rTLC manual

Version: 23/11/2015

Table of Contents

Introduction

rTLC is an application made to facilitate the exploitation of HPTLC pictures and especially statistics exploitation.
Different features are available:

- Chromatograms extraction from pictures
- Chromatograms preprocessing
- Variables selection
- Exploratory statistics
    - PCA
    - Cluster
    - Heatmap
- Predictive statistics
    - Model training
        - Parameters tunning
        - Cross validation
        - Regression/classification
    - New data prediction
- Report Output

The application could be found at this url:
http://shinyapps.ernaehrung.uni-giessen.de/rtlc/

Analytical Pipeline

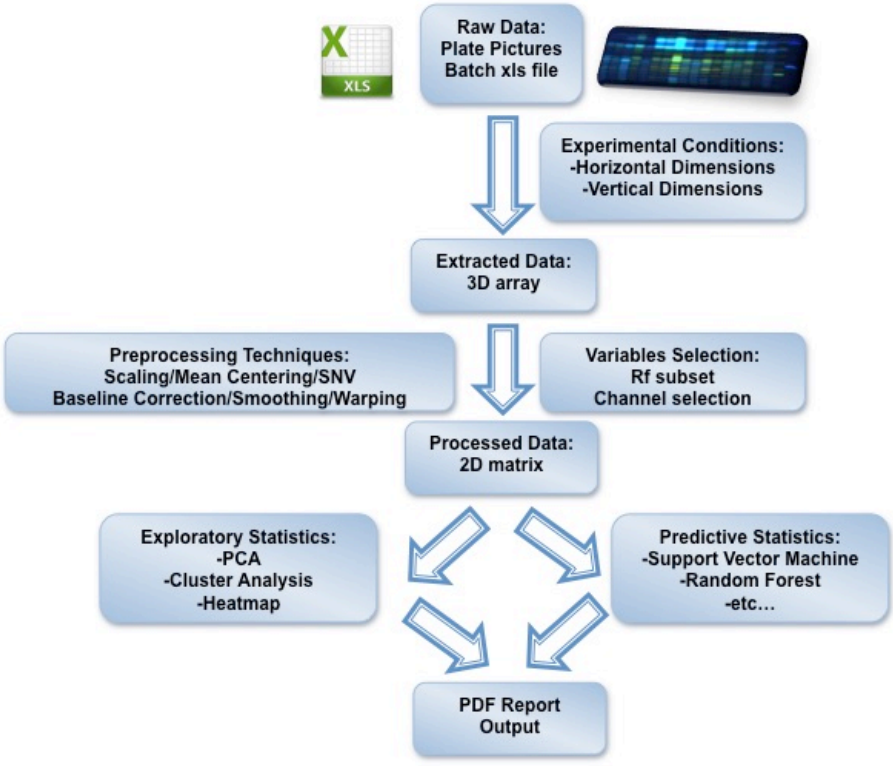

Figure 1. Analytical Pipeline

Data Input

Chromatogram extraction

Demonstration Data

In the tab Data Input, select one of the demo files in the *data to use* menu on the left (Figure 3-1).
A picture should appear on the page (Figure 3-2), as well as a *Plate choice* menu (Figure 3-3) and a table named *Horizontal Dimension* (Figure 3-4).

Horizontal Dimensions

A chromatogram will be extracted between each pair of red and green vertical lines on the central picture by taking the horizontal mean of pixels on each of the red, green and blue channels of the picture (Figure 2 and Figure 3-2).

