# WebSpecmine

## USER GUIDE

# Contents

# Tutorial

# 1. Introduction

The website consists in providing means of analysing metabolomics data, as well as allowing the sharing of metabolomics experimental data between users. The name chosen for the website is based on the name of the core package *specmine* where functionalities are implemented: *WebSpecmine*.

## 1.1 Website architecture

The website starts with a home page, where users can enter their user account or do the analysis without logging in, although some features will not be available in the last scenario. These features would mainly consist on saving, into the account, experimental data, so it can be used later, reports and the current work (named workspace, and consists on data and results), that the user could later return to and continue the analysis being made.

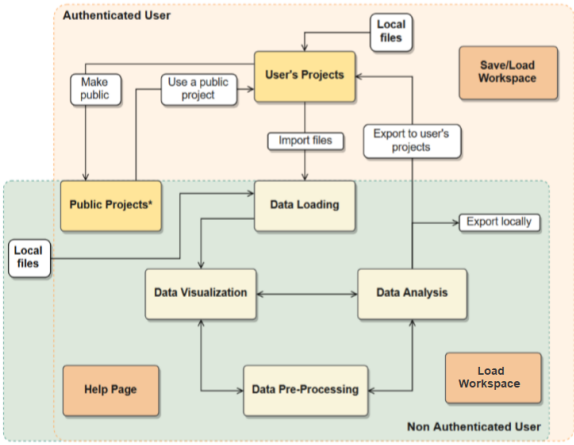



Figure 1.1: Graphical representation of the application's structure, portraying what is accessible for both non authenticated and authenticated users (green rectangle) or only for the latter (yellow rectangle). *Non authenticated users can only view the information contained within the Public Projects page, unless a workspace is associated with that project is available.

## 1.2 Website Layout

The overall appearance of the wesbite is the following:

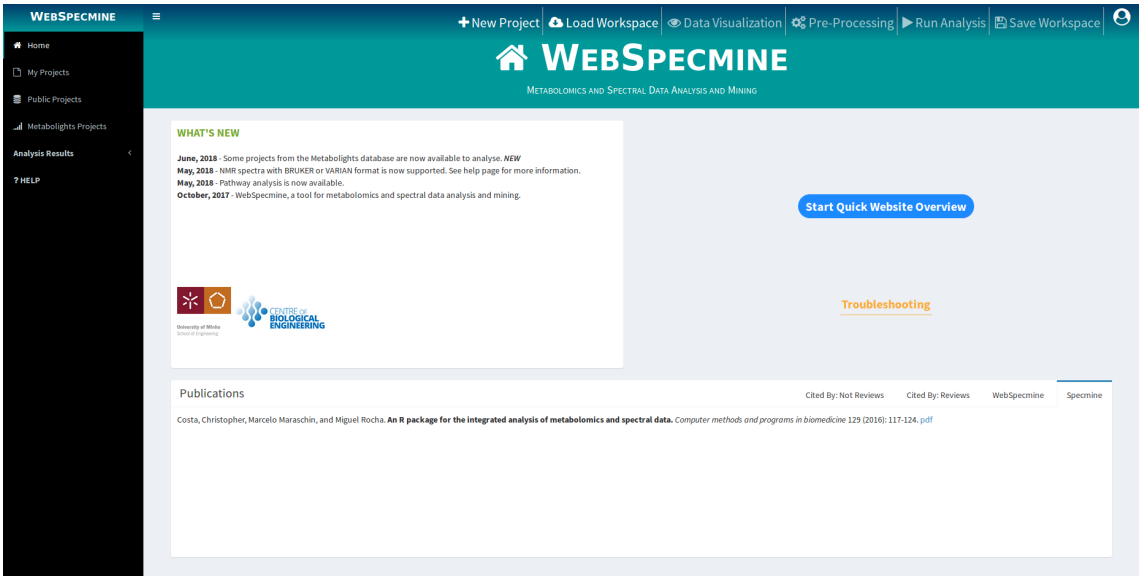



Figure 1.2: Layout of the WebSpecmine Analysis App.

The **header panel** contains the website name and a button that allows to show and hide the sidebar. Furthermore, in this panel, it can be seen the links to pages or pop-up windows that carry out actions:

- *New Project* or *Choose Project*: Load data files for analysis;
- *Load Workspace*: Load data and results previously saved on the website for analysis;
- *Data Visualization*;
- *Pre-Processing*;
- *Run Analysis*;
- *Save Workspace*: Save data and results into the user's account;
- *Account Authentication* icon: Handle the authentication of the user and his account options.

The **sidebar panel** has five tabs, which lead to pages that show the respective information:

- *Home*;
- *My Projects*;
- *Public Projects*;
- *Analysis Results*;
- *Help*.

# 2. Website Functionalities

## 2.1 Supported Data

Various types of data are supported, in many formats. The website considers that each **data** file represents **one** distinct sample, with exception for when one csv file of UV-VIS, IR and Raman Spectra is given and for the data file of concentrations data.

### 2.1.1 NMR and GC/LC-MS Peak Lists

The peak lists data files must have the CSV format. Each CSV file must represent a sample and have two columns: the first one corresponds to the chemical shifts (in ppms) or the mass/charge ratios and the second one the intensities of those peaks. Part of a CSV file of a peak list:

```
ppm,intensity
0.74,0.0001
0.89,0.0004
0.90,0.0007
0.91,0.0005
0.91,0.0008
0.92,0.0004
0.94,0.0003
0.95,0.0004
0.96,0.0009
```

### 2.1.2 MS Spectra

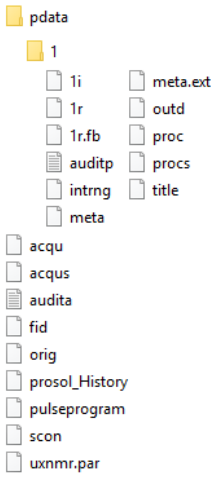
The MS spectral data files must either have .mzXML, .netCDF or mzData formats.

### 2.1.3 NMR Spectra

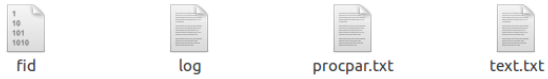
There are two NMR spectra formats that are supported.

The *BRUKER* format is supported, if the processed spectra are given. Each spectrum data has to be in a different folder. Each folder has to have the following structure:

*At least the files procs and 1r have to be present. They have to be inside spectrumfolder-name/pdata/1.*




The *VARIAN* format is supported, only if the raw fid file is given, alongside with the procpar file. Each spectrum data has to be in a different folder. Each folder has to have the following structure:




### 2.1.4 UV-VIS, IR and Raman Spectra

The data files of these type of spectra must be in one of the following formats: CSV, (J)DX, SPC or MS EXCEL (.xlsx).

For data in MS EXCEL or CSV files, each file must have two columns: the first one representing the wavenumber, wavelength or raman shift, according to the type of spectra, and the second one the value of the measurements.

When only one CSV file is given, the structure as to be similar to the following example (the first column corresponds to the wavenumber, wavelength or raman shift, according to the type of spectra):

```
,sampleName1,sampleName2
200,0.085956648,0.04830468
201,0.067182627,0.017316359
202,0.044842223,0.026930633
203,0.051335963,0.041539431
```

### 2.1.5 Concentrations Data

Concentrations data must be a CSV or TSV file with the samples names in the first column (each line then corresponds to a sample) and the concentrations values for each metabolite in the following columns. Alternatively, samples names can be in the first line (each column then corresponds to a sample) and the concentrations values for each metabolite in the following lines.

Part of a CSV example file of concentrations file:

```
Patient ID,1.6-Anhydro-beta-D-glucose,1-Methylnicotinamide,2-Aminobutyrate
PIF_178,40.85,65.37,18.73
PIF_087,62.18,340.36,24.29
```

```
PIF_090,270.43,64.72,12.18
NETL_005_V1,154.47,52.98,172.43
PIF_115,22.2,73.7,15.64
```

### 2.1.6 Metadata File

As regards to the metadata file, it can either have CSV or TSV format. Each line should correspond to a sample, where the first column represents the names of such samples, and the remaining ones the metadata classes.

The first column corresponds to the names of the samples. **For the cases where more than one data file is given, the names of the samples have to correspond to the names of the data files.**

Here you have an example of a metadata file:

```
Sample Name,Seasons
July2010,Winter
September2010,Spring
October2010,Spring
November2010,Spring
February2011,Sum/Aut
March2011,Sum/Aut
April2011,Sum/Aut
may2011,Sum/Aut
June2011,Winter
July2011,Winter
August2011,Winter
September2011,Spring
October2011,Spring
```

## 2.2 Projects

### 2.2.1 What is a project?

A project consists on a study, or group of studies, and contains the data and metadata used, as well as reports that were obtained throughout the analysis of such data.

The projects are saved in the user's account and can be stored as private, so that only the user can see them and analyse them, or made public, where everyone that accesses the website is able to see the project, without making any changes on it. However, logged in users can copy other users' public projects to their own account and then analyze it and save changes, as it won't compromise the original project.

### 2.2.2 The structure of a project

Each project is organized in different types of folders, as follows:

- ***Data***: stores data folders, with each one of them with data files that are used in an analysis;
- ***Metadata***: stores metadata files, where each one can be used in an analysis;
- ***Reports***: stores the reports generated by the analysis of a certain data from the corresponding project;

### 2.2.3 My Projects

This page is accessed through the sidebar panel and it is only accessible for logged in users, as it is the page that contains the information on the projects that were stored in the user's account.

When you firstly access this page, only a box at the left side of the page appears, named "List of Projects", with the list of projects that you have on the account.

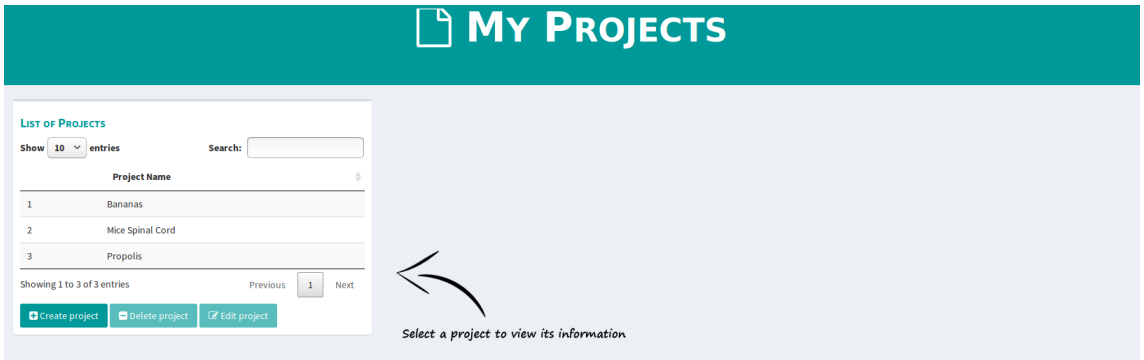



Figure 2.1: Layout of the "My Projects" page when the user enters it for the first time.

When you select a project by clicking on its name it the table with the list of projects, you will be able to see its information at the right and all the tasks available to perform will be done on the selected project.

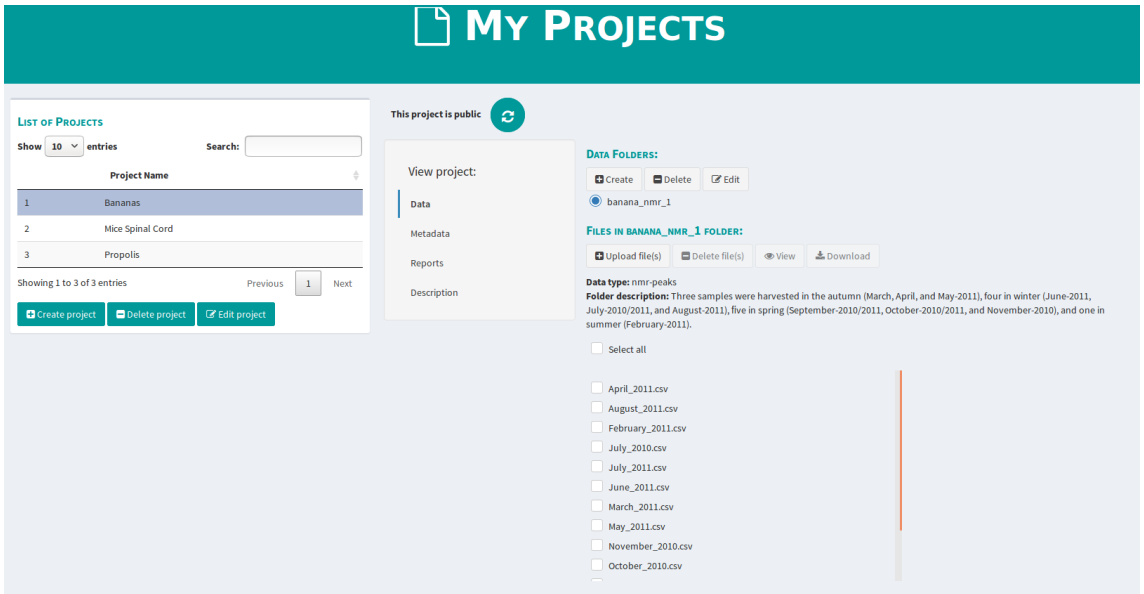
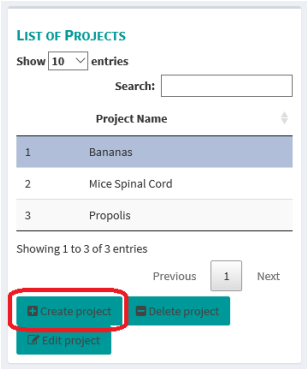



Figure 2.2: Layout of the "My Projects" page when the user selects a project.

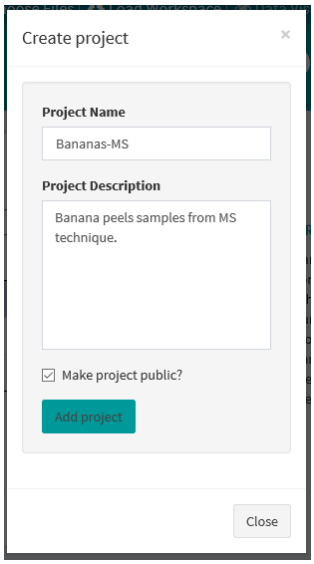
Following, are the different tasks that can be performed in this page, regarding creating and editing the projects.

### Create a project

To create a new project, you have to click the button "Create project" in the "List of Projects" box:



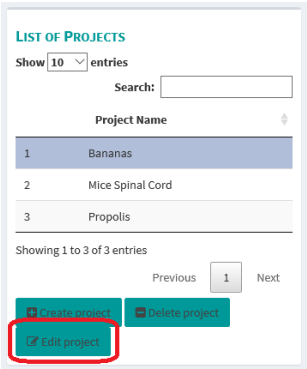

Once clicked, a pop up window appears, where you only need to give the project name and description:



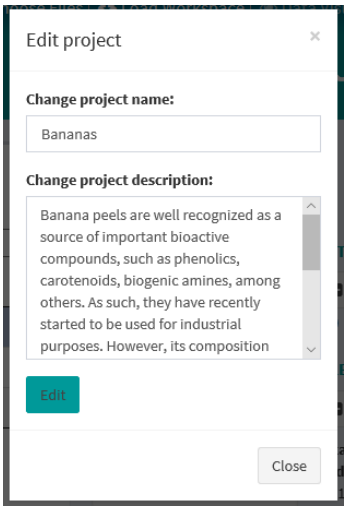

You can also choose if you want to make the project public or not, although this feature and the other ones can be changed latter on.

### Edit project information

To edit a project's information, you can click the button "Edit project" in the "List of Projects" box, when you have that project selected.

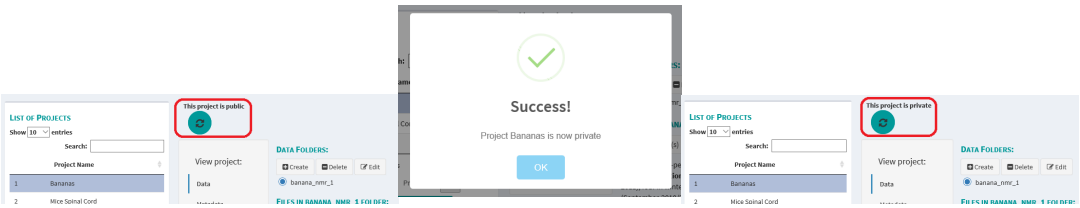


Once clicked, a pop-up window will appear so that you can change the project's name and description:




### Set project to public or private

To change the project between being private and public, you will only need to click a button present at the top middle of the page and a pop-up window appears to say that the task was successfully performed:




(a) Here, the project is public.

(b) Pop-up window appears to inform that the project was successfully changed to private.

(c) Now, the project is private.

### Project's Data

To edit any information regarding data folders of a project, you will have to select the "Data" tab, after selecting the project, from the list of tabs present in the middle of the page. All the information regarding the data folders of the project will appear at the right.

At the top, three buttons that allow to perform tasks on data folders are present ("Create", "Delete" and "Edit"). Below these, a set of options with the data folders present in the project are shown, if any, so that the user can select a data folder and perform tasks on it.

When a data folder is selected, all information regarding this folder appears below, which consists on four buttons that allow to perform tasks on the data files in the folder ("Upload file(s)", "Delete file(s)", "View" and "Download"), the data type, the folder description and the list of files present.

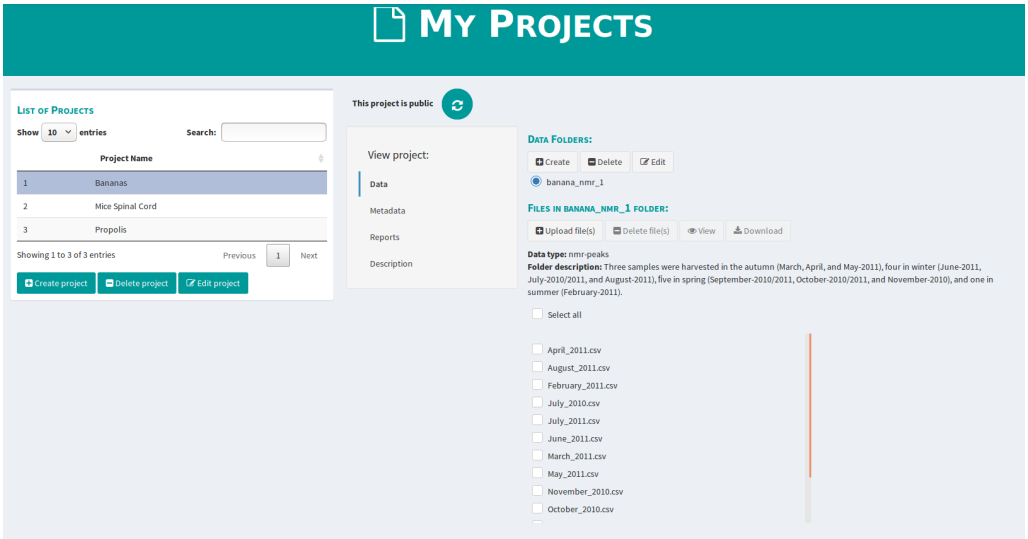
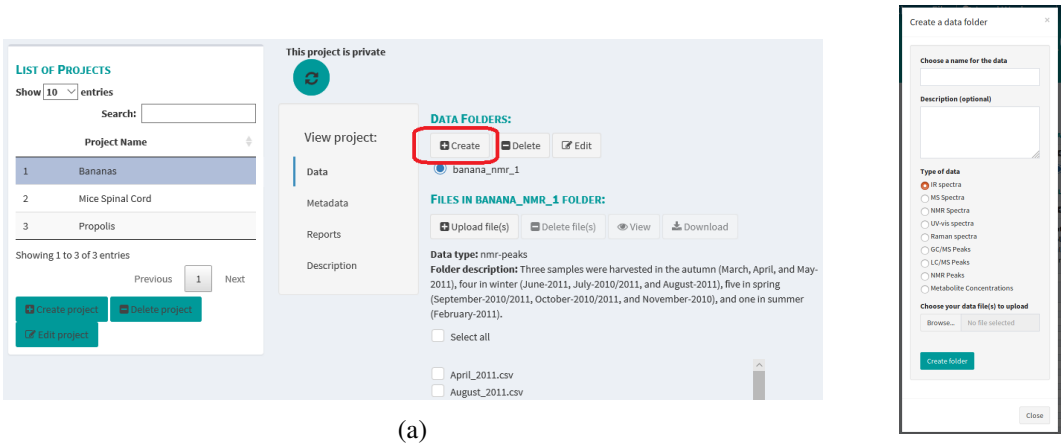



Figure 2.4: Layout of "My Projects" page when information on a data folder is showing.

