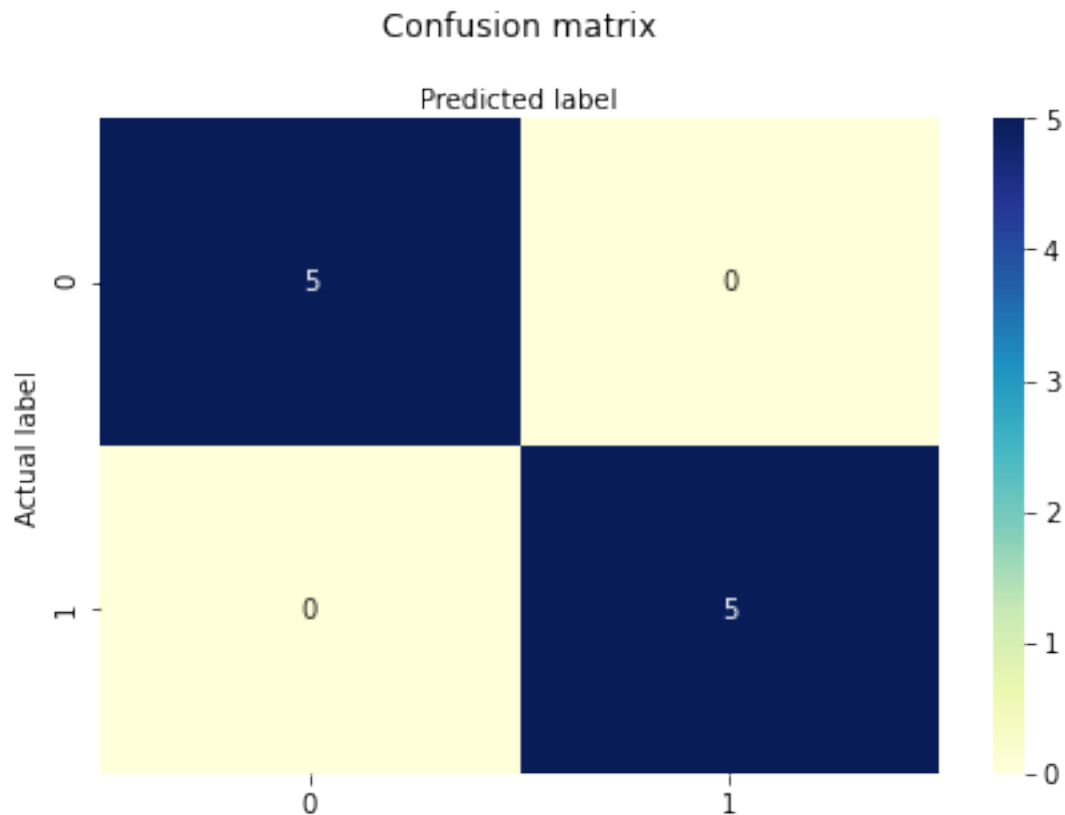
5   Regression

5.1   Linear Regression

Linear regression fits a linear plane between the dependant variables and the response. The linear plane models the relationship between them and allows for prediction or explain variation.

R2 score between NSCLC_A549_2 and NSCLC_H3122_1 is 0.37739250415156134
R2 score between NSCLC_H1437_1 and NSCLC_H2228_2 is 0.29329760536722105
R2 score between NSCLC_H522_1 and NSCLC_HCC4006_1 is 0.38171401707539343
R2 score between NSCLC_H358_1 and NSCLC_H358_2 is 0.4156127581178537
R2 score between NSCLC_PC9_1 and NSCLC_H322_2 is 0.3677940179661904
R2 score between NSCLC_PC9_1 and NSCLC_HCC4006_1 is 0.40681273290339537