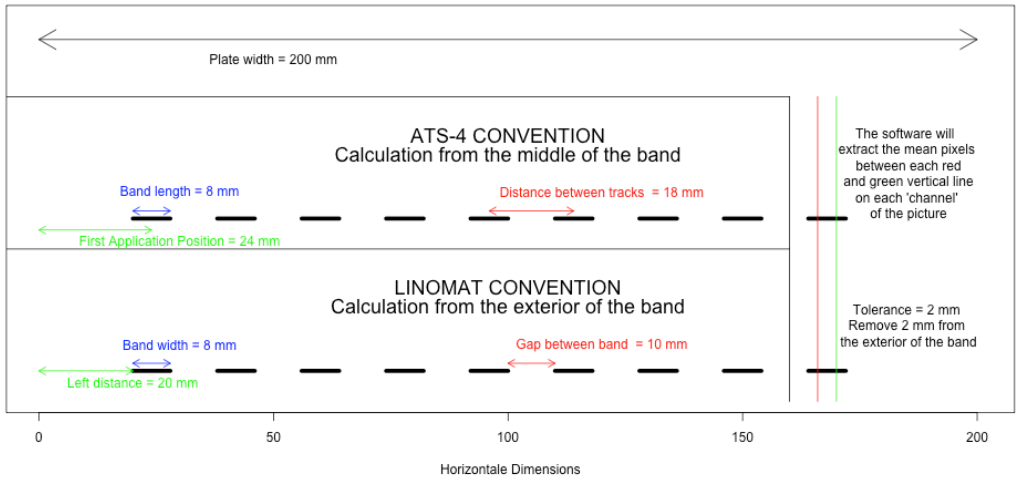

Figure 2. Illustration of the chromatograms extraction

You must modify the number in the Horizontal Dimension table in order to match each band of your picture between a pair of red and green lines.
If the dimensions are available from the manipulation AND there wasn't **unnecessary cropping** of the pictures, this step should be straightforward (Be careful with the frame option in the default setting of winCats, the default option is to remove 2 mm of the plate so you must adapt the dimension or put this option to 0 mm).
It is possible to choose 2 conventions for those dimensions: *LINOMAT* or *ATS*, depending if the dimensions are considered from the outside of the band (LINOMAT) or from the center (ATS) (Figure 3-5).
The *tolerance* dimension is here to control the zone of the band you want to extract, a value of 0 will extract all the band, whereas a bigger value will help to take only the center of the band.
This operation must be done for each plate of the study, the picture could be chosen in the *Plate choice* menu on top of the picture (Figure 3-3). If your study contains 3 plates, there will be 3 choices in the drop-box menu and therefore, 3 rows in the *Dimension Table.*
It's possible to save a dimension table as an excel file to use later, for example with the same study but with pictures under a different light.

Vertical Dimensions

The vertical dimension table is here to calculate the Retention Front value and redimension the number of pixels of the picture and therefore the number of points in the chromatogram (Figure 3-6).

Batch Table

Visit the batch table to visualize the batch (Figure 3-7).
You can edit this table and choose to remove some samples from the study (standard or outlier for example).

Chromatograms/Band Comparison/Chromatograms Comparison

In these three tabs, you can visualize the chromatograms extracted from the plates (Figure 3-8).

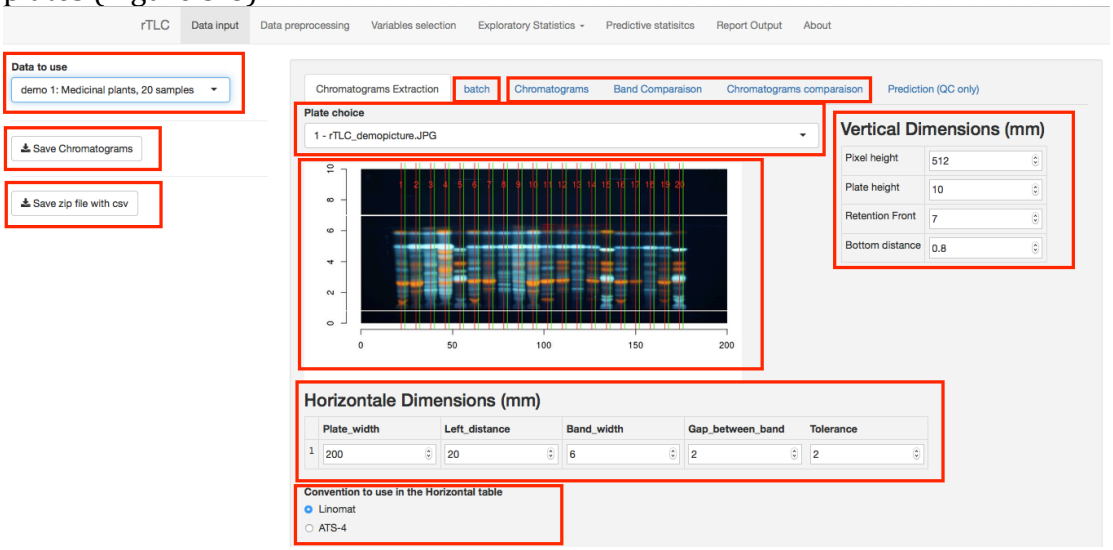

Figure 3: Data Input. Demo file

Your Own Data

Now in the tab Data Input, choose to use *Your Own Data* (Figure 4-1).
There are two parts:

- the independent variables: plate pictures with the band
- the dependent variables: batch file (in excel) with information on each band

You can upload your(s) plate(s) in the *Browse* that appears on the left (Figure 4-2).
Proceed to the extraction like for the demonstration data.
For the batch, there are two choices, it's possible to upload an excel file on the left side of the page or it's possible to edit directly the batch file in the *batch* tab, the number of rows will correspond to the number of extracted chromatograms (Figure 4-3).
In case a excel file is uploaded, it must contain the information in the first sheet and have this kind of format:

Id	Colname 1	Colname 2	Colname 3	…
1	…	…	…	…
2	…	…	…	…
3	…	…	…	…
…	…	…	…	…

The first row must be the name of the columns with at least 4 columns, the first column will be the id column used for the application to track the samples. There must be the same number of rows (without the first one) as chromatograms extracted.
In case one of the constraints is not respected, a green message will appear showing the user what is the problem.
This kind of format is called tidy data in the field of Data sciences and it's a good practice to adopt this kind of format for data collection: 1 observation is the association of independent variables and dependent variables, therefore, taking tracks of the data in a consistent way from the beginning of a study (or even multiple study) can save a lot of time later in the phase of data analysis.

Save the data extracted

In order to avoid the step of chromatogram extraction for a future session, it's possible to save a file containing the chromatograms and the batch table with the *Save Chromatograms* button on the left of the page (Figure 4-4).
In another session, choose to use *Saved Data* in the tab *Data Input*. And upload the file saved precedently in the *browse* button (Figure 5).

Save csv file for each channel

To export the chromatograms to another software for further exploitation, it's possible to save each channel as a CSV file with observation as row and R_F as column. The files use ";" as separator. The download buttons are on the left part of the page (Figure 4-5).