To **create a new data folder**, you need to press the "Create" button, which will lead to a pop-up window where the following options have to be set:

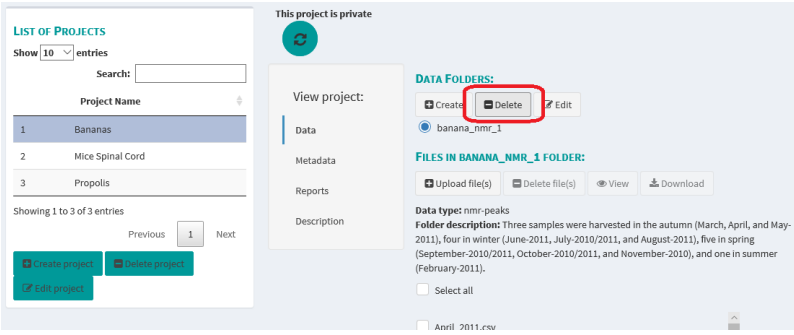
- Data folder name;
- Description;
- Type of data: whose available options are the ones supported ("IR spectra", "MS spectra", "NMR Spectra", "UV-vis spectra", "Raman spectra", "GC/MS Peaks", "LC-MS Peaks", "NMR Peaks", "Metabolite Concentrations");
- Upload the data files.



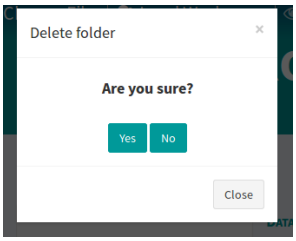

(a) (b)

Figure 2.5: You must (a) click the button "Create" so that (b) you can create the data folder.

To **delete a data folder**, you will only need to select the folder to delete and press the button "Delete". A pop-up window will appear asking if you are sure you want to delete the selected data folder:

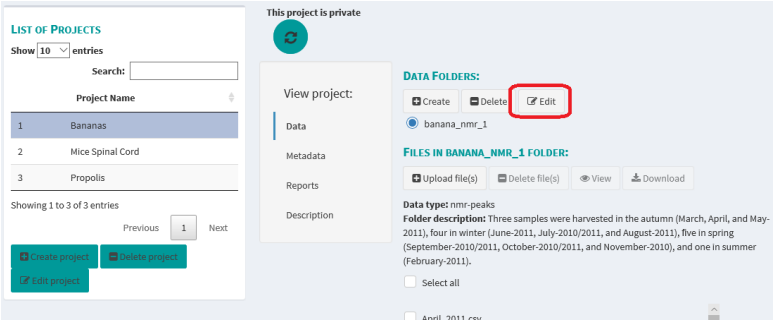



(a)

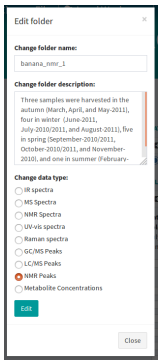



(b)

To **edit the information of a data folder**, you will only need to select the folder to edit and press the button "Edit". A pop-up window will appear so that you can change the folder's name, description and/or type of data:

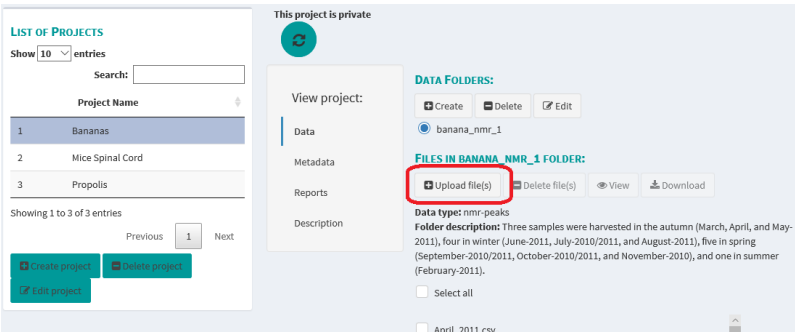



(a)

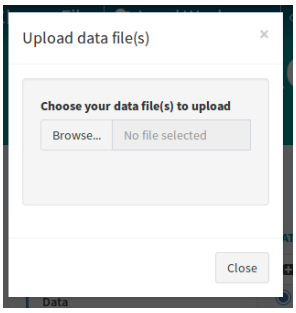



(b)

To **upload files to a data folder already created**, you will need to click the "Upload file(s)" button. A pop-up window will appear, so that you can upload the file(s) wanted:

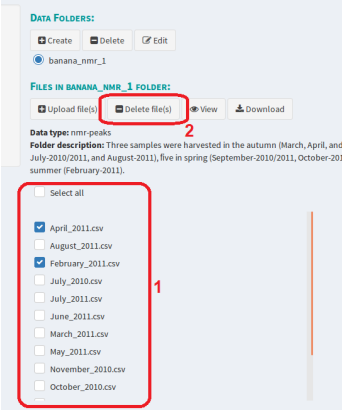



(a)

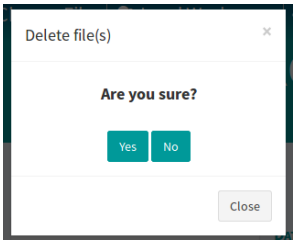



(b)

To **delete files from a data folder**, you will need to select the files to delete, from the list of data files, and press the button "Delete file(s)". A pop-up window will appear asking if you are sure you want to delete the selected data files:



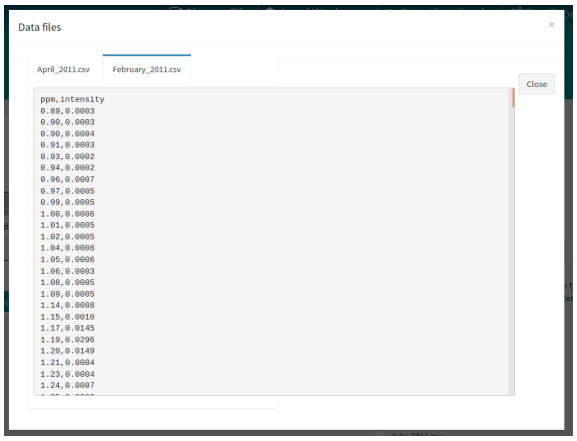

(a)




(b)

To **view the content of data files**, you will have to select the file(s), from the list of data files, and press the button "View". A pop-up window will appear with the content of the selected file(s). Not all types of files are yet supported in this task:


(a)




(b)

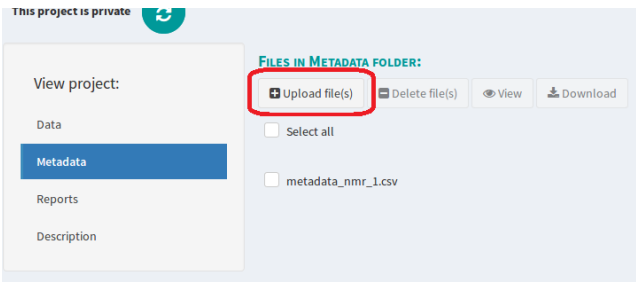
To **download data file(s)**, you will only need to select the data file(s), from the list of data files, and press the button "Download":

### Project's Metadata

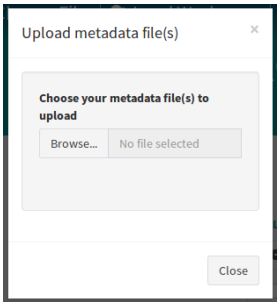
To edit any information regarding metadata files of a project, you will have to select the "Metadata" tab, after selecting the project, from the list of tabs present in the middle of the page. All information regarding the metadata files of the project will appear at the right.

At the top, four buttons that allow to perform tasks on the metadata files are present ("Upload file(s)", "Delete file(s)", "View" and "Download"). Bellow this, there is a list of the metadata files.

To **upload metadata file(s)**, you will need to click the "Upload file(s)" button. A pop-up window will appear, so that you can upload the file(s) wanted:

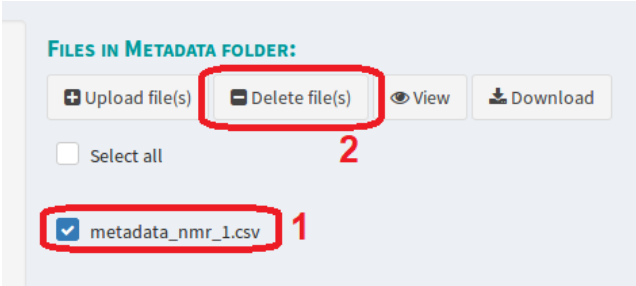



(a)

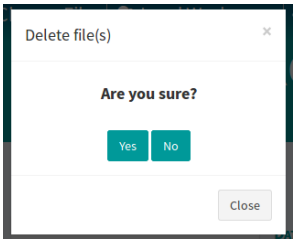



(b)

To **delete metadata file(s)**, you will need to select the files to delete, from the list of metadata files, and press the button "Delete file(s)". A pop-up window will appear asking if you are sure you want to delete the selected metadata files.

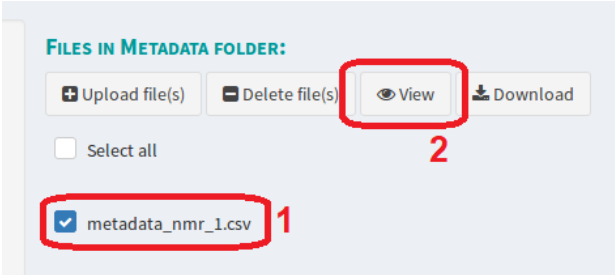



(a)




(b)

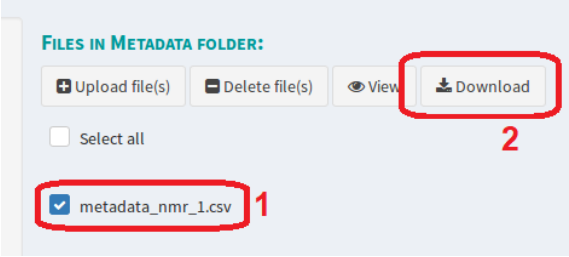
To **view the content of a metadata file**, you will have to select the file, from the list of metadata files, and press the button "View". A pop-up window will appear with the content of the selected file. As the metadata files must have CSV or TSV format, all metadata files can be seen:




(a)


(b)

To **download metadata file(s)**, you will only need to select the metadata file(s), from the list of metadata files, and press the button "Download":

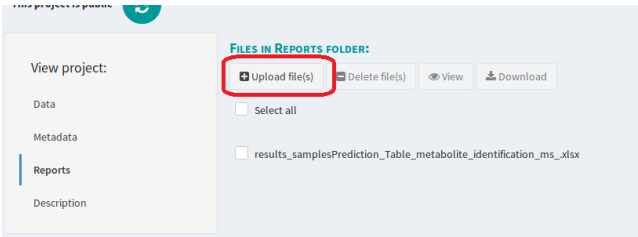



### Project's Reports

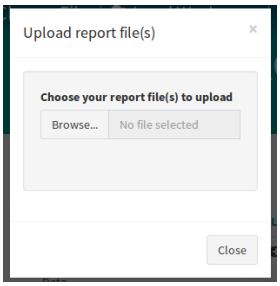
To see any information regarding reports files of a project, you will have to select the "Reports" tab, after selecting the project, from the list of tabs present in the middle of the page. All information regarding the reports files of the project will appear at the right.

At the top, four buttons that allow to perform tasks on the reports files are present ("Upload file(s)", "Delete file(s)", "View" and "Download"). Bellow this, there is a list of the reports files that were previously uploaded to the account or saved after an analysis was performed.

To **upload report file(s)**, you will need to click the "Upload file(s)" button. A pop-up window will appear, so that you can upload the file(s) wanted:

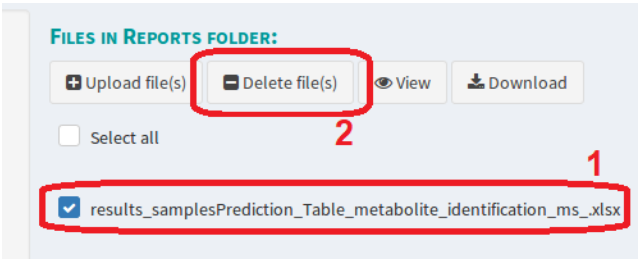



(a)

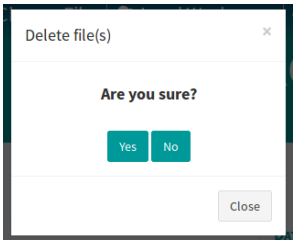



(b)

To **delete report file(s)**, you will need to select the files to delete, from the list of report files, and press the button "Delete file(s)". A pop-up window will appear asking if you are sure you want to delete the selected report files.

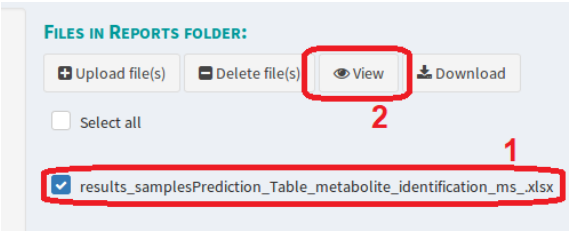



(a)

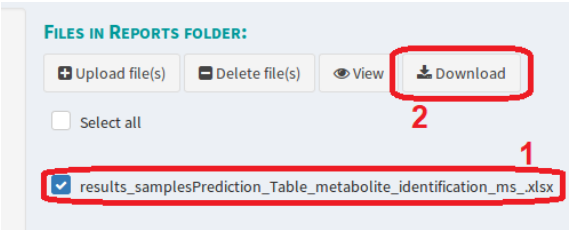



(b)

To **view the content of a report file**, you will have to select the file, from the list of report files, and press the button "View". As the only format here supported is HTML, a new tab in the web browser with the report will open.

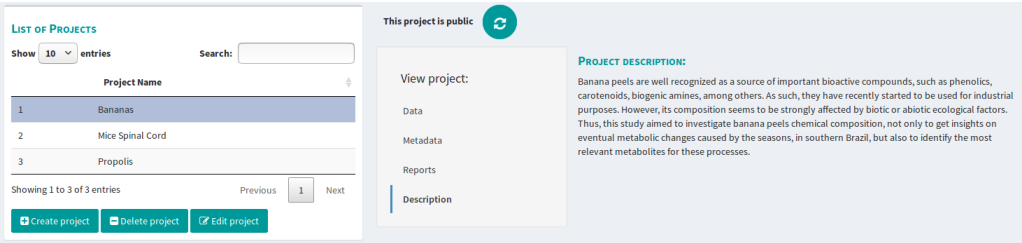



To **download report file(s)**, you will only need to select the report file(s), from the list of report files, and press the button "Download".



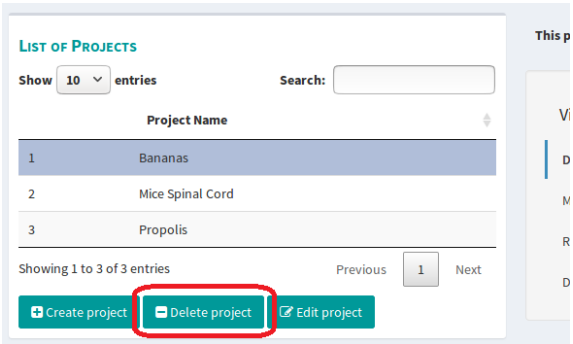

### View Project's Description

One final information that you can see on each project is the description given to the project, accessible through the "Description" tab, from the list of tabs present in the middle of the page:

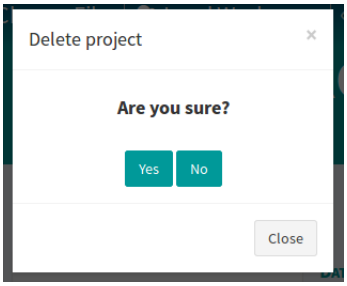



### Delete a project

To delete a project, you will only need to click the button "Delete project", when you have the project you want to delete selected:




Once clicked, a pop-up window will appear asking if you are sure you want to delete the project. All files and information on the project will be deleted:




### 2.2.4 Public Projects

This page is accessed through the sidebar panel and it is accessible to whoever enters the website. It is here where the users can access information on public projects that are stored in the website database.

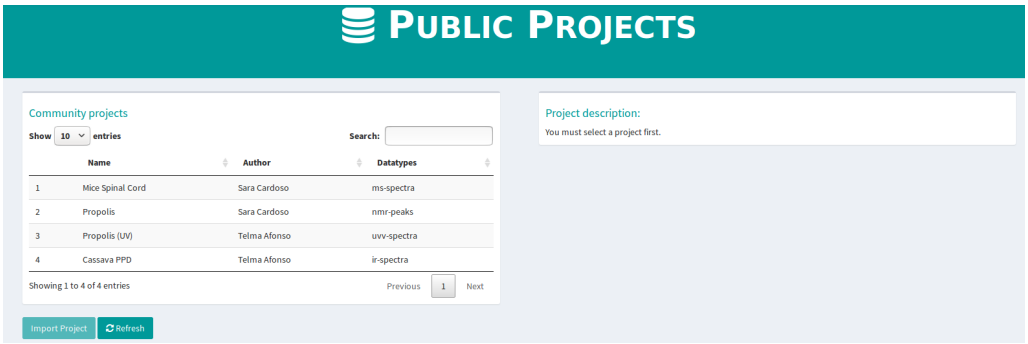



Figure 2.16: Layout of the "Public Projects" page.

#### View general information on all public projects

In the left side of the page, the users can see a table with general information on each project (one line corresponds to one project), in the box named "Community projects". This table has information on the name given to the project, the author of such project (name of the user that created the project) and the types of data that are stored.

Below this table, a "Refresh" button is provided, to obtain the latest list of public projects.

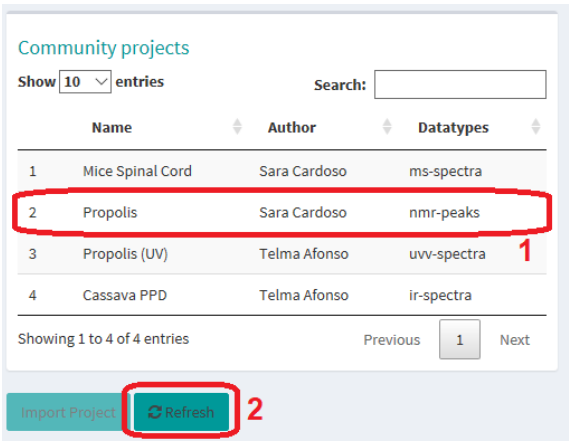
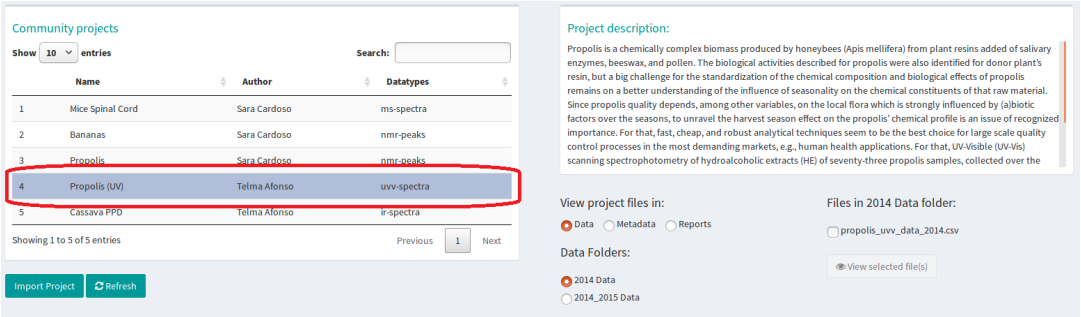



Figure 2.17: "Community projects" box. The project highlighted with rectangle 1 is named *Propolis*, created by author *Sara Cardoso*, and the data files here present are of only one type of data: *NMR peaks*. In rectangle 2 is highlighted the "Refresh" button.

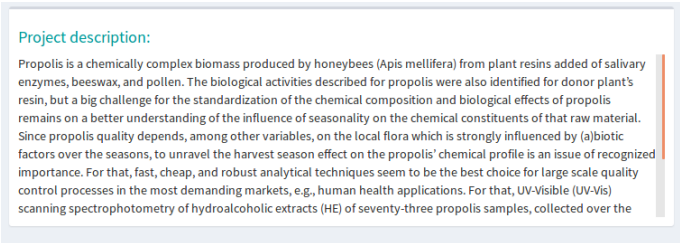
### View detailed information on a public project

By clicking in one project of this table, a more detailed information on the project appears at the right side of the page:

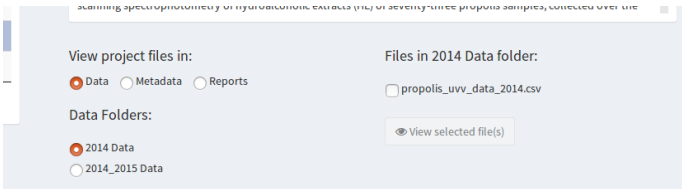



Here, the user can see the:

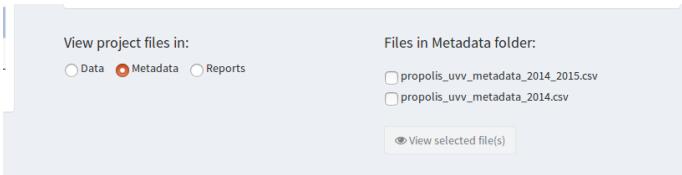
- Project Description;



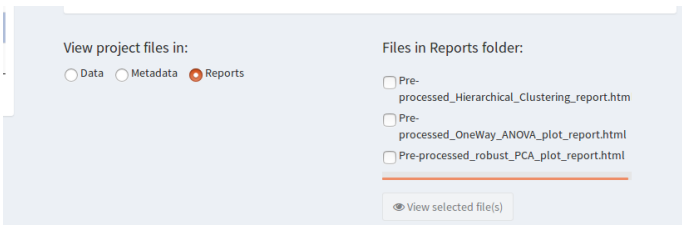

- Data files in each data folder;




- Metadata files;



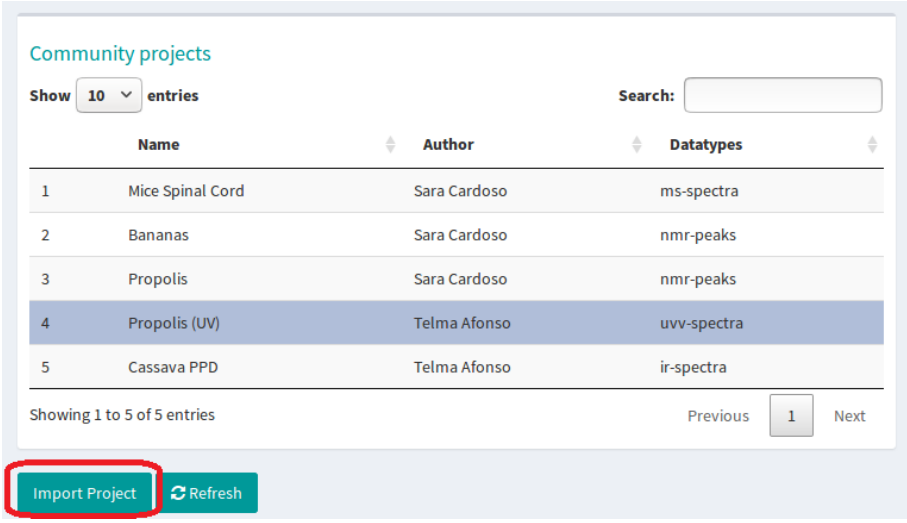

- Reports the project owner saved into it.



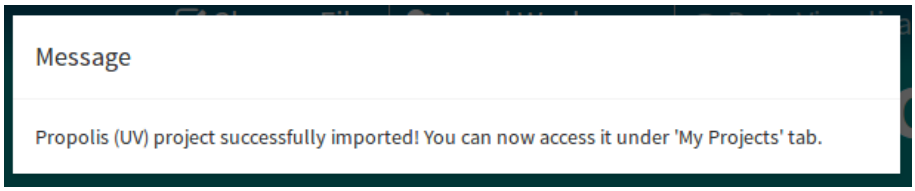
### Copy a public project to the personal account

As stated before, any project can be imported into the user's private projects collection, given that the user is authenticated and the project itself is not owned by the user nor does he/she already own a project by that name.

To do this, the user only needs to click in the button saying "Import Project", present below the projects table:




(a)




(b)

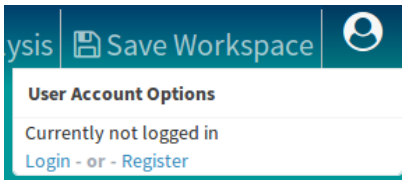
## 2.3 User Account

Any of the following tasks related to the user account are performed by accessing through the account authentication icon, present at the top right corner of the webpage:

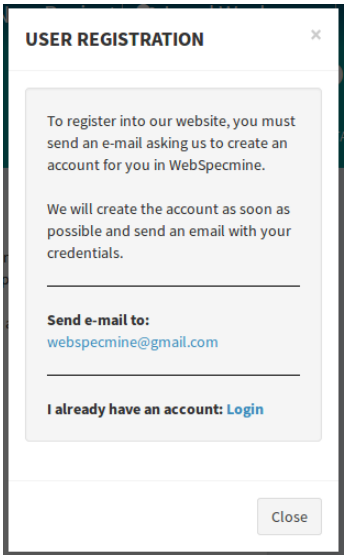

Figure 2.19: Account authentication icon.

### 2.3.1 User Registration

To create an account in *WebSpecmine*, you will need to click on the user icon and press the "Register" link:

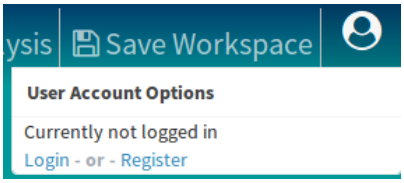


After this, a pop up window will appear so that you can register yourself, by giving the first and last name, your e-mail and password wanted:

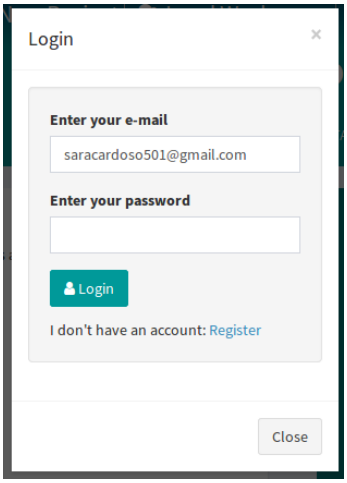



### 2.3.2 Login

To log in, you must click on the user icon and press the "Login" link:

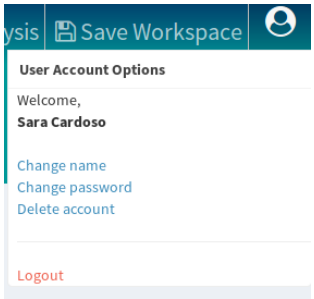



After this, a pop up window will appear so that you can log in:



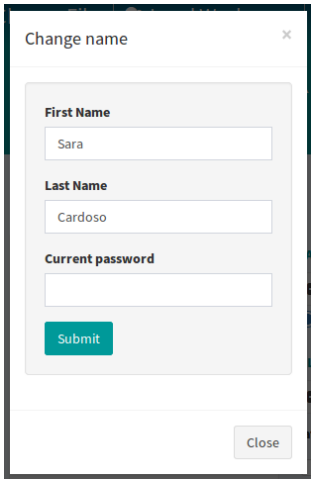

### 2.3.3 Account options

All the options available are accessible once the user logs in. By clicking in the user icon, the account options panel appears below:



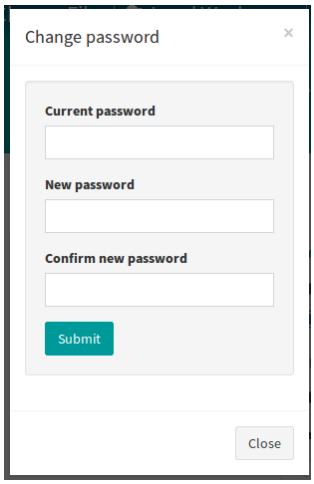

### Change user name

To change the user name, you will need to, besides providing the new name wanted, insert the correct password of your account, so that the task can be performed:



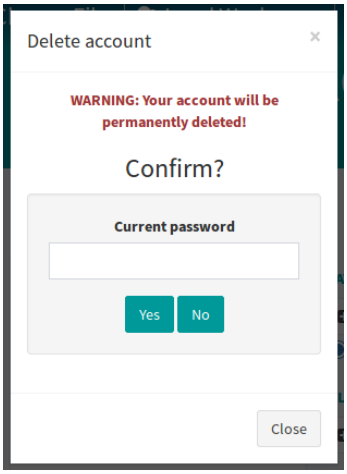

### Change password

To change the password, you will need to insert the correct password of your account, the new one wanted and repeating the new password, to confirm is right, so that the task can be performed:



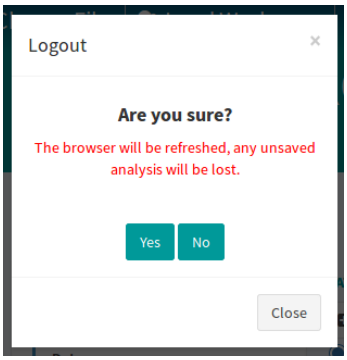

### Delete account

To change the password, a warning pop-up window will inform you that the task is not reversible and ask you to give the correct account password, so that the task can be performed:




### 2.3.4 Logout

When logging out, a pop-up window will appear warning that any unsaved analysis will be lost:




## 2.4 Workspaces

### 2.4.1 What is a workspace?

A workspace consists on all data the user is working on at the moment and the possible results that have already been obtained. Each data folder from a project can have a workspace associated. Therefore, only users with an account can save workspaces.

However, not only users that hold an account can load a workspace, but also users with no account. The difference remains in the fact that logged in users upload workspaces from their account or from public projects, even if the project to where they belong was not yet copied to the account, whilst logged out users can only upload public workspaces.

See section 2.5.8 on how to save and load workspaces.

## 2.5 Load Data for analysis

There are three different ways of choosing data for analysis.

- ***New Project***: available for users not logged in, you will have to upload here the files to analyze;
- ***Choose Files***: available for users logged in their accounts, you will have to choose here the files to analyze, from the ones that you have previously saved into the account;
- ***Load Workspace***: for any type of user, you can load a workspace (data and/or results) available to analyze

### 2.5.1 New Project

When the user clicks the "New Project" button, a pop-up window to submit the files for analysis appears.

In this window, according to the the data type chosen on the top of the window, either "MS Spectra (.mzXML, .netCDF, mzData)", "NMR or MS peaks lists", "Concentrations" or "Spectral Data", the user has to set some options about the data and metadata files, present under the data type choices. Optional information can also be given.

Here is an example of what can be seen in this window:

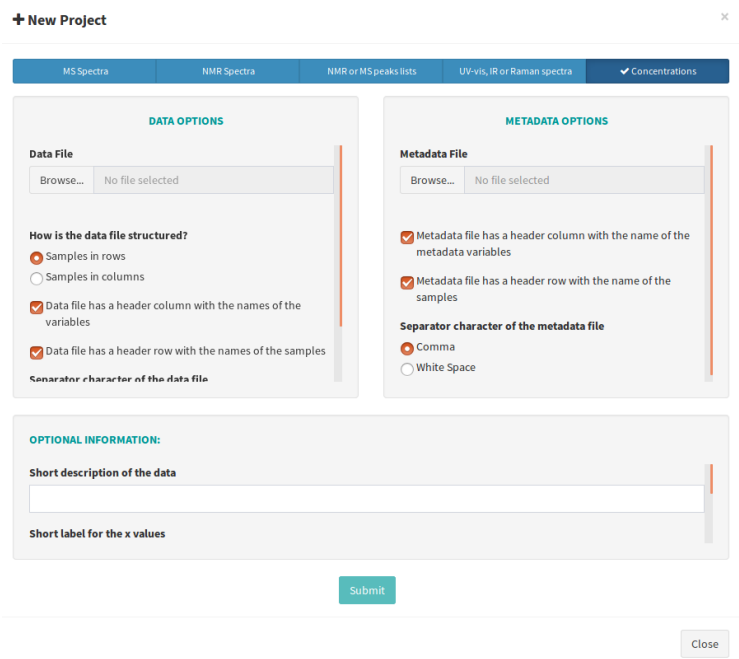



Figure 2.20: Layout of the submission of a new project of concentrations data.

For cases when more than one data file needs to be uploaded, a .zip file with the data files must be submitted. Concentrations data must be only one file. For spectral data it can either be a .zip file with files of the supported format or only one file (CSV format).

The "Submit" button present at the bottom of the window is only enabled when both the data and metadata are submitted. After clicking the button "Submit", the files are processed and the corresponding data stored under the dataset name *OriginalData*.

Then, the window disappears and the page "Run Analysis" appears. All the other buttons in the header panel will be made available, except for the "Save workspace" one, only available for logged in users.

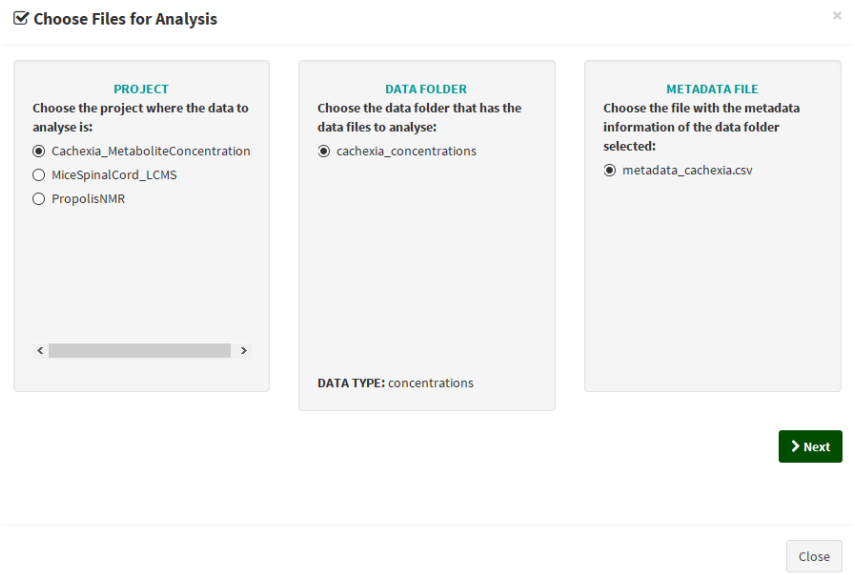
Also, the tab "Dataset being used" appears on the sidebar panel, with one selected option, "OriginalData", which means that this is the dataset being currently used for analysis.

Each time a new submission of files for analysis is done, all the work that the user may have done before is lost.

The uploaded files are not stored anywhere.

### 2.5.2 Choose Files

When you click the "Choose Files" button, a pop-up window to choose the files from your account for analysis appears:

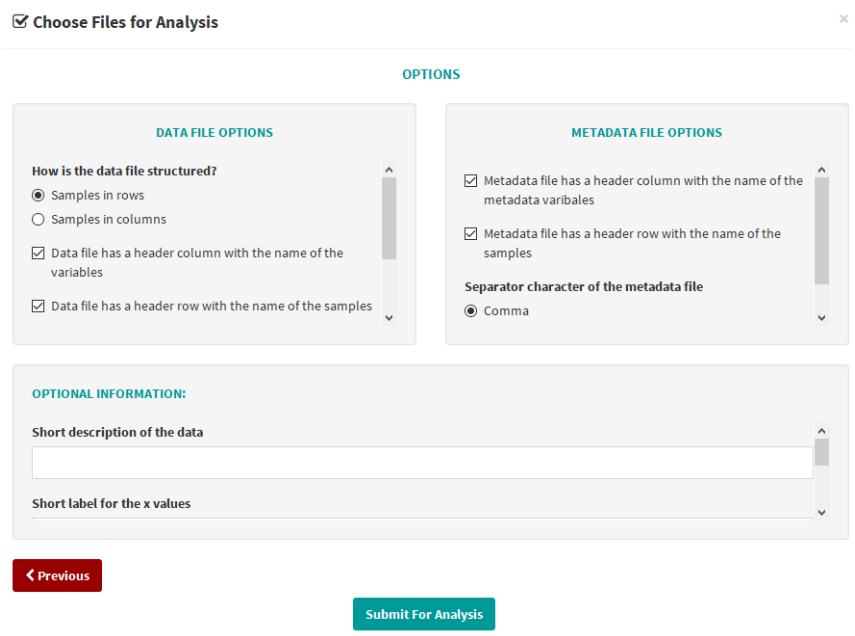



Initially, three boxes appear on the window, so you can choose:

- The project to work with;
- The data folder from the chosen project that contains the data files to analyse;
- The metadata file from the chosen project that contains the metadata information about the data to analyse.

Only the projects that do not have empty Data and Metadata folders can be selected for analysis.

After doing so, the user will have to set some options regarding the data type in question. After this, the user is able to submit the chosen files for analysis, by clicking the "Submit for Analysis" button:




Everytime a different project is chosen, the workspace regarding the previous project is lost, unless the user saves it first.

### 2.5.3 MS Spectra Options

The **data options** made available concern the feature (peak) detection in the chromatographic time domain. The user must choose:

- The profile generation method: "bin", "binlin", "binlinbase" or "intlin";
- The full width at half maximum (fwhm) of matched filtration gaussian model peak: commonly 30 for LC-MS spectra and 4 for GC-MS spectra;
- The bandwidth (standard deviation or half width at half maximum) of the Gaussian smoothing kernel, to apply to the peak density chromatogram: commonly 30 for LC-MS spectra and 5 for GC-MS spectra;
- The peak intensity measure: "Integrated area of original (raw) peak", "Integrated area of filtered peak", "Maximum intensity of original (raw) peak" or "Maximum intensity of filtered peak".

The available **metadata options** concern the way how the metadata file is formatted. The user must say if the file:

- Has a header column with the name of the metadata variables;
- Has a header row with the name of the samples;
- Has a comma or white space separating the data.

The user can also provide, optionally, a short description of the data.

### 2.5.4 NMR Spectra Options

Regarding the **data options**, the user must say:

- Data format: "BRUKER processed data" or "VARIAN raw data";
- If each data folder given inside the zip provided in *New Project* also compressed (.zip). This option only appears in *New Project*;
- If VARIAN format is selected, you will also have to choose if you want to perform or not zero filling and/or apodization.

The available **metadata options** concern the way how the metadata file is formatted. The user must say if the file:

- Has a header column with the name of the metadata variables;
- Has a header row with the name of the samples;
- Has a comma or white space separating the data.

Optional information can also be given by the user, such as a short description of the data and short labels for the x and y values.

### 2.5.5 NMR or MS peaks lists Options

Regarding the **data options**, the user must say:

- The type of data peaks submitted: "NMR", "GC-MS" or "LC-MS" peaks;
- If the file has a header row with the names of the data variables;
- If the file has a comma or white space separating the data;
- If the character used for decimal points is a comma (",") or a point (.)

The available **metadata options** concern the way how the metadata file is formatted. The user must say if the file:

- Has a header column with the name of the metadata variables;
- Has a header row with the name of the samples;
- Has a comma or white space separating the data.

Optional information can also be given by the user, such as a short description of the data and short labels for the x and y values.

When this type of data is chosen, there is no "Submit"/"Submit for Analysis" button initially, but a "Next" button, which, in the "New Project" feature, is only enabled when both data and

metadata files are submitted.

When the user clicks next, a set of **options to do the alignment of peaks**, after processing the files, is provided. The user is able to choose between the MetaboAnalyst and Specmine algorithms. When the specmine algorithm is chosen, the user must give the size of the step, in ppms. On the other hand, when the MetaboAnalyst method is chosen, the metadata variable to be used can be chosen. The user can then press the "Submit"/"Submit for Analysis" button.

### 2.5.6 Concentrations Options

The **data options** made available concern the way how the file is formated. The user must say if the samples are distributed over the rows or columns. According to what he responds to this option, the user must say if the file has a header column with the names of the variables or samples and a header row with the names of the samples or variables. If the user says that the file does not have the samples names, he will be asked to give them, by writing them, separating each one by a comma. Finally, the user must specify the data separator character ("Comma" or "White space").

The available **metadata options** concern the way how the metadata file is formatted. The user must say if the file:

- Has a header column with the name of the metadata variables;
- Has a header row with the name of the samples;
- Has a comma or white space separating the data.

Moreover, the user is able to provide, optionally, a short description of the data and short labels for the x and y values.

### 2.5.7 Spectral Data Options

The **data options** made available concern the way how the file is formated. The user must say:

- The type of spectral data: "UV-Vis", "Infrared" or "Raman";
- Type of file(s) submitted: "CSV file", "CSV folder", "DX folder", "SPC folder", "XLSX folder".

If the user is submitting a CSV file it must also say:

- What is the character that is separating the data values: "Comma", "Semicolon" or "Tab";
- If the samples are distributed over the rows or columns;
- If the file has a header column with the name of the samples or variables;
- If the file has a header row with the name of the samples or variables.

If the user is submitting CSV files it must also say:

- What is the character that is separating the data values: "Comma", "Semicolon" or "Tab";
- If the files have row headers;
- If the files have experimental info in the first lines and, if yes, how many lines are, so that they can be skipped when the website reads the data files.

If the user is submitting SPC files it must also say if the website should read the subheaders or not.

If the user is submitting XLSX files, it must also say if the files have row headers or not.