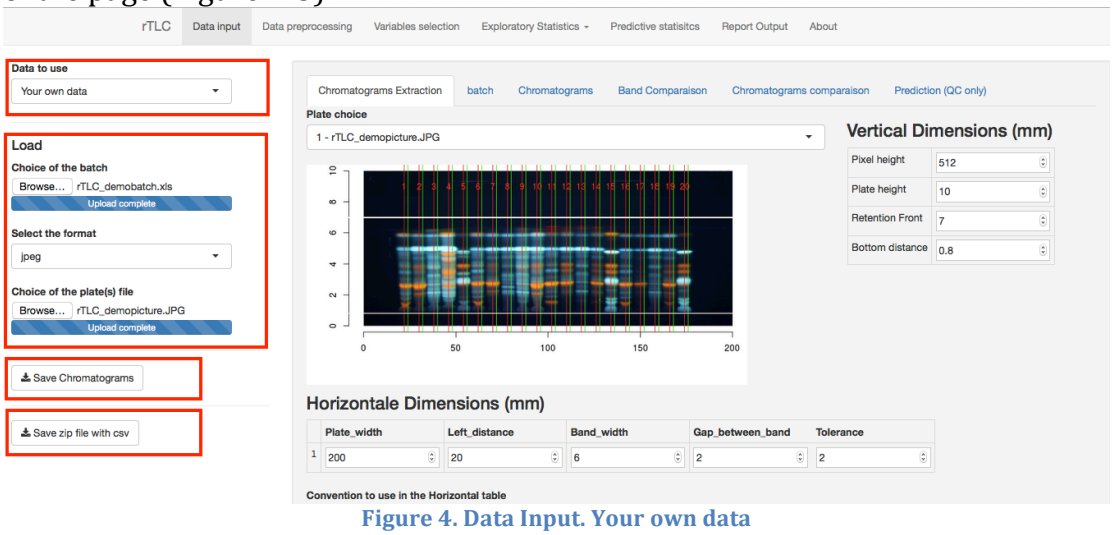

Figure 4. Data Input. Your own data

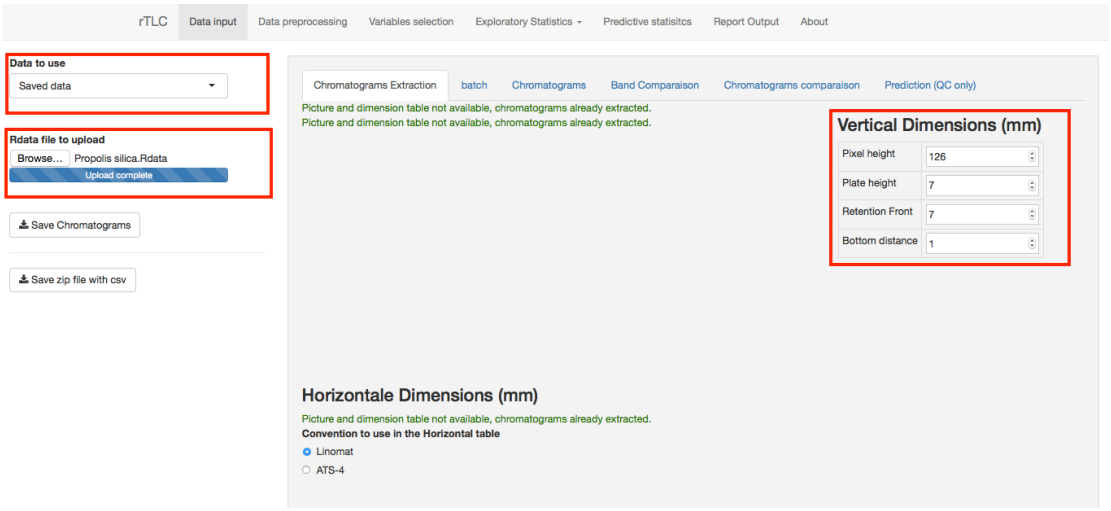

Figure 5: Data Input. Saved Data

Data preprocessing

This tab allows different preprocessing in order to prepare the data for further analysis.

Preprocess Order

In the left side of the page, choose the order the preprocessing should appear (Figure 6-1). Available preprocessing are:

- Smoothing: Savitzky-Golay transformation
- Warping: Peak alignment (experimental)
- Baseline correction
- Scaling
- Standard Normal Variate
- Mean centering

Preprocess Details

For each preprocessing, a set of options are available, in each case, a link leads to an exhaustive explanation of the features (Figure 6-2).

Chromatograms/ Chromatograms Comparison

In these two tabs, you can visualize the results of the preprocessing (Figure 6-3).

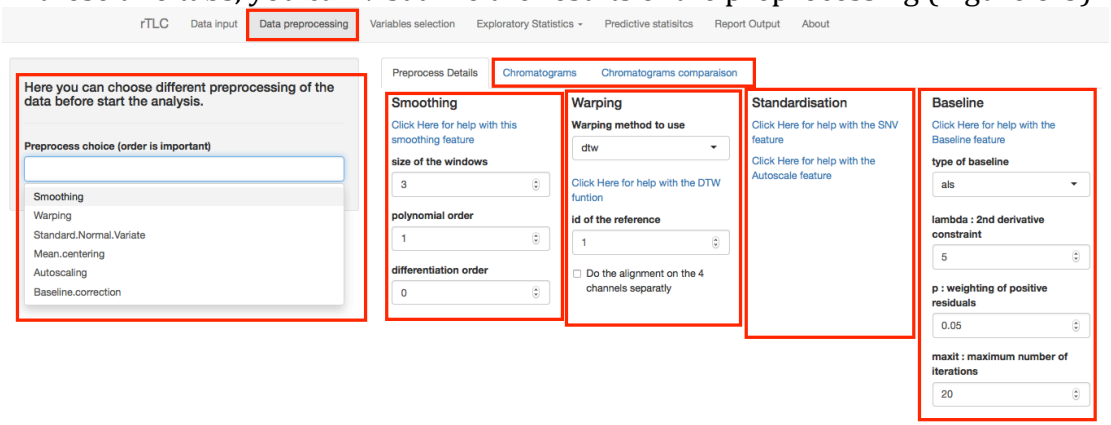

Figure 6. Data Preprocessing

Variables Selection

This tab allows for variable selection in order to choose a channel or part of a channel. There are 12 possibilities to choose a channel, a range and to include or not this range in the study (Figure 7-1). After this step, all selected data are combined into one data set that will be used for statistical study. The two plots on the left should help the user to understand the feature (Figure 7-2).