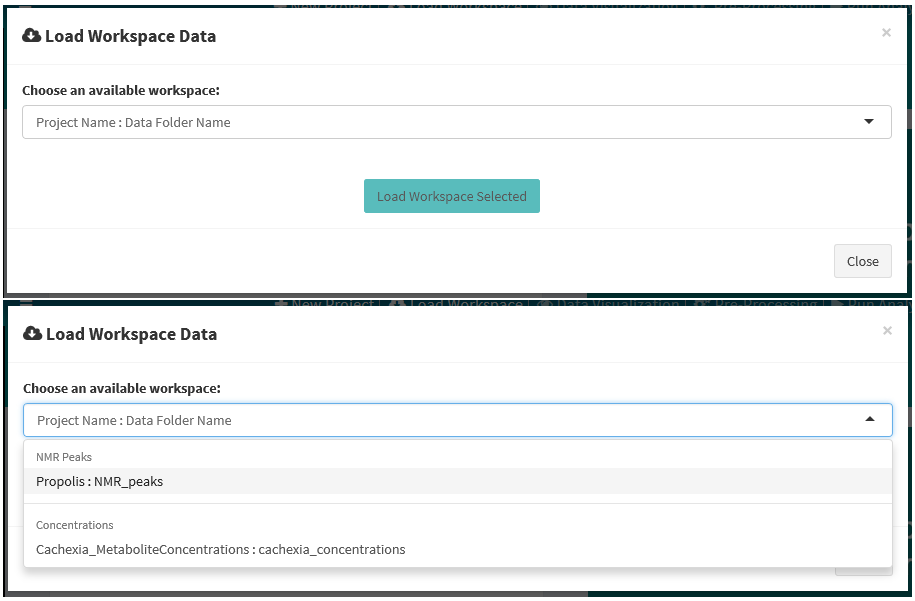
As regards to the **metadata options**, the user must say if the file:

- Has a header column with the name of the metadata variables;
- Has a header row with the name of the samples;
- Has a comma, white space or semicolon separating the data.

### 2.5.8 Load and Save Workspaces

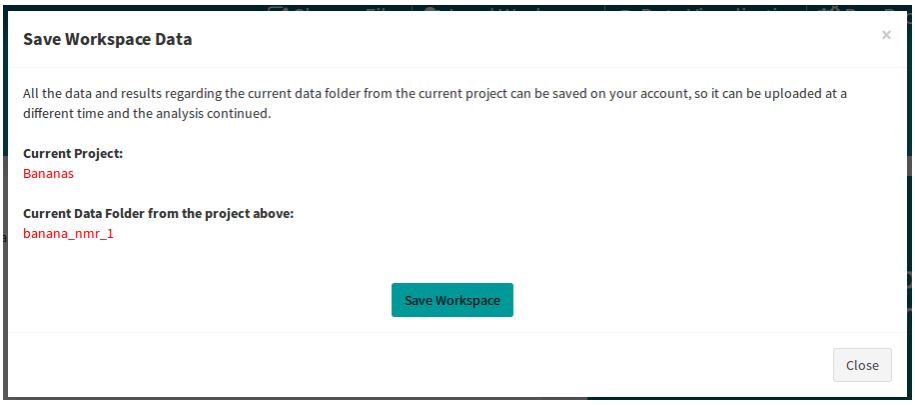
The load and save workspace features are both available through the header panel. After clicking in one of these buttons, the respective window appears above the website.

In the load workspace window, the available workspaces to load are presented to the user grouped by the respective type of data, in order to make easier the search for the wanted data. Furthermore, each workspace is identified by the project and data folder that it corresponds to:




Similarly to what happens with the submission or choice of different projects, loading a different workspace implies loosing the workspace the user is working at that time, unless it is saved.

When saving a workspace, the pop-up window that appears informs the user what is the project and data folder that the workspace is related with:




**Remember that** workspaces can only be saved into someone's account. Therefore, this feature is only available for users that are logged in.

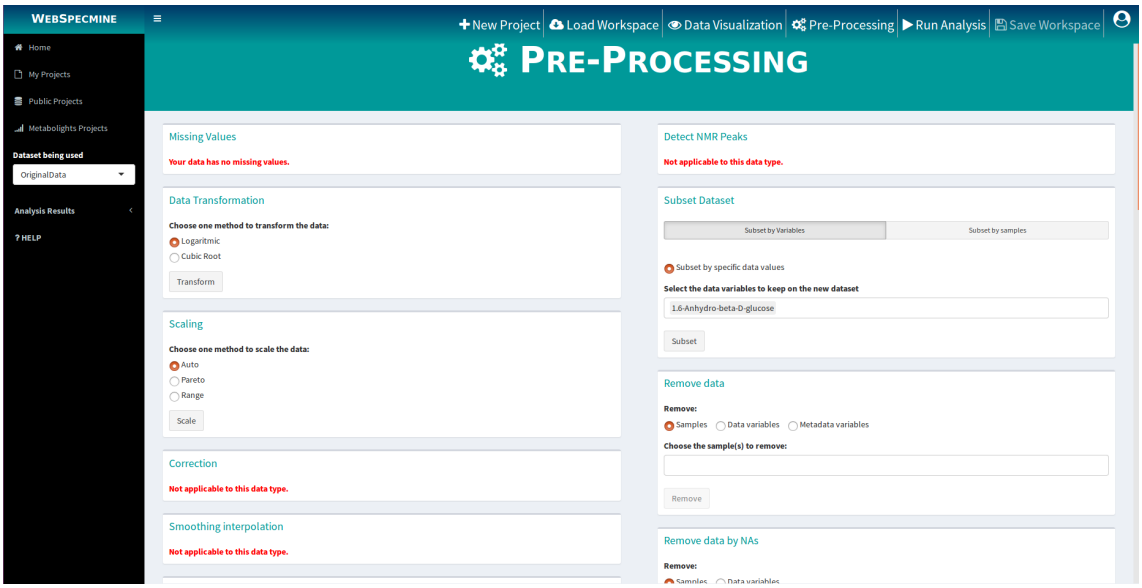
After loading a public workspace, the logged in user has to copy the public project associated with the public workspace loaded to be alowed to save the workspace, as a warning message appears in the pop-up window of "Save Workspace" if the user has not yet copied the public project.

Only the user that "owns" the public project can save and, therefore, change the actual associated public workspace(s).

## 2.6 Data Pre-processing

When the user clicks the "Pre-Processing" button, the page that allows the user to pre-process datasets appears.

This page was organized into two columns with the different types of pre-processing in each box:

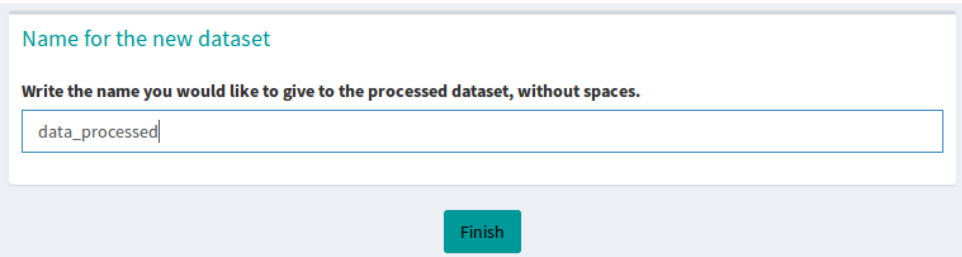



**Some pre-processing boxes will only be available for the user when it comes to spectral data: "Correction", "Smoothing Interpolation", "First Derivative", "Multiplicative Scatter Correction" and "Low-level data fusion".**

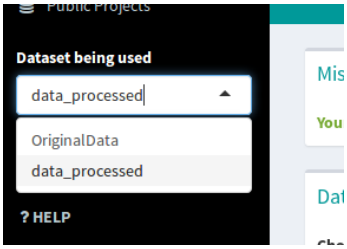
Each box is further discussed in this chapter, below.

The processing is done over the dataset being currently used, and it can be done in any desired order, applying the wanted tasks. Various datasets can be generated, with different pre-processing pipelines, which allows to compare different results of the same analysis, according to the processing pipeline applied.

At the end of the page, the "Finish" button, only enabled when the dataset name input is filled, allows the user to indicate that the processing pipeline is defined:



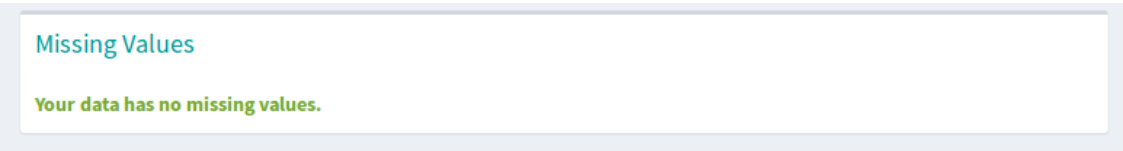

After naming the pre-processing, the name of the dataset will appear on the sidebar panel, in the section "Dataset being used", so that the user can choose the new dataset for further analysis:



If the chosen name already exists, the site will not allow the user to save the pre-processing done when he clicks the "Finish" button and the message "A dataset with that name already exists! Please choose a different one" will appear, giving the opportunity of renaming the dataset.

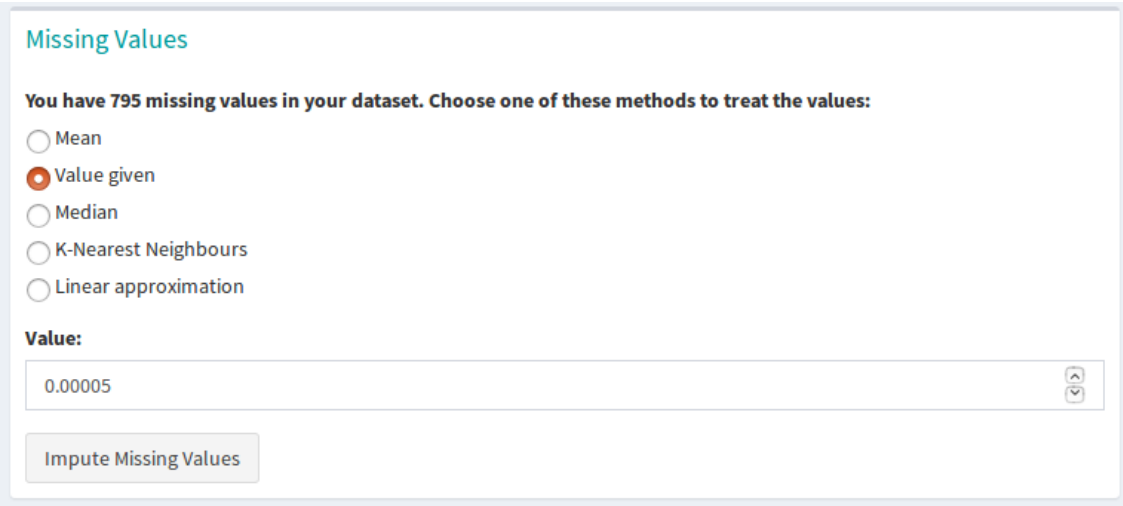
### 2.6.1 Missing Values

In the "Missing Values" box, a message saying "Your data has no missing values" appears if the dataset in use does not have missing values:




If it has, the options to treat missing values will appear. These include replacing the missing values by:

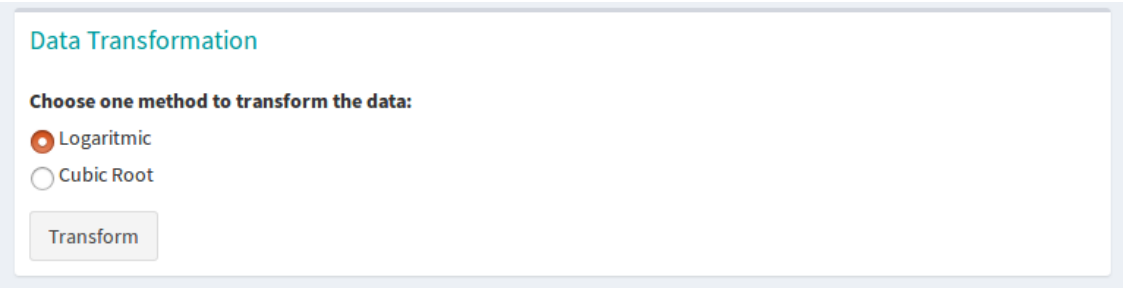
- The mean;
- The median;
- A value given by the user: when this choice is selected, an input will appear below so that a value can be specified;
- Calculating the K-Nearest Neighbours: when this choice is selected, an input will appear below so that a number K can be specified;
- Doing a linear approximation.




### 2.6.2 Data Transformation

The methods made available are:

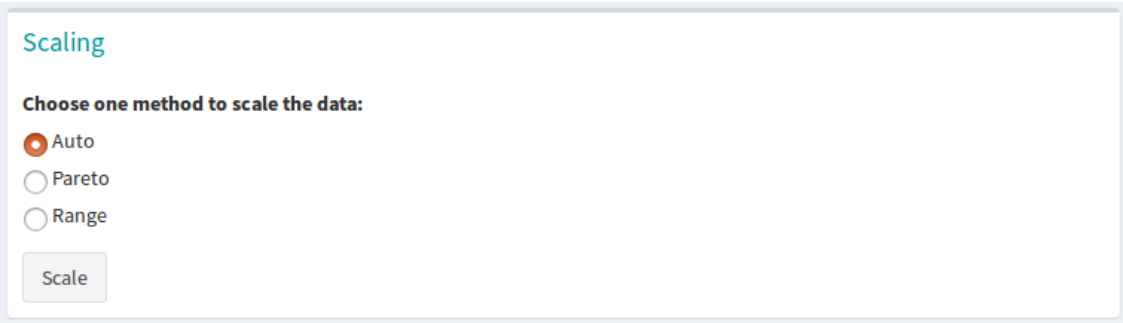
- Logaritmic;
- Cubic Root.




### 2.6.3 Scaling

The methods made available are:

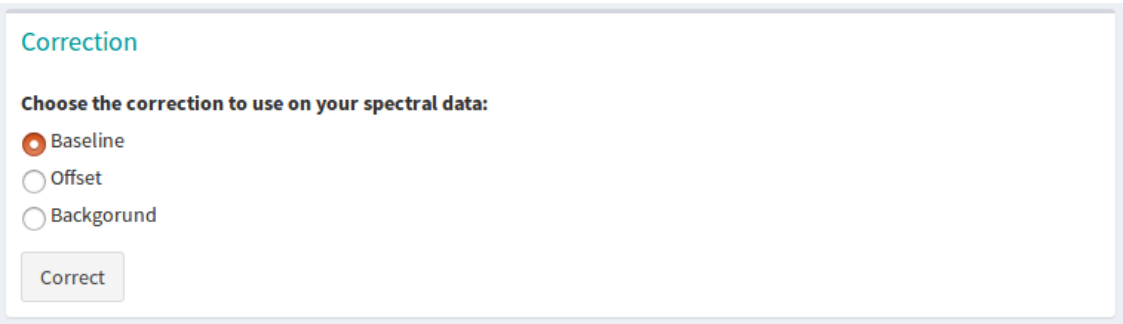
- Auto;
- Pareto;
- Range.




### 2.6.4 Correction

The methods made available to perform correction are:

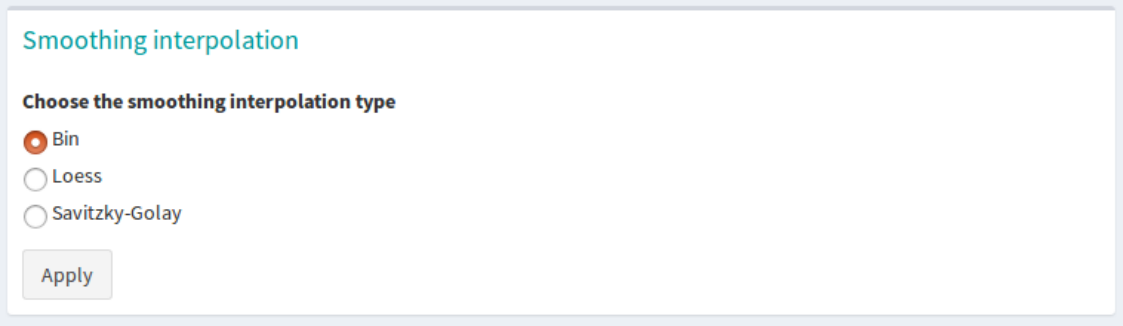
- Baseline;
- Offset;
- Background;




### 2.6.5 Smoothing Interpolation

The types of smoothing interpolation made available are:

- Bin;
- Loess;
- Savitzky-Golay.

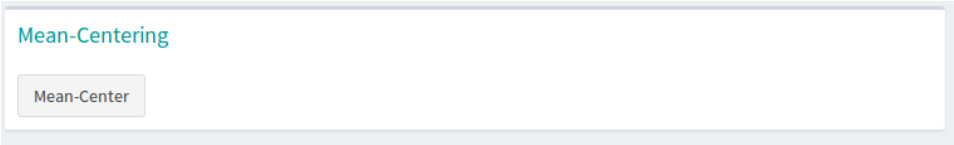



### 2.6.6 Convert to Factor

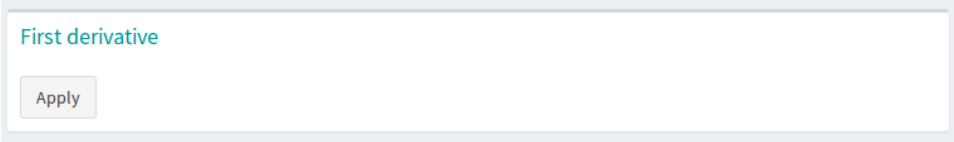
Metadata variables can be converted to factors here, if they are not already:


This feature is specially important if the user wants to perform machine learning, as it is only possible to perform classification problems.

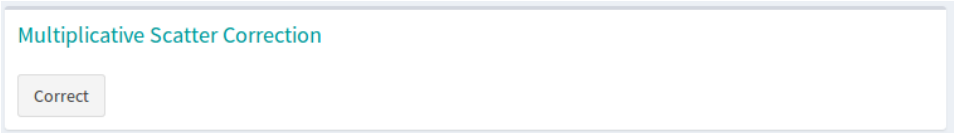
### 2.6.7 Mean Centering




### 2.6.8 First Derivative




### 2.6.9 Multiplicative Scatter Correction

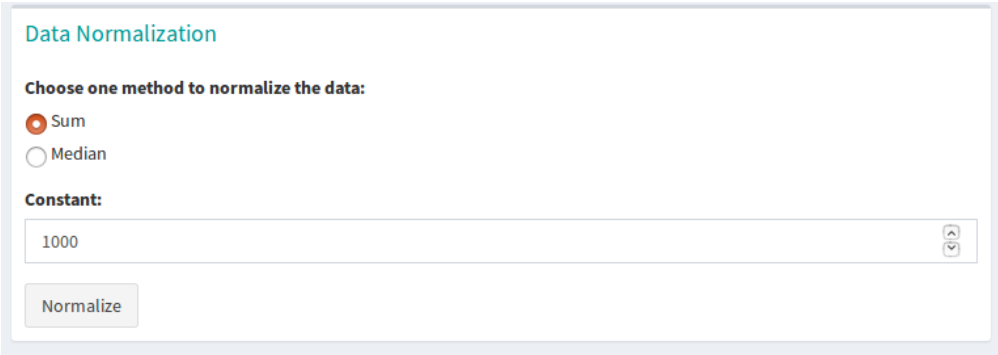



### 2.6.10 Data Normalization

The normalization of the data can be done by the sum of:

- A constant given by the user: when this choice is selected, an input will appear below so that the constant can be specified;

- The median.




### 2.6.11 Detect NMR Peaks

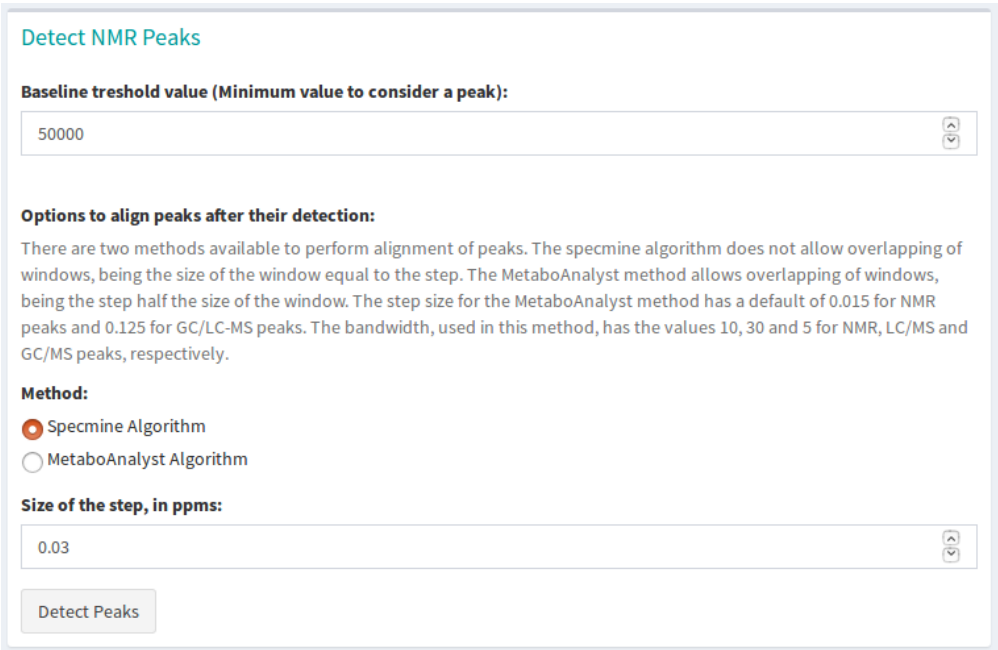
The NMR peak detection box, as the name suggests, is only available for NMR spectra data.

To detect the peaks, you will only need to set the following options:

- Baseline treshold value: it's the minimum intensity value to consider a peak as a detected peak;

After the peaks are detected, they are aligned and the following options have to be set:

- Peak alignment method: "MetaboAnalyst Algorithm" or "Specmine Algorithm";
- Size of the steps (in ppms), if "Specmine Algorithm" is chosen;
- Metadata variable to use in the alignment, if "MetaboAnalyst Algorithm" is chosen;

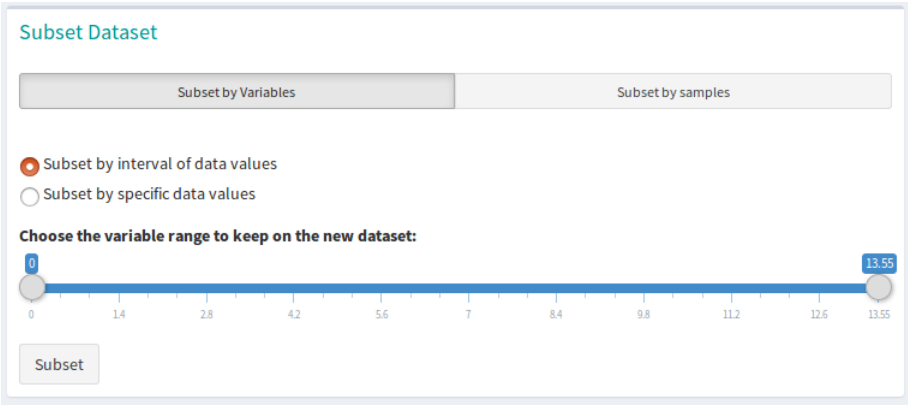



### 2.6.12 Subset Dataset

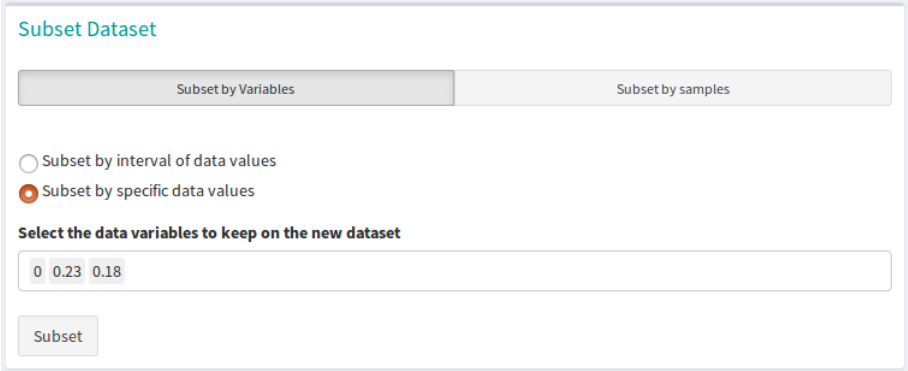
You can get a new dataset with only certain data variables ("Subset by Variables") and/or certain samples ("Subset by samples").

When subsetting the dataset by variables, you can choose between:

- An interval of data variables to keep on the new dataset;

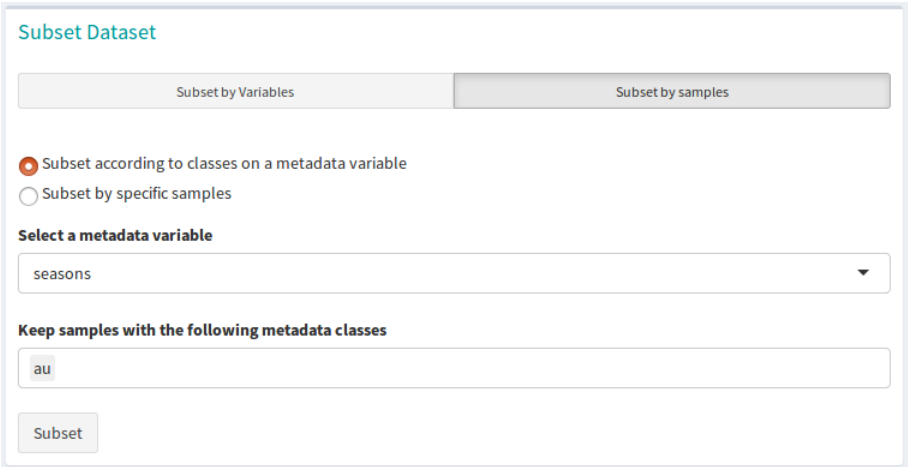



- Keep specific data variables.

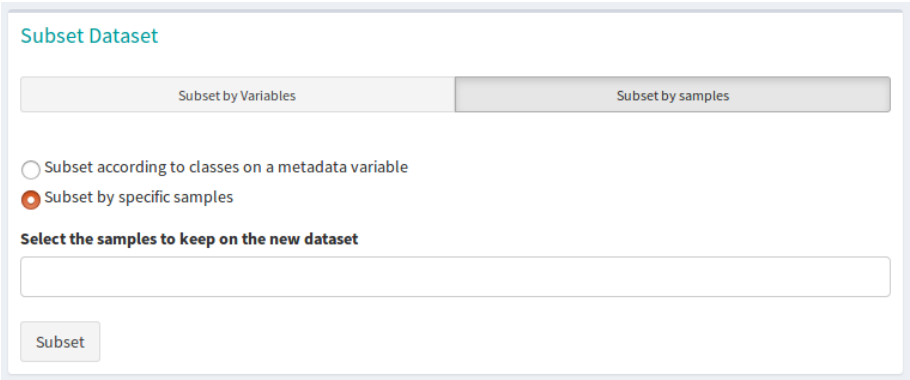



When subsetting by samples, you can choose between:

- Subsetting according to classes on a metadata variable, i.e., only samples that have a certain value(s) for a metadata variable will be kept on the new dataset;



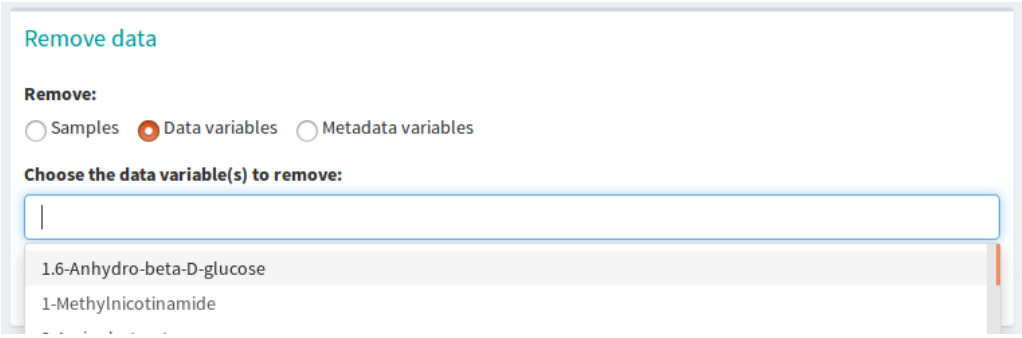

- Keep specific samples.



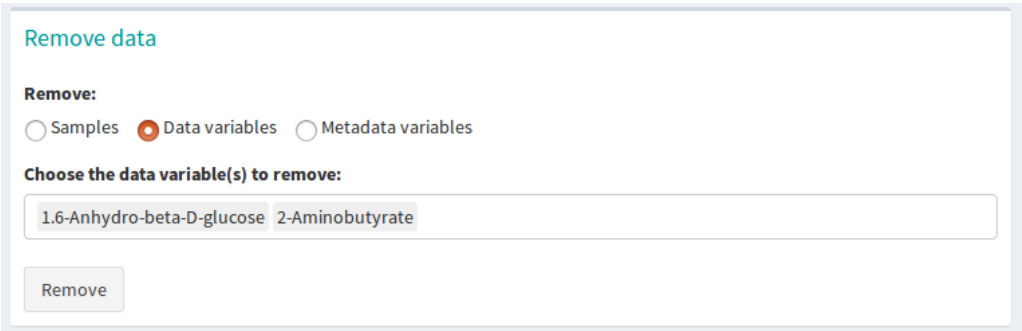

When the dataset being used is of concentrations or MS spectra, subsetting the dataset by an interval of data values is not accessible.

### 2.6.13 Remove Data

Specific samples, data variables and/or metadata variables can be removed:




(a)




(b)

Figure 2.22: Remove data box, (a) when the choices showing the data, metadata or samples names show when the user clicks on the input to insert what to remove and (b) after selecting two data variables names to remove.

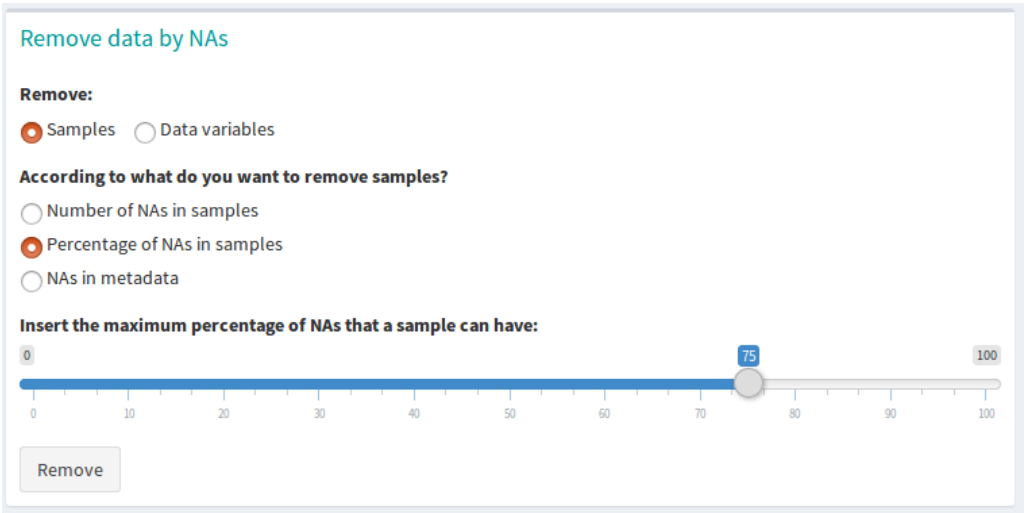
### 2.6.14 Remove data by NAs

Samples and data variables can be removed according to the missing values they have.

In both cases, the data can be removed by the number or percentage of missing values. When the first is chosen, an input will appear, asking the user to give the maximum number of missing values a sample or data variable can have. On the other hand, when the percentage option is chosen,

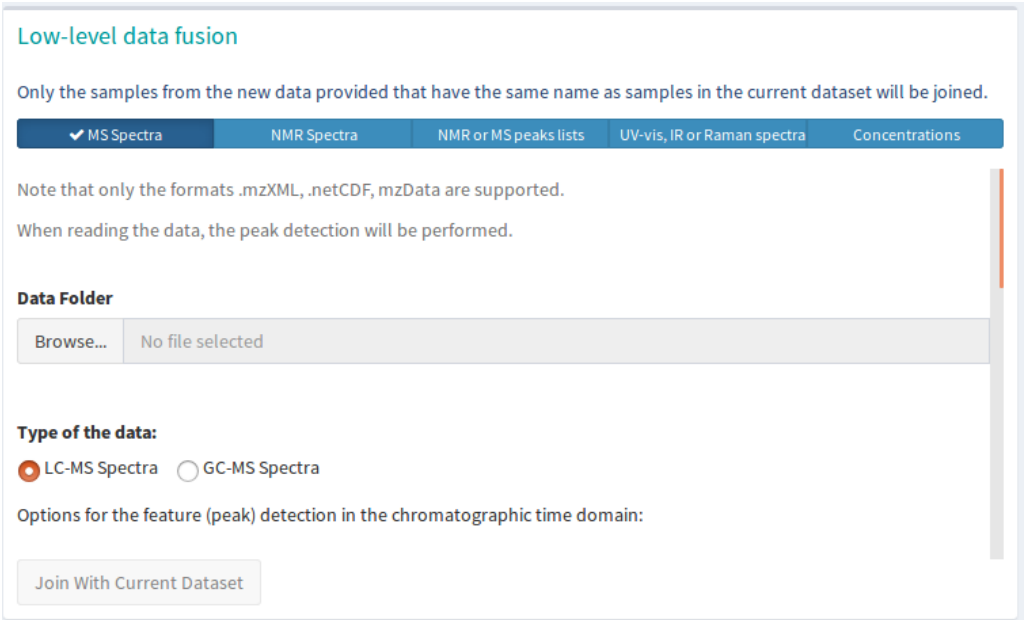
a slider input will appear, allowing the user to set the maximum percentage of missing values a sample or data variable can have.

The samples have an additional option of removing samples if they have missing values in the respective metadata variables.




### 2.6.15 Low-level data fusion

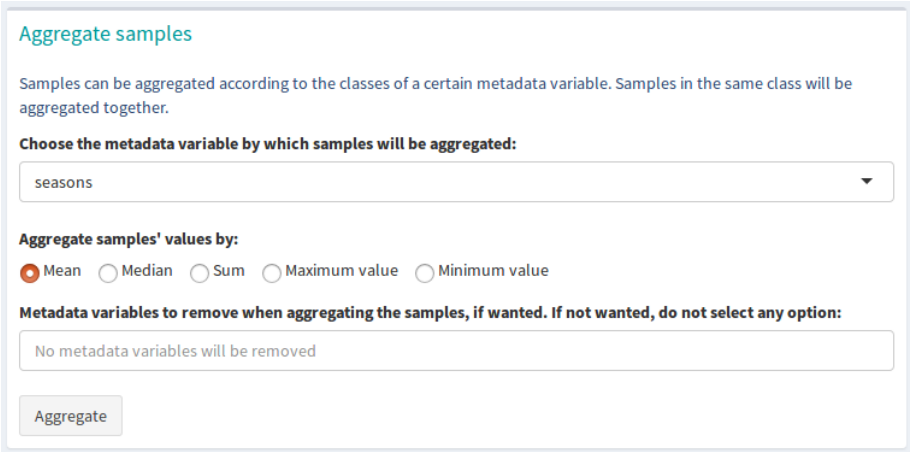
Datasets from different types of data can be merged in one dataset. To merge the dataset you are currently working on with another one of another type of data, you only need to select the type of data to upload in the blue options at the top of the box, upload the file(s) and set the options. The options available are similar to those seen when loading a data for the first time through *New Project* or *Choose Files* (see section 2.5 for more information)




### 2.6.16 Aggregate Samples

Samples can be aggregated (joined in one) according to a metadata variable, i.e., samples that belong to the same class of a metadata variable are joined together in only one sample. Samples can be joined by calculating the mean, median, sum, maximum value or minimum value of the

samples. Furthermore, you can also choose metadata variables to remove, in case they stop making sense once the samples are aggregated.




### 2.6.17 Flat Pattern Filter

Six functions are available for selection in the "Flat Pattern Filters" processing box, including:

- Interquantile range;
- Relative Standard Deviation;
- Standard Deviation;
- Median absolute deviation;
- Mean;
- Median.

The values can be filtered by:

- Percentage: when selected, a slider input with numbers between 0 and 100 appears;
- Treshold: when selected, a numeric input appears;
- Number of variables to remove are calculated automatically.

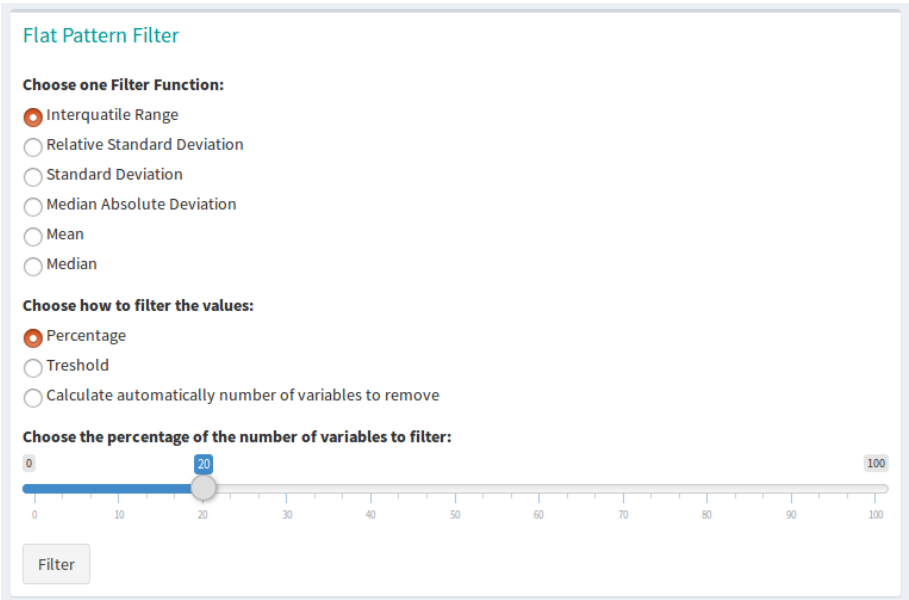



Figure 2.23: Flat Pattern Filter box.

## 2.7 Visualize the data

The "Data Visualization" feature is available through the header panel, and allows the user to see the data and some of its characteristics.

At the top left of this page, the buttons present correspond to what the user can see about the dataset in question. At the right, the content that belongs to the button clicked is shown.

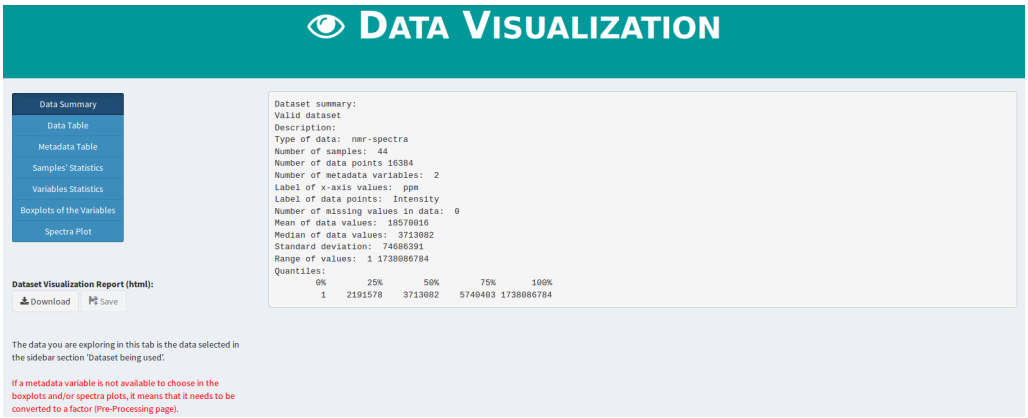
Below these, the possibility to download or save the data visualization report is made available.

The information the user is able to see in this page corresponds to the dataset being currently used, chosen in the sidebar tab "Dataset being used".

### 2.7.1 Data Summary

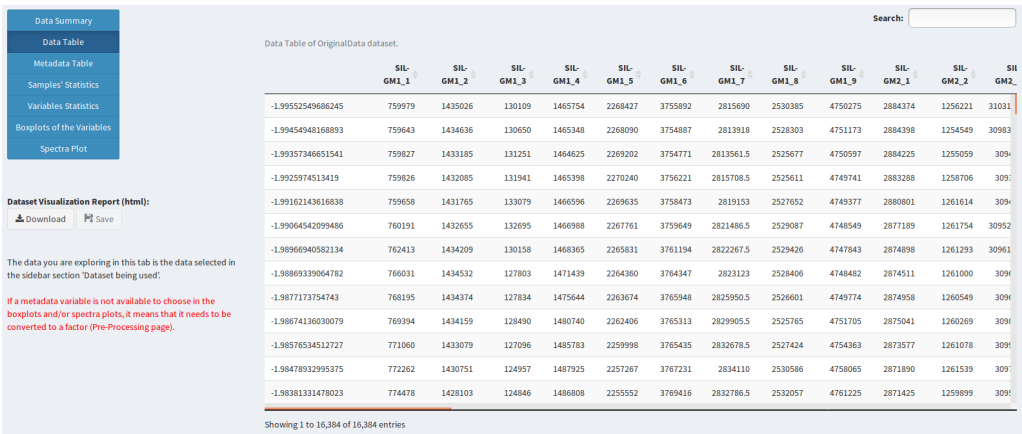
Contains information such as:

- Short description of the data that was provided by the user while submitting the project for analysis;
- Type of data;
- Number of samples, data points and metadata variables;
- XX and YY axis labels;
- Number of missing values;
- Statistics: mean, median, standard deviation, range and quantiles.