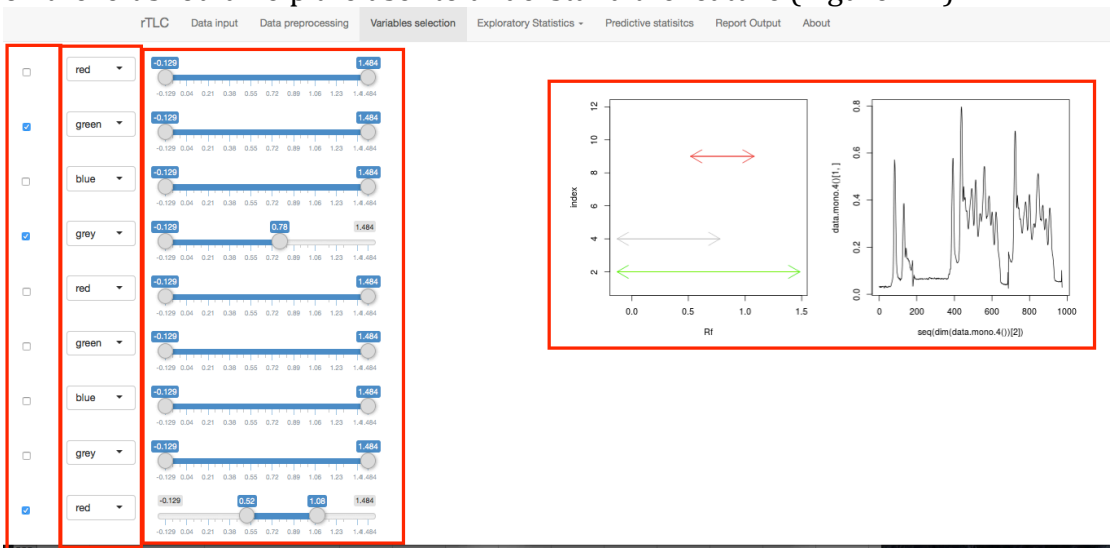

Figure 7. Variable Selection

Exploratory Statistics

PCA

This feature allows to perform Principal Component Analysis on the dataset.

PCA tab

The principal plot is the score plot, a few options are available:
- choose the color/shape/label of the point according to one of the variable of the batch (Figure 8-1)
- choose the component to plot (Figure 8-2)
- choose to calculate the ellipse and to plot it (Figure 8-3)
- a few aesthetics parameters (Figure 8-4):
    - hjust and vjust move the label on the plot
    - point size scales the point size on the plot
- title of the plot (Figure 8-5)

On this page, there is also a table of the batch and the first 4 components of the analysis and a Summary of the model which shows the cumulative variables of the first 5 and the 10th components.

Loading Plot

This tab shows the loading plot of the PCA (Figure 8-6).
It's possible to choose the component to study and to plot or not the minimum and maximum point on the graph according to a *neighborhood* value. The resulting maximum and minimum values are shown in the field bellow.

Outlier

This tab is for outlier detection, i.e. points that should be removed because they are too different from the dataset.
It's possible to choose the number of components of the PCA to include in the test and the quantile to use for the cutoff. The Mahalanobis distance is used and the classical and robust tests are calculated (Figure 8-7).

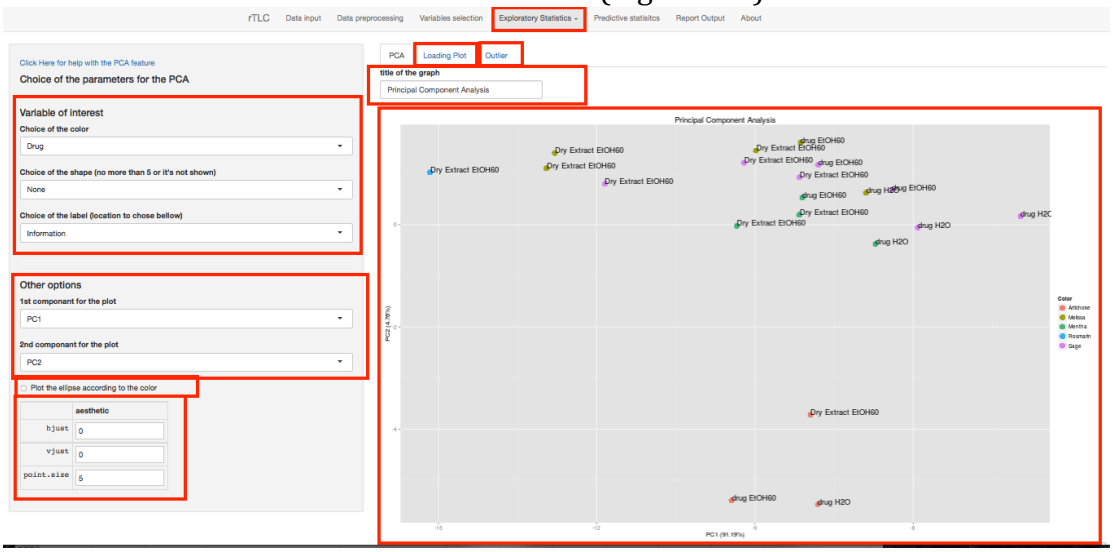

Figure 8. Exploratory statistics. PCA

Cluster

This feature allows to perform cluster analysis on the dataset.
The options available are:
- Choice of the variable of interest in the batch (Figure 9-1).
- Choice of the clustering options (Figure 9-2):
    - Choice of the distance method (Euclidean is the more used).
    - Choice of the cluster method.
    - Number of clusters to cut the tree in.

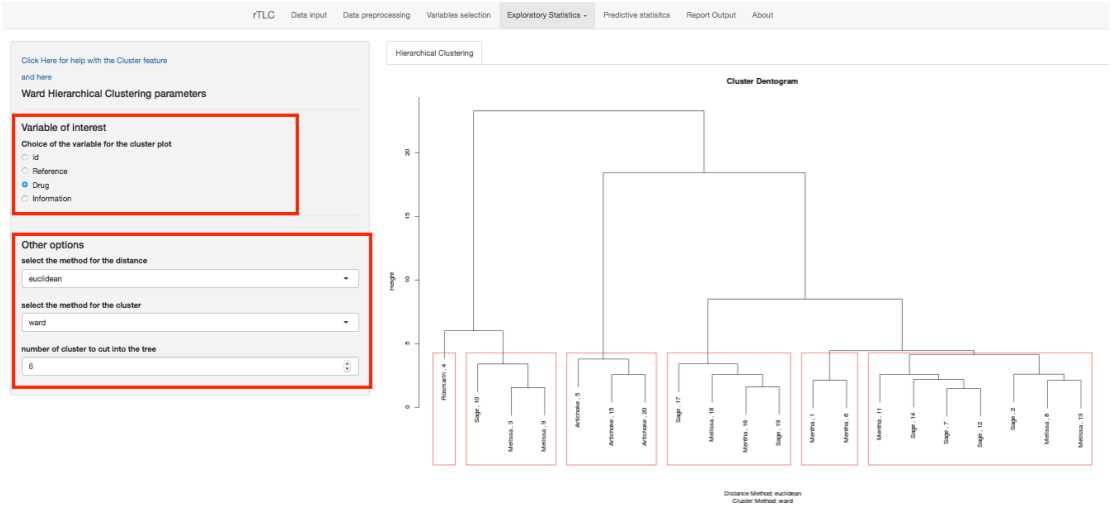

Figure 9. Exploratory Statistics. Cluster

Heatmap

This feature allows to perform and visualize the heatmap, choose the variable of interest and visualize the result, either with the normal heatmap, or with the interactive heatmap.

Predictive Statistics

This tab allows you to train a predictive model for classification or regression (Figure 10).