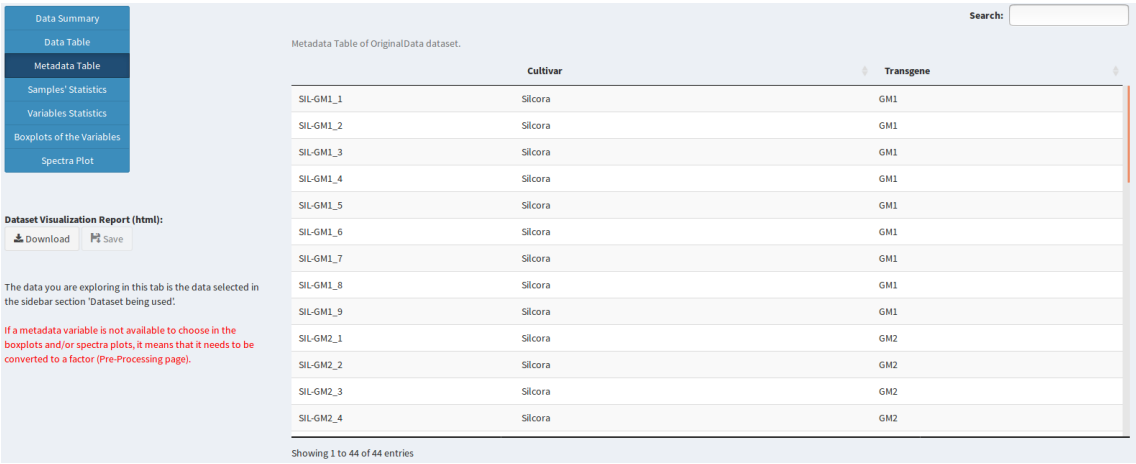

### 2.7.2 Data and Metadata Tables

The data section shows a table where each sample is represented by a column and each data variable by a row:




| | SIL-GM1_1 | SIL-GM1_2 | SIL-GM1_3 | SIL-GM1_4 | SIL-GM1_5 | SIL-GM1_6 | SIL-GM1_7 | SIL-GM1_8 | SIL-GM1_9 | SIL-GM2_1 | SIL-GM2_2 | SIL GM2_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -1.99552549686245 | 759979 | 1435026 | 130109 | 1465754 | 2268427 | 3755892 | 2815690 | 2530385 | 4750275 | 2884374 | 1256221 | 31031 |
| -1.99454948168893 | 759643 | 1434636 | 130650 | 1465348 | 2268090 | 3754887 | 2813918 | 2528303 | 4751173 | 2884398 | 1254549 | 30983 |
| -1.99357346651541 | 759827 | 1433185 | 131251 | 1464625 | 2269202 | 3754771 | 2813561.5 | 2525677 | 4750597 | 2884225 | 1255059 | 309 |
| -1.9925974513419 | 759826 | 1432085 | 131941 | 1465398 | 2270240 | 3756221 | 2815708.5 | 2525611 | 4749741 | 2883288 | 1258706 | 309 |
| -1.99162143616838 | 759658 | 1431765 | 133079 | 1466596 | 2269635 | 3758473 | 2819153 | 2527652 | 4749377 | 2880801 | 1261614 | 309 |
| -1.99064542099486 | 760191 | 1432655 | 132695 | 1466988 | 2267761 | 3759649 | 2821486.5 | 2529087 | 4748549 | 2877189 | 1261754 | 30952 |
| -1.98966940582134 | 762413 | 1434209 | 130158 | 1468365 | 2265831 | 3761194 | 2822267.5 | 2529426 | 4747843 | 2874898 | 1261293 | 30961 |
| -1.98869339064782 | 766031 | 1434532 | 127803 | 1471439 | 2264360 | 3764347 | 2823123 | 2528406 | 4748482 | 2874511 | 1261000 | 309 |
| -1.9877173754743 | 768195 | 1434374 | 127834 | 1475644 | 2263674 | 3765948 | 2825950.5 | 2526601 | 4749774 | 2874958 | 1260549 | 309 |
| -1.98674136030079 | 769394 | 1434159 | 128490 | 1480740 | 2262406 | 3765313 | 2829905.5 | 2525765 | 4751705 | 2875041 | 1260269 | 309 |
| -1.98576534512727 | 771060 | 1433079 | 127096 | 1485783 | 2259998 | 3765435 | 2832678.5 | 2527424 | 4754363 | 2873577 | 1261078 | 309 |
| -1.98478932995375 | 772262 | 1430751 | 124957 | 1487925 | 2257267 | 3767231 | 2834110 | 2530586 | 4758065 | 2871890 | 1261539 | 309 |
| -1.98381331478023 | 774478 | 1428103 | 124846 | 1486808 | 2255552 | 3769416 | 2832786.5 | 2532057 | 4761225 | 2871425 | 1259899 | 309 |

The metadata section shows a table where each sample is represented by a row metadata variable by a column:

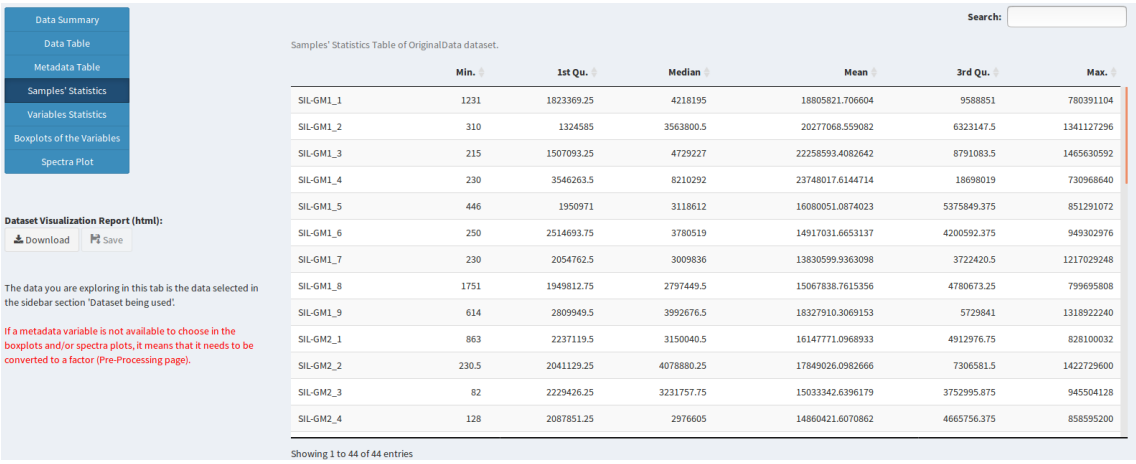



| | Cultivar | Transgene |
|---|---|---|
| SIL-GM1_1 | Silcora | GM1 |
| SIL-GM1_2 | Silcora | GM1 |
| SIL-GM1_3 | Silcora | GM1 |
| SIL-GM1_4 | Silcora | GM1 |
| SIL-GM1_5 | Silcora | GM1 |
| SIL-GM1_6 | Silcora | GM1 |
| SIL-GM1_7 | Silcora | GM1 |
| SIL-GM1_8 | Silcora | GM1 |
| SIL-GM1_9 | Silcora | GM1 |
| SIL-GM2_1 | Silcora | GM2 |
| SIL-GM2_2 | Silcora | GM2 |
| SIL-GM2_3 | Silcora | GM2 |
| SIL-GM2_4 | Silcora | GM2 |

The tables can be scrolled down and right if the size of the table is big. While scrolling down, the columns' names are fixed.

### 2.7.3 Variables and Samples Statistics

These two sections show a statistical summary for each variables and sample of the dataset, respectively. The statistical information consists on the minimum value, first quantile, median, mean, third quantile and maximum value.




| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| SIL-GM1_1 | 1231 | 1823369.25 | 4218195 | 18805821.706604 | 9588851 | 780391104 |
| SIL-GM1_2 | 310 | 1324585 | 3563800.5 | 20277068.559082 | 6323147.5 | 1341127296 |
| SIL-GM1_3 | 215 | 1507093.25 | 4729227 | 22258593.4082642 | 8791083.5 | 1465630592 |
| SIL-GM1_4 | 230 | 3546263.5 | 8210292 | 23748017.6144714 | 18698019 | 730968640 |
| SIL-GM1_5 | 446 | 1950971 | 3118612 | 16080051.0874023 | 5375849.375 | 851291072 |
| SIL-GM1_6 | 250 | 2514693.75 | 3780519 | 14917031.6653137 | 4200592.375 | 949302976 |
| SIL-GM1_7 | 230 | 2054762.5 | 3009836 | 13830599.9363098 | 3722420.5 | 1217029248 |
| SIL-GM1_8 | 1751 | 1949812.75 | 2797449.5 | 15067838.7615356 | 4780673.25 | 799695808 |
| SIL-GM1_9 | 614 | 2809949.5 | 3992676.5 | 18327910.3069153 | 5729841 | 1318922240 |
| SIL-GM2_1 | 863 | 2237119.5 | 3150040.5 | 16147771.0968933 | 4912976.75 | 828100032 |
| SIL-GM2_2 | 230.5 | 2041129.25 | 4078880.25 | 17849026.0982666 | 7306581.5 | 1422729600 |
| SIL-GM2_3 | 82 | 2229426.25 | 3231757.75 | 15033342.6396179 | 3752995.875 | 945504128 |
| SIL-GM2_4 | 128 | 2087851.25 | 2976605 | 14860421.6070862 | 4665756.375 | 858595200 |

### 2.7.4 Boxplots of the variables

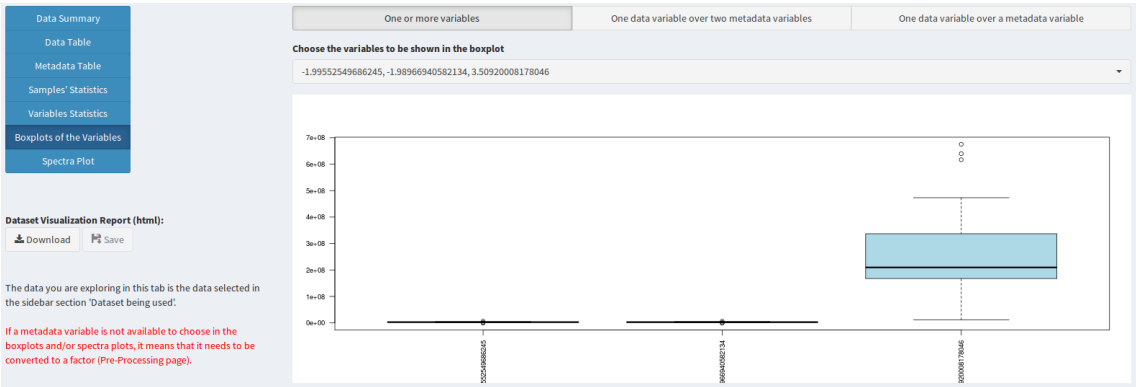
Three different types of boxplots are available. You can see a boxplot of one or more data variables, a boxplot of one data variable over two different metadata variables and a boxplot of one data variable over one metadata variable. Each of these plots can be accessed through their respective buttons that appear at the top of the page when the user clicks "Boxplots of the variables" at the left.

The second plot mentioned will not be available if the data in question does not have two or more metadata variables.

### One or more variables

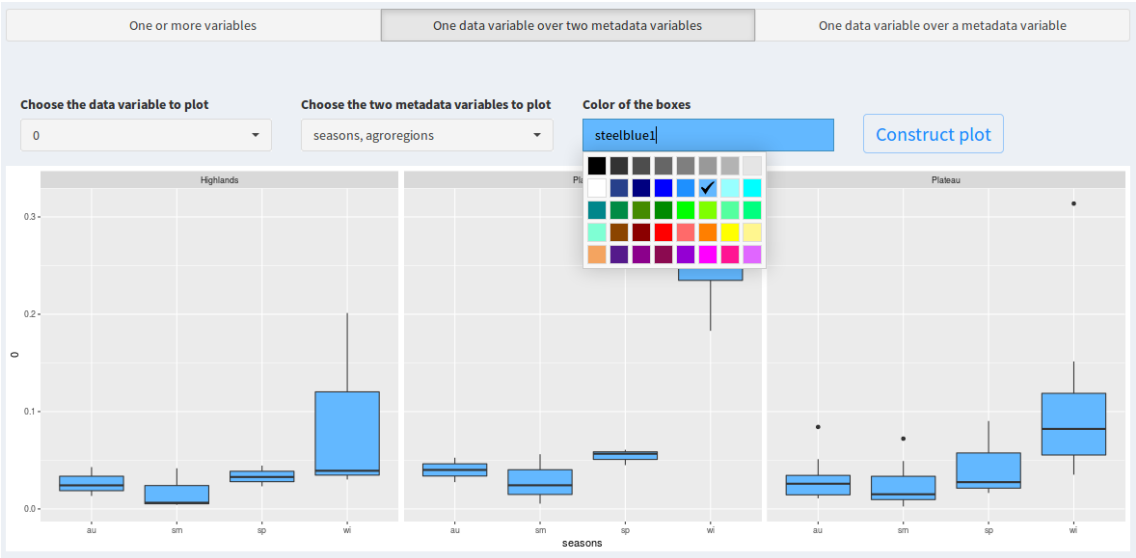
The variables to appear in the boxplot can be chosen by the user, through the select input above the plot.

So that all variables appear in the boxplot, you can click in the button "Select All" at the top of the choices in the selection input. However, if there are a lot of variables, the plot may not be readable.



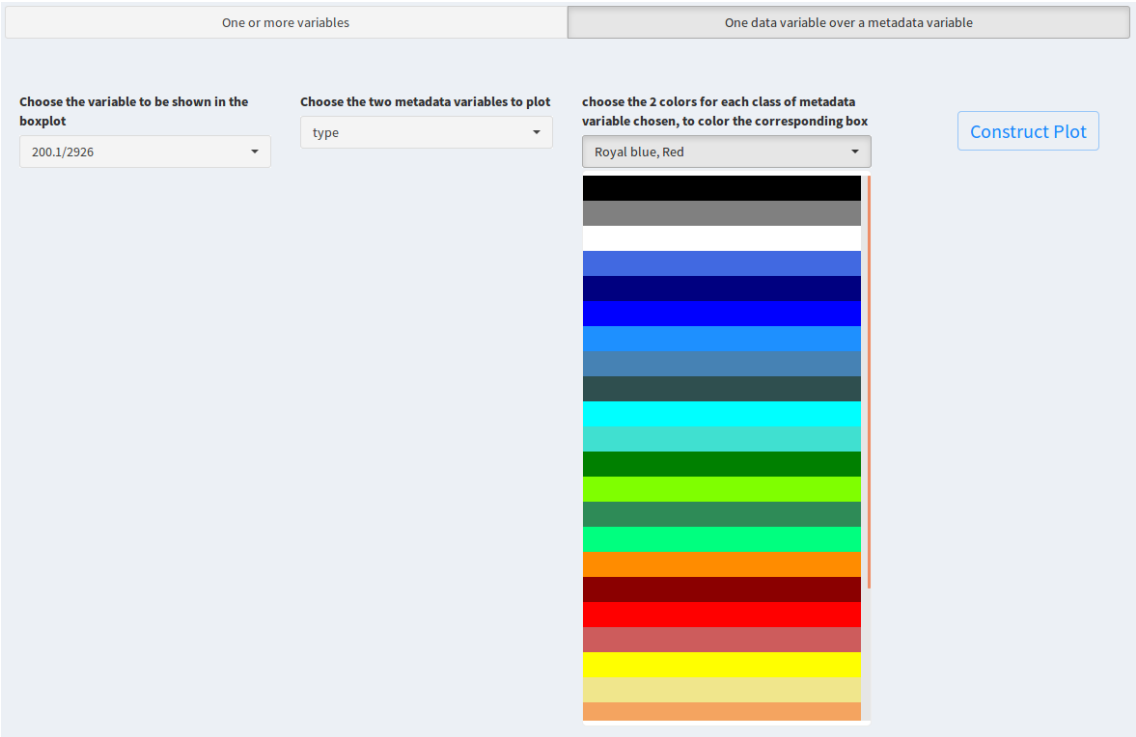

### One data variable over two metadata variables

To visualize this plot, you have to choose the data variable and the two metadata variables in their respective inputs and the color for the boxes in the plot. After that, you can click "Construct Plot" to see the new plot below.




### One data variable over a metadata variable

To visualize this plot, you must choose not only the data and metadata variables in their respective inputs, but also the different colors of the boxes that will appear in the plot (one for each metadata variable):




After setting the options, you can click "Construct Plot" and the new plot appears below the options:


### 2.7.5 Spectra/ Peaks plot

This plot is only available, as the name suggests, for datasets of spectral type or peaks data.

There are some changes that the user can perform to personalize the plot:

- The plot can show the spectra/peaks of one up to all samples, selected by the user. Again, if no samples are chosen in the select input, the spectra for all samples is plotted;
- Color the plot according to the different values of the metadata variable chosen;

- Choose the range of the xx axis;
- And choose if the values in this axis appear in descendent or ascendent order.

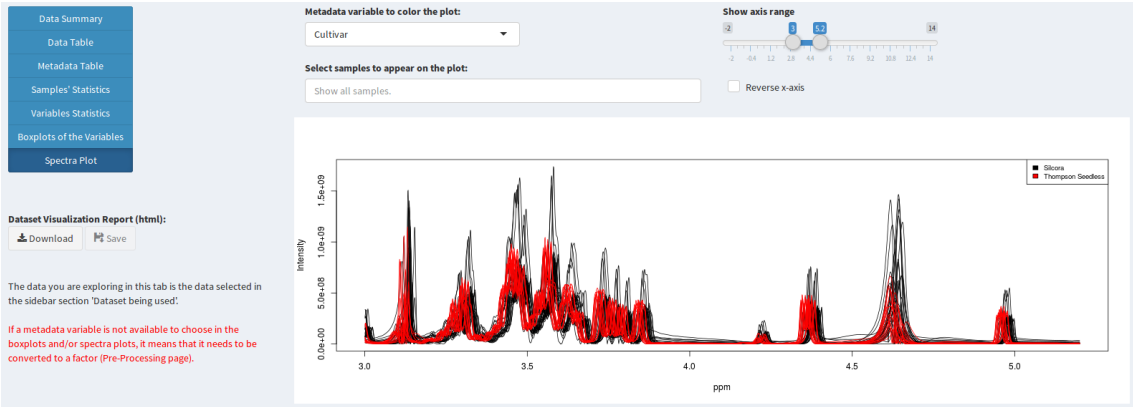



Figure 2.24: Spectra Plot section.

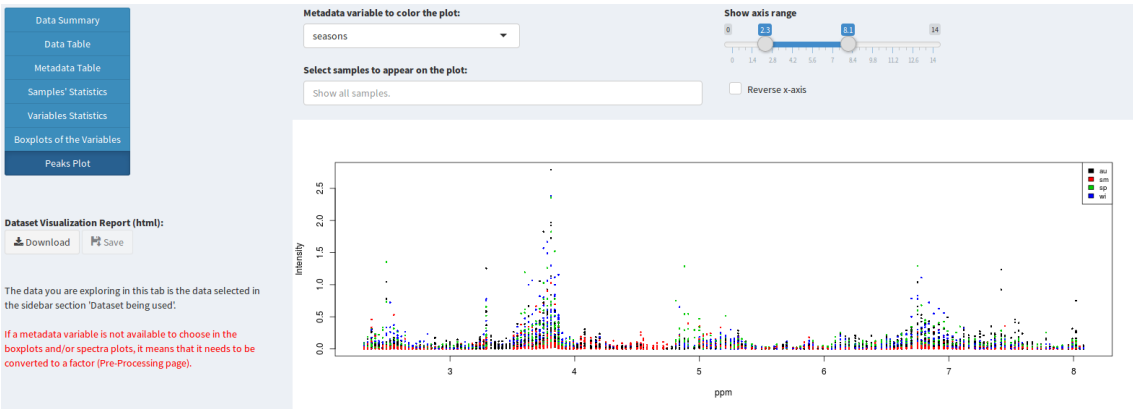



Figure 2.25: Peaks Plot section.

### 2.7.6 Get a report of the data visualization

The plots appear in the report like they are in the page in the moment of the report file creation. This means that only the variables that are selected to appear in the boxplot at the time of the report file creation will appear in the file plot, for example, and the same happens with other inputs that may change the plots.