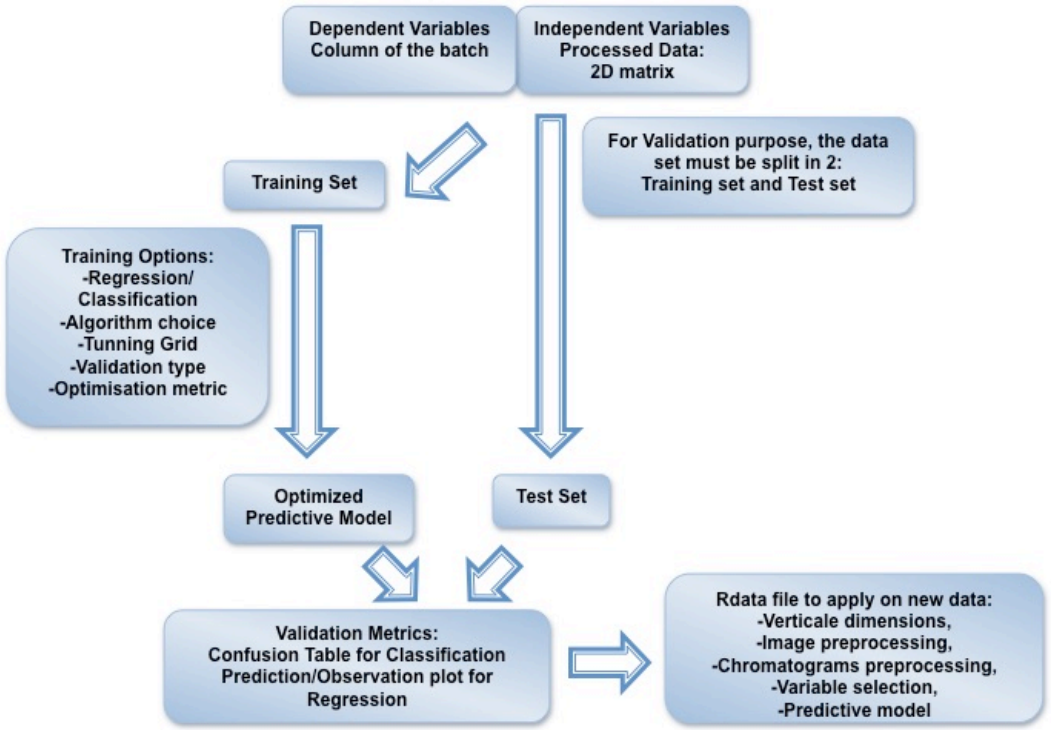

Figure 10. Pipeline for predictive statistics

Options

Training/Test split

In a first time, the data set should be split in two, the test set and the training set. The training set will be used to train the data and the test set will be used to verify the result of the training on an independent part of the dataset (Figure 11-1).

Classification/Regression

Depending on the problem, one option should be chosen in order to train the system on the good type of data.
(Figure 11-2).

Choice of the variable of interest

Choose the variable to be trained with from the batch. What should be predicted. It must be in accordance with the Classification/Regression choices, otherwise an error will be returned, for example if regression is asked on non-numeric data (Figure 11-3).

Algorythm

Choose which machine learning algorithm should be used, some of them are only available for classification or for regression (Figure 11-4).

Only a subset of available algorithms is available, others could be added, just contact us. The list of all models available could be found here:
http://topepo.github.io/caret/modelList.html

Tuning Options

The training will try every combination of every parameters of the grid in order to optimize the performance of the model and choose the better parameters.

Validation

- Validation method (Figure 11-5):
    - bootstrap
    - repeated cross validation
    - leave one out cross validation
- Summary metrics (Figure 11-6):
    Which summary metrics to use for the tuning
- Cross validation k-fold or resampling iterations (Figure 11-7):
    Number of k-fold or resampling
- Number to repeat (k-fold only) (Figure 11-8):
    Number of times to repeat the validation process

Grid

This area contains the tuning length, i.e. the maximum number of parameters to test on each parameters. It is also possible to choose the different parameters manually in the Grid table for fine tuning (Figure 11-9).

Launch

Once all the options are chosen, press the *Train* button to launch the analysis, note that you must visit another tab to really launch the analysis (Figure 11-10).

Results

Validation Metrics

This tab is used to verify the performance of the model (Figure 11-11), a confusion matrix is shown for the classification problem and a plot of predicted values against the real value is shown for the regression problem.

It's possible to choose to visualize the result for the Test data, the Training data and the cross-validation data, i.e. the data used during the optimization phase of the training.

Prediction table

This tab shows the prediction table for all data, it's possible to filter according to the use in the training set or not, to the prediction class etc… (Figure 11-12)

Algorithm information

This tab gives more information about the algorithm used during the training, in particular, what are the tuning parameters (Figure 11-13).

Model Summary

This tab summarizes important information of the tuning, it's possible to extract the information for each row of the tuning grid and for each of the metrics.
Also important information describes how the tuning took place (Figure 11-14).

Tuning Curve

This tab shows the evolution of the metric chosen for the tuning depending on each parameter of the algorithm (Figure 11-15).