If the data being visualized is of concentrations type, no spectra/peaks plot will be present in the report, similarly to what happens in the website:

## Data Visualization Report

```
## Report generated on 2018-06-08 11:54:34 using WEBSPECMINE
```

### Dataset

OriginalData

### Dataset Summary

```
## Dataset summary:
## Valid dataset
## Description:
## Type of data:  concentrations
## Number of samples:  77
## Number of data points 63
## Number of metadata variables:  1
## Label of x-axis values:  Compounds
## Label of data points:  Concentrations
## Number of missing values in data:  0
## Mean of data values:  347.3735
## Median of data values:  51.42
## Standard deviation:  1500.838
## Range of values:  0.79 33860.35
## Quantiles:
##       0%      25%      50%      75%     100%
##     0.79    17.46    51.42   160.77 33860.35
```

### Data Table

Search:

| | PIF_178 | PIF_087 | PIF_090 | NETL_005_V1 | PIF_115 | PIF_110 | NETL_019_V1 | N |
|---|---|---|---|---|---|---|---|---|
| 1.6-Anhydro-beta-D-glucose | 40.85 | 62.18 | 270.43 | 154.47 | 22.2 | 212.72 | 151.41 | |
| 1-Methylnicotinamide | 65.37 | 340.36 | 64.72 | 52.98 | 73.7 | 31.82 | 36.6 | |
| 2-Aminobutyrate | 18.73 | 24.29 | 12.18 | 172.43 | 15.64 | 18.36 | 8.67 | |
| 2-Hydroxyisobutyrate | 26.05 | 41.68 | 65.37 | 74.44 | 83.93 | 80.64 | 42.52 | |
| 2-Oxoglutarate | 71.52 | 67.36 | 23.81 | 1199.91 | 33.12 | 47.94 | 223.63 | |
| 3-Aminoisobutyrate | 1480.3 | 116.75 | 14.3 | 555.57 | 29.67 | 17.46 | 56.26 | |
| 3-Hydroxybutyrate | 56.83 | 43.82 | 5.64 | 175.91 | 76.71 | 31.82 | 11.59 | |
| 3-Hydroxyisovalerate | 10.07 | 79.84 | 23.34 | 25.03 | 69.41 | 35.16 | 25.79 | |

Showing 1 to 63 of 63 entries

### Metadata Table

Search:

| | Muscle.loss |
|---|---|
| PIF_178 | cachexic |
| PIF_087 | cachexic |
| PIF_090 | cachexic |
| NETL_005_V1 | cachexic |
| PIF_115 | cachexic |
| PIF_110 | cachexic |
| NETL_019_V1 | cachexic |
| NETCR_014_V1 | cachexic |
| NETCR_014_V2 | cachexic |

Showing 1 to 77 of 77 entries

### Samples' Statistics

Search:

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| PIF_178 | 5.58 | 52.72 | 154.47 | 699.855714285714 | 416.235 | 16481.6 |
| PIF_087 | 7.69 | 78.66 | 208.51 | 708.302380952381 | 412.095 | 15835.35 |
| PIF_090 | 4.44 | 31.5 | 141.17 | 771.794444444444 | 308.03 | 24587.66 |
| NETL_005_V1 | 25.03 | 102.51 | 247.15 | 1021.28206349206 | 673.705 | 20952.22 |
| PIF_115 | 4.53 | 44.26 | 84.77 | 441.219682539683 | 196.615 | 6836.29 |
| PIF_110 | 5.05 | 35.34 | 113.3 | 537.475079365079 | 325.58 | 15677.78 |
| NETL_019_V1 | 2.1 | 26.725 | 91.84 | 400.849206349206 | 223.63 | 8022.46 |
| NETCR_014_V1 | 1.73 | 7.14 | 18.17 | 82.7695238095238 | 52.525 | 2208.35 |
| NETCR_014_V2 | 2.41 | 14.63 | 39.65 | 207.801904761905 | 102 | 6634.24 |

Showing 1 to 77 of 77 entries

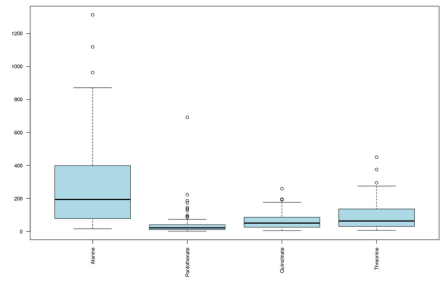
### Variables' Statistics

Search:

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| 1.6-Anhydro-beta-D-glucose | 4.71 | 28.79 | 45.6 | 105.63038961039 | 141.17 | 685.4 |
| 1-Methylnicotinamide | 6.42 | 15.8 | 36.6 | 71.5736363636364 | 73.7 | 1032.77 |
| 2-Aminobutyrate | 1.28 | 5.26 | 10.49 | 18.159740259740 3 | 19.49 | 172.43 |
| 2-Hydroxyisobutyrate | 4.85 | 15.8 | 32.46 | 37.2506493506494 | 54.6 | 93.69 |
| 2-Oxoglutarate | 5.53 | 22.42 | 55.15 | 145.087142857143 | 92.76 | 2465.13 |
| 3-Aminoisobutyrate | 2.61 | 11.7 | 22.65 | 76.7563636363636 | 56.26 | 1480.3 |
| 3-Hydroxybutyrate | 1.7 | 5.99 | 11.7 | 21.717012987013 | 29.96 | 175.91 |
| 3-Hydroxyisovalerate | 0.92 | 5.26 | 12.55 | 21.5477922077922 | 30.27 | 164.02 |
| 3-Indoxylsulfate | 27.66 | 82.27 | 144.03 | 218.879220779221 | 333.62 | 1043.15 |

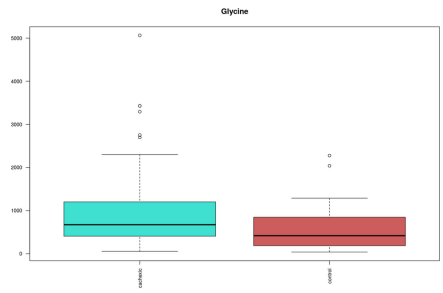
Showing 1 to 63 of 63 entries

### Boxplot of the Variables




### Data variable Glycine over metadata variable Muscle.loss

1




## End of report

The user will always be able to download the report to their computer, but can only save it if logged in the account.

## 2.8 Run an Analysis

The "Run Analysis" feature is available through the header panel and leads to the page that allows the user to do the analysis of the datasets.

There are a total of 7 boxes of analysis provided.

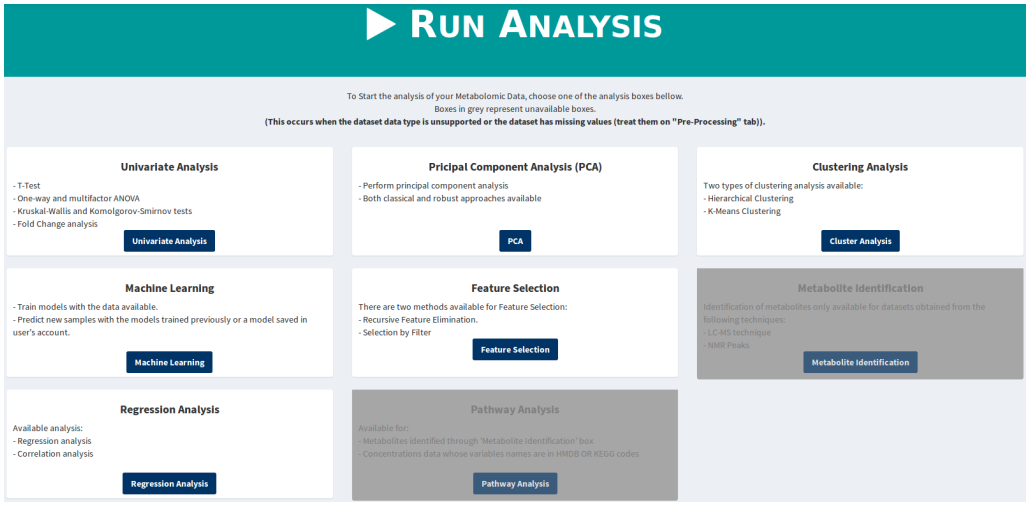



Figure 2.27: Run Analysis page layout.

- ***Metabolite Identification***, only available for spectral data from the LC-MS technique or peaks lists data from the NMR technique;
- ***Pathway Analysis***, only for concentrations data whose metabolites are represented by KEGG or HMDB codes, or for identified metabolites from NMR or LC-MS data;
- ***Univariate Analysis***, where t-tests, one-way and multifactor analysis of variance (ANOVA), Kruskal-Wallis and Komolgorov-Smirnov tests, and fold change analysis can be done;
- ***Regression Analysis***, where regression and correlation analysis are made available.
- ***PCA***, both classical and robust approaches;
- ***Clustering Analysis***, where hierarchical and k-means clustering are available;
- ***Machine Learning***, where it is possible to train models and predict new samples;
- ***Feature Selection***, where two methods are available, namely recursive feature elimination and selection by filter;

The analysis boxes might not be accessible if the dataset currently in use contains missing values. Another example is the box "Metabolite Identification", which might also be unaccessible if the dataset type is not supported, i.e., if the dataset is not spectral data from the LC-MS technique or NMR peaks lists data. In these cases, the respective boxes remain in grey, inaccessible, until the desired conditions are met.

All analyses done must have a name associated to them, given by the user. These names must differ, although the input text box where this name has to be given comes with a default value.

### 2.8.1 Univariate Analysis

Regarding univariate data analysis, the web application is able to perform either one-way or multi-factor ANOVA, T-Tests, Kruskal-Wallis and Kolmogorov-Smirnov tests, and fold change analysis.

Therefore, after entering the Univariate Analysis page, the user sees a sidebar panel with each tab leading to the respective type of analysis mentioned.

After clicking in of these tabs, the options that must be given to perform the analysis are shown at the right of this tab.

By default, the options regarding the first type of analysis in the sidebar panel appear automatically when first entering the page.

### T-Test

The available **options** to set are:

- Analysis Name;
- Metadata variable to use (to create the groups of samples based on the different values of the selected variable, that will be compared between each other);
- P-value Treshold: defaults to 0.01.

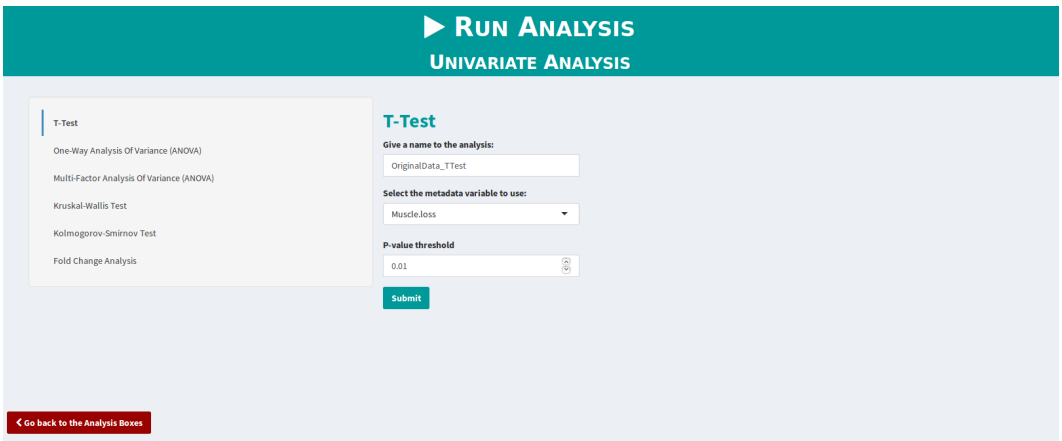



Figure 2.28: T-Test analysis page layout.

### One-way Analysis of Variance (ANOVA)

The available **options** to set are:

- Analysis Name;
- Metadata variable to use (to create the groups of samples based on the different values of the selected variable, that will be compared between each other);
- If the TukeyHSD test should also be performed, alongside with ANOVA.

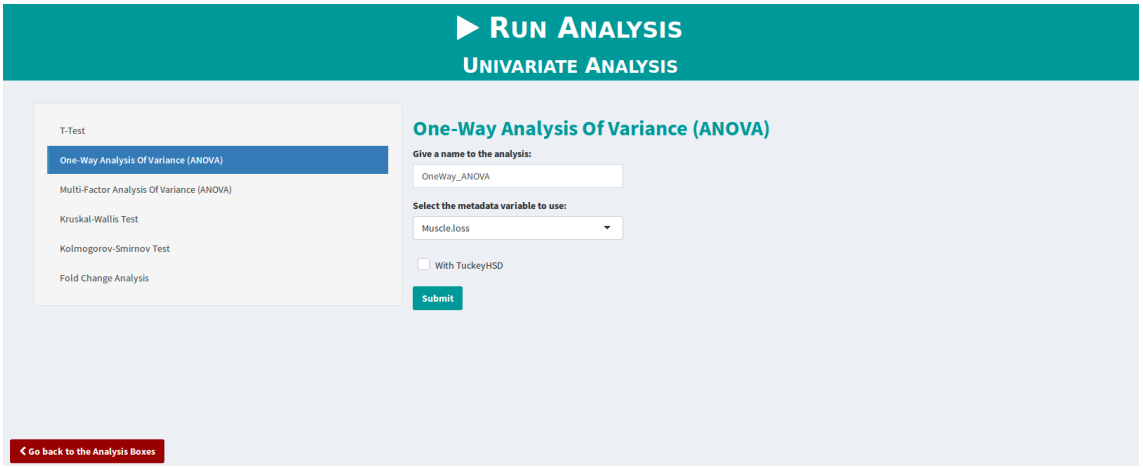



Figure 2.29: One-Way ANOVA analysis page layout.

### Multi-factor Analysis of Variance (ANOVA)

As this type of analysis can only be performed on datasets with more than one metadata variables, this analysis won't be available for datasets that do not fill this requirement.

The available **options** to set are:

- Analysis Name;
- Metadata variables to use (to create the groups of samples based on the different values of the selected variables, that will be compared between each other);
- Write the formula specifying the model, using the names of the metadata variables chosen.

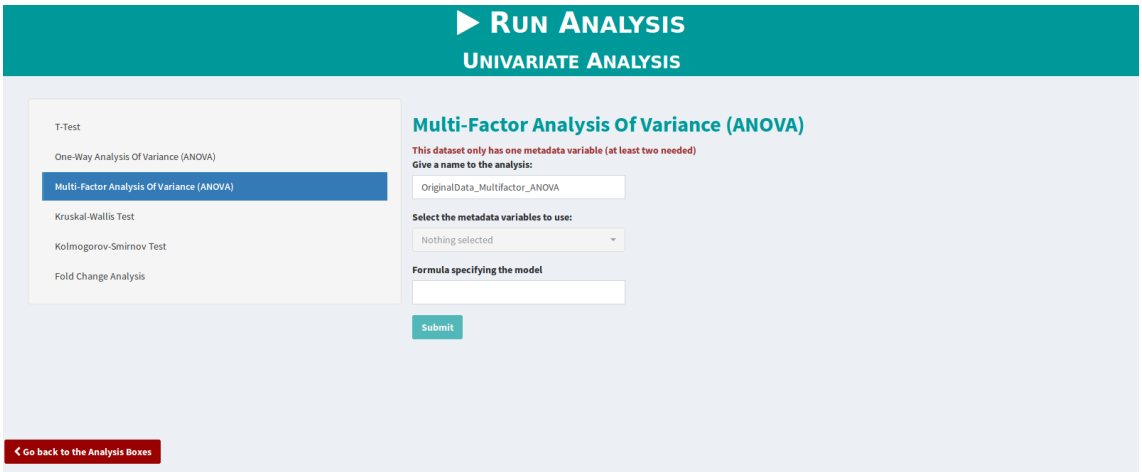



Figure 2.30: Multi-Factor ANOVA analysis page layout.

### Kruskal-Wallis Test

The available **options** to set are:

- Analysis Name;
- Metadata variable to use (to create the groups of samples based on the different values of the selected variable, that will be compared between each other);
- P-value treshold.

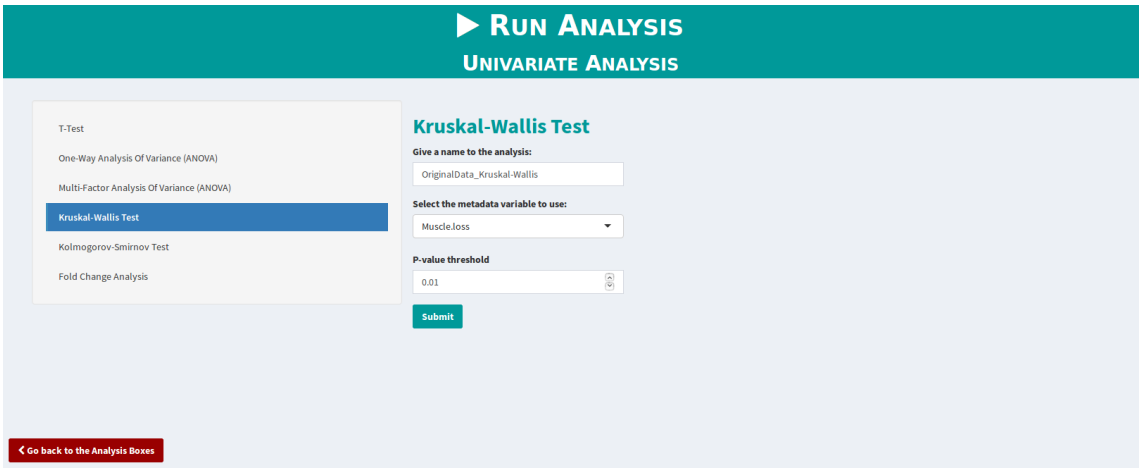



Figure 2.31: Kruskal-Wallis Test page layout.

### Kolmogorov-Smirnov Test

The available **options** to set are:

- Analysis Name;
- Metadata variable to use (to create the groups of samples based on the different values of the selected variable, that will be compared between each other);
- P-value treshold.

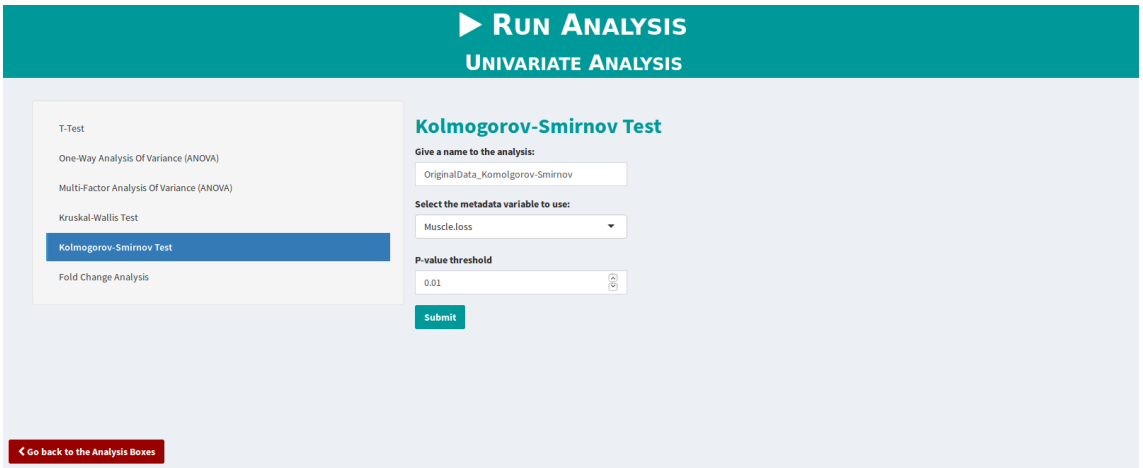



Figure 2.32: Kolmogorov-Smirnov Test page layout.

### Fold Change Analysis

The available **options** to set, in order to perform fold change analysis on the entire dataset (difference of the variables on two groups) are:

- Analysis Name;
- Metadata variable to use (to create the groups of samples based on the different values of the selected variable, that will be compared between each other);
- One of the possible values of the metadata variable chosen, to use as reference value.

If the user choose to perform an additional fold change analysis on two variables (difference of the groups on two variables), the following options must be set:

- Select the two data variables to use.