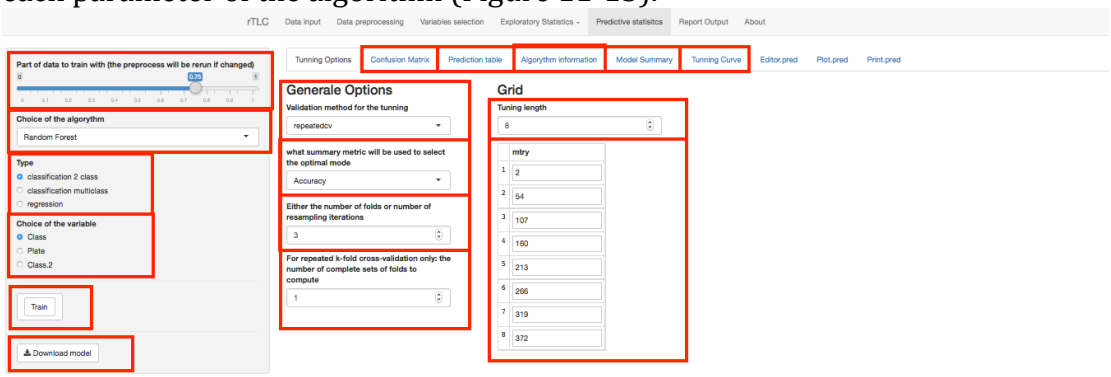

Figure 11. Predictive Statistics

Model Download and New data prediction

Once the good model with the good preprocessing, the good variable selection, the good tuning parameters is made. It's possible to download a file that could be then uploaded at the beginning of the process (Figure 11-15).
In the first tab Data Input, choose to use *Predicted data – QC* (Figure 12-1).
Upload the batch and picture file as previously and also a model file created in another session (Figure 12-2).
Proceed to the chromatograms extraction with the dimension table and visit the tab *Prediction (QC only).* The prediction for each chromatogram should appear (Figure 12-3).

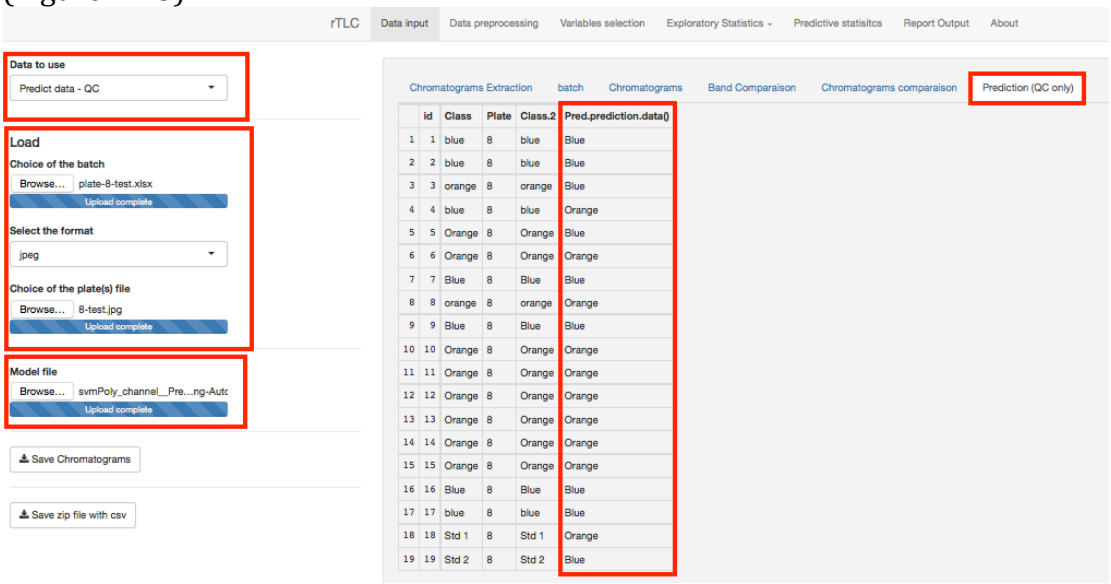

Figure 12. Predict new data

Report output

In this tab, it's possible to download a pdf report, choose the content of this document and click the *Download the report* button (Figure 13).

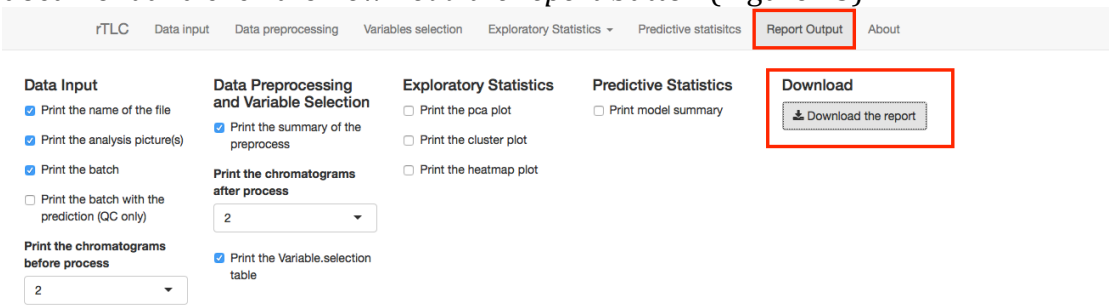

Figure 13. Report Ouput