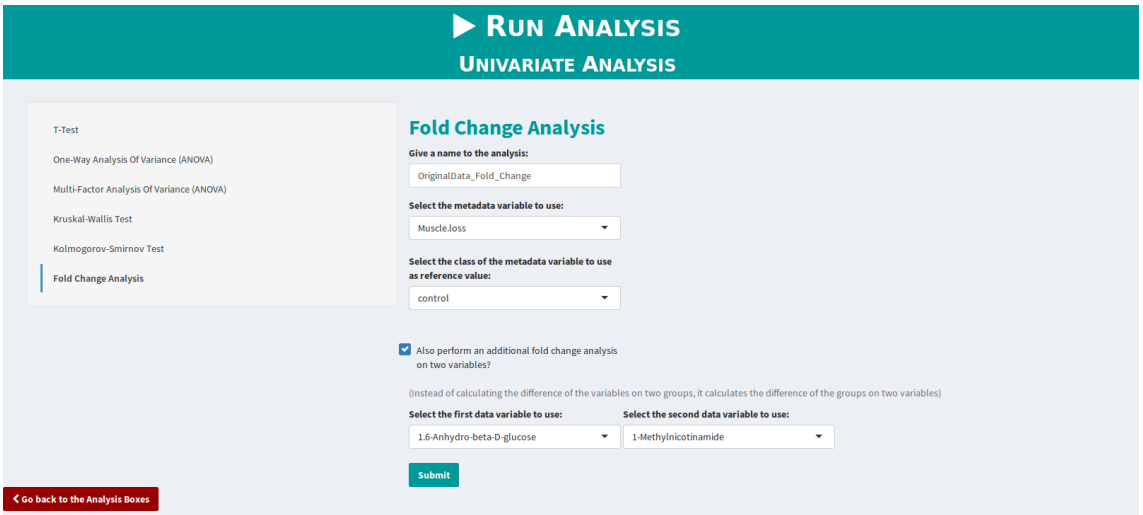



Figure 2.33: Fold Change analysis page layout.

## 2.8.2 Principal Components Analysis (PCA)

Both classical and robust PCA are available to perform.

### Normal PCA

The following **options** must be set in order for the analysis to be performed:

- Analysis Name;
- Choose if variables should be scaled and/or centered.

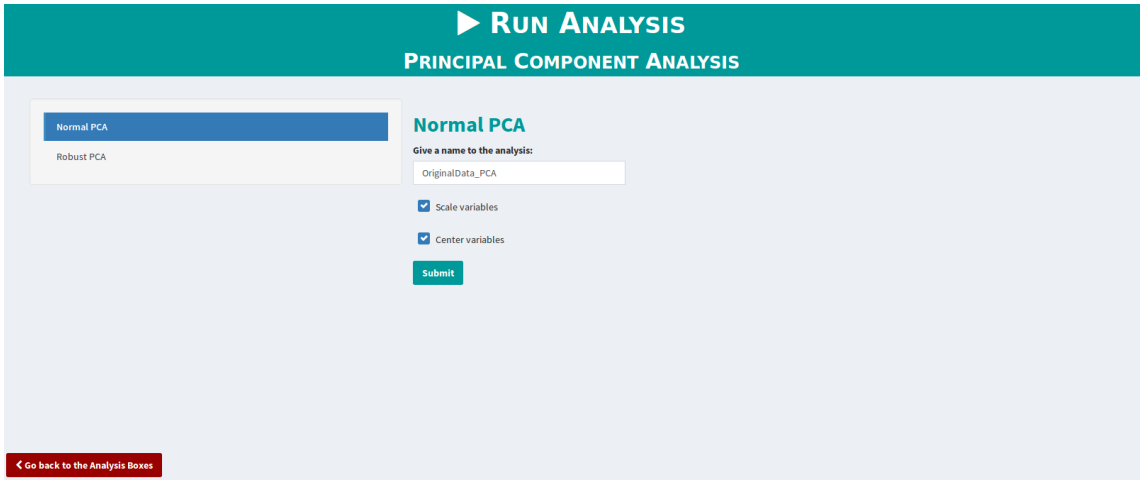



Figure 2.34: Normal PCA analysis page layout.

### Robust PCA

The following **options** must be set in order for the analysis to be performed:

- Analysis Name;
- Choose the method by which to center the variables: "Mean" or "Median";
- Choose the method by which to scale the variables: "Standard deviation ratio" or "Mean absolute deviation";
- Number of components to obtain.

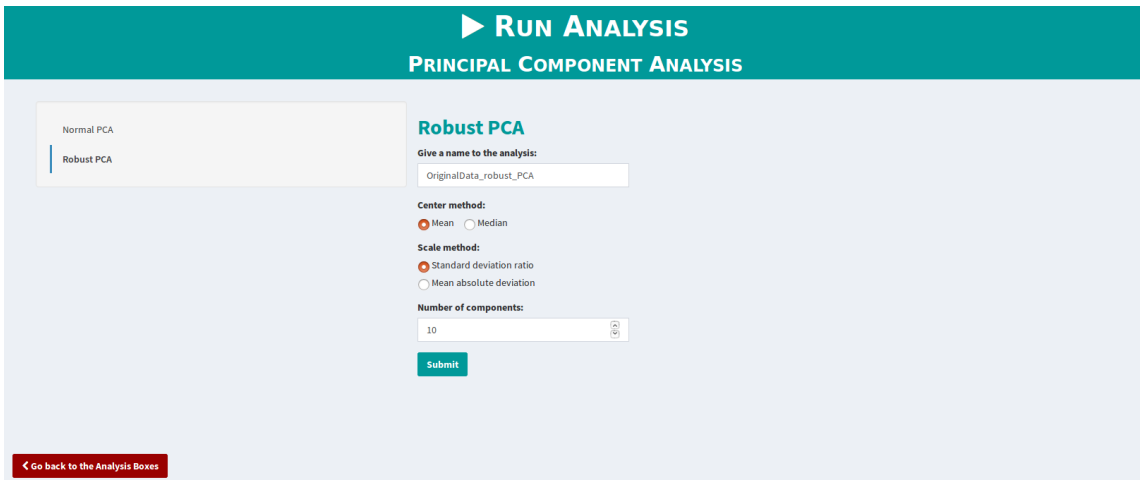



Figure 2.35: Robust PCA analysis page layout.

### 2.8.3 Clustering Analysis

Both hierarchical and K-means clustering are available.

#### Hierarchical Clustering

The **options** to set here are:

- Analysis Name;
- Distance Measaure: "Euclidean", "Manhattan", "Pearson" or "Spearman";
- Agglomeration method: "Complete", "Ward", "Single", "Average", "Mcquitty", "Median" or "Centroid";
- If distance should be calculated between samples or variables;

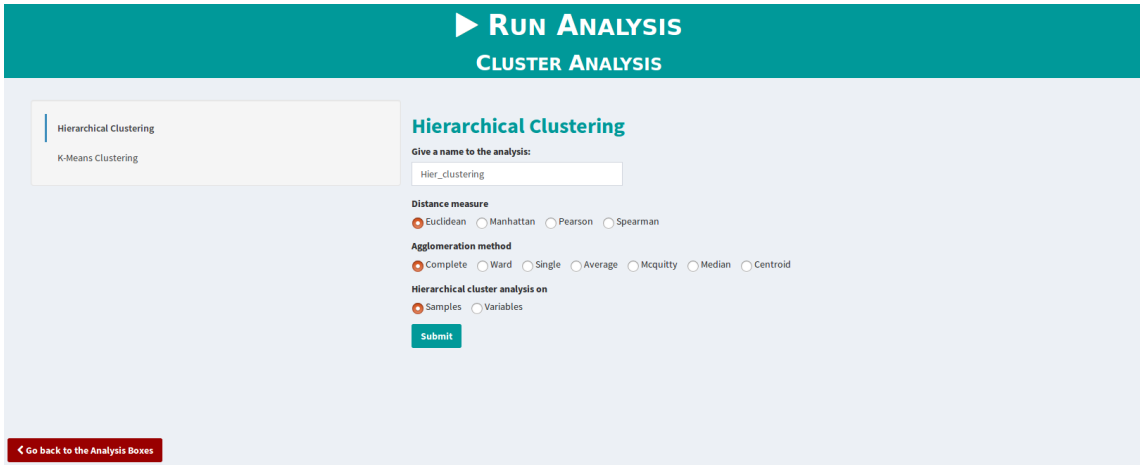



Figure 2.36: Hierarchical Clustering analysis page layout.

#### K-Means Clustering

The only **options** to set in this case are:

- Analysis Name;
- If distance should be calculated between samples or variables;
- The number of clusters in which to group the samples/variables;

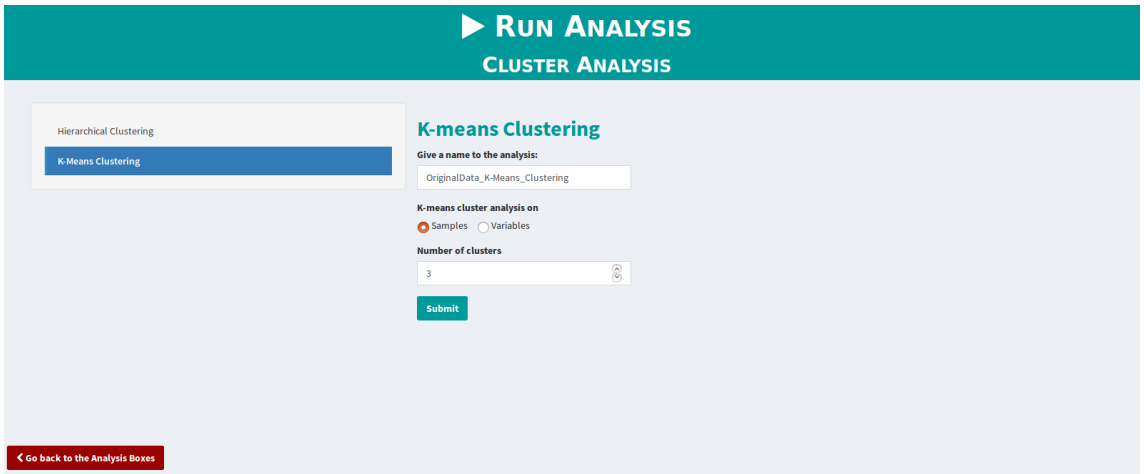



Figure 2.37: K-Means Clustering analysis page layout.

### 2.8.4 Machine Learning

At the top of the Machine Learning page, there are two buttons that can lead to either model training, if the button "Train Models" is clicked, or to samples prediction, if the button "Predict New Samples" is chosen. The options that can be chosen for each type of analysis appear below these buttons.

The "Predict New Samples" option is not accessible if no model was previously trained using the dataset currently in used.

#### Model Training

Model training is only available for classification models.

The **options** available to set are:

- Analysis Name;
- Choose one or more types of models to train: "Partial Least Squares (PLS)", "Decision Tree (C4.5)", "Rule-Based Classifier", "Support Vector Machine (SVM) with linear kernel", "Random Forests", "Linear Discriminant Analysis (LDA)" and "Neural Network";
- Name of the metadata variable to predict;

As regards to the **options for parameter optimization**, the user can either choose:

- Give all the values that will be tested for each parameter of each chosen model;
- Or define the number of different values that will be tested for each parameter, whose values will be set automatically.

For the **model validation options**, the following have to be set:

- Method: "Resampling", "Cross-validation", "Repeated Cross-validation", "Leave One Out Cross-validation" and "Leave Group-out Cross-validation";
- Number of resampling iterations: if "Resampling" method is chosen;
- Number of validation folds: if any of the other methods is chosen;
- Number of repeats: if the selected method is "Repeated Cross-validation";
- Validation Metric: "Accuracy" or "ROC".

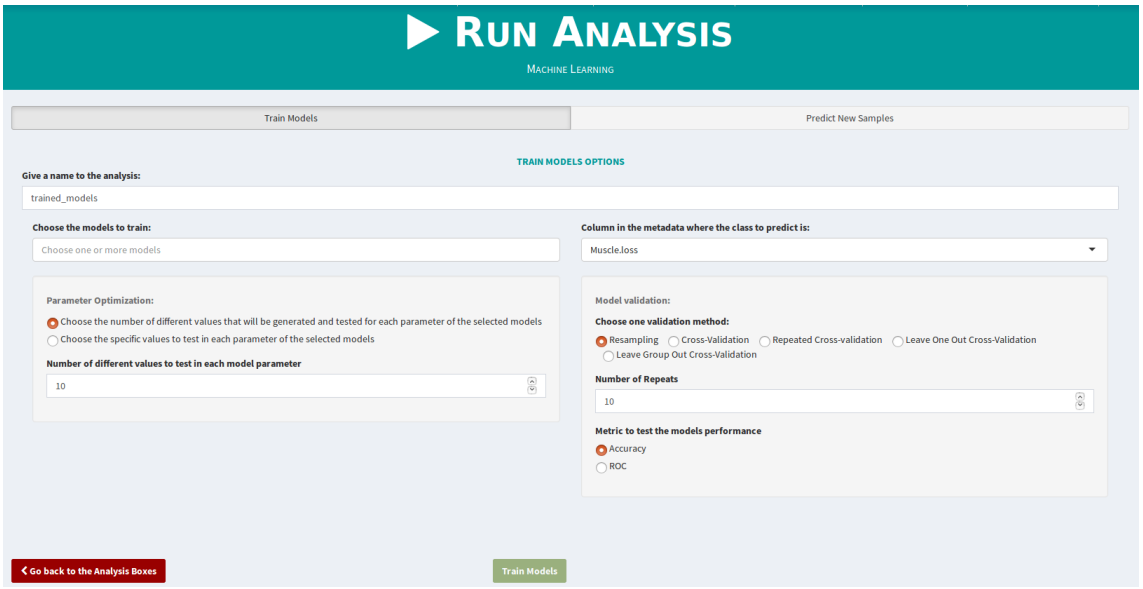
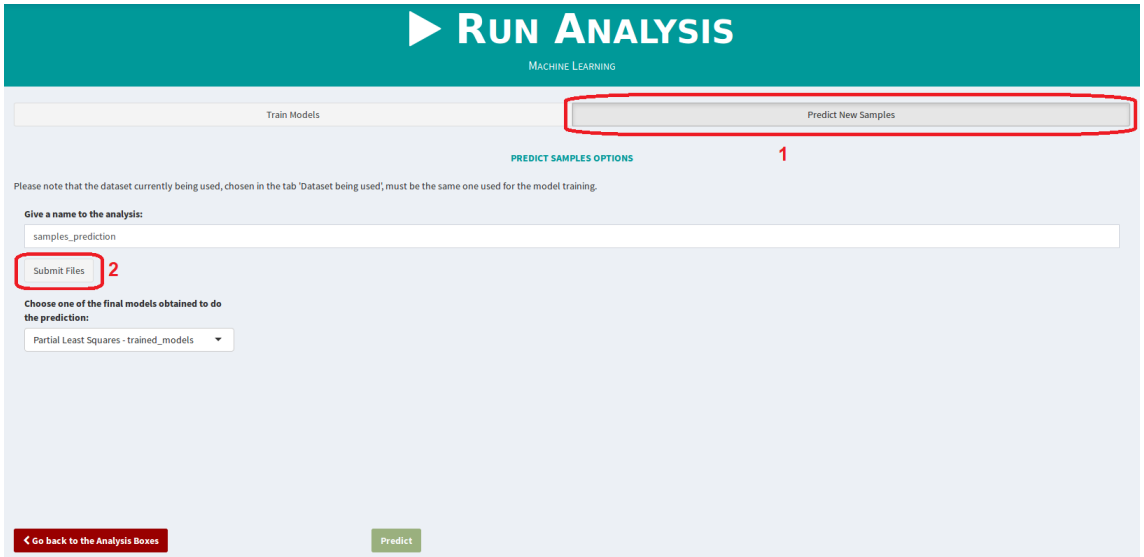



Figure 2.38: Train Models analysis page layout.

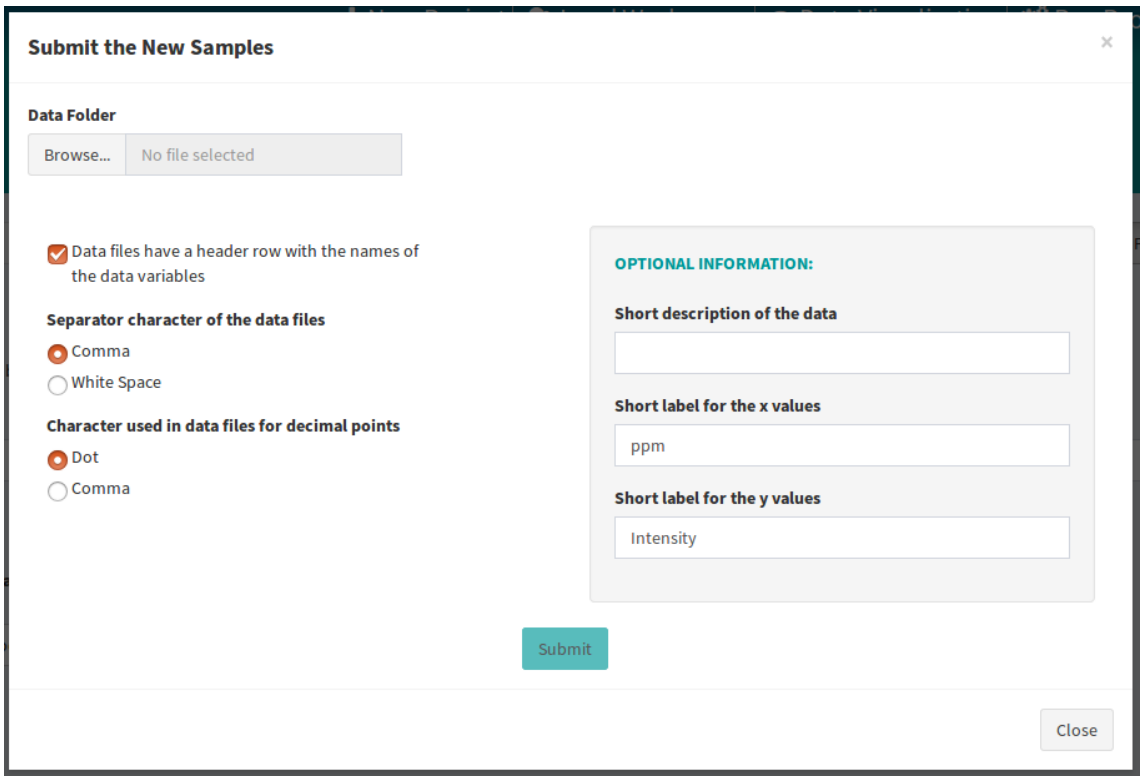
### Sample Prediction

First, the user should **submit the new samples file(s)**:

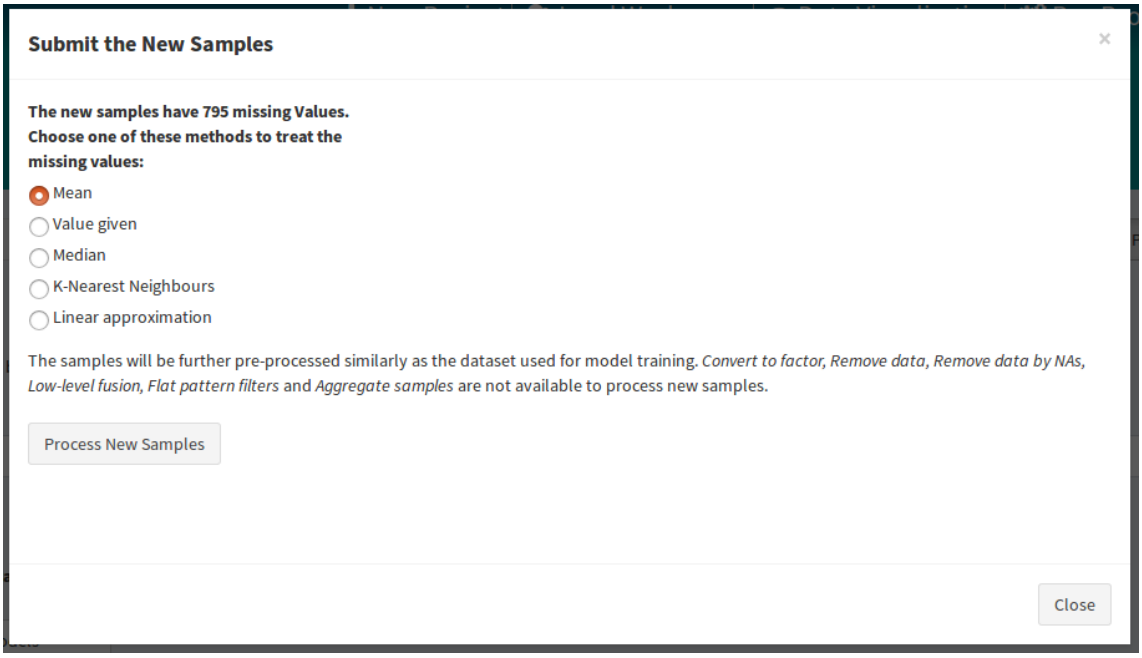
1. Click the button to start the submission of the file(s);




2. A pop-up window will appear, with the options to process the data file(s), according to the type of data that must be submitted (the same type of data used to train the models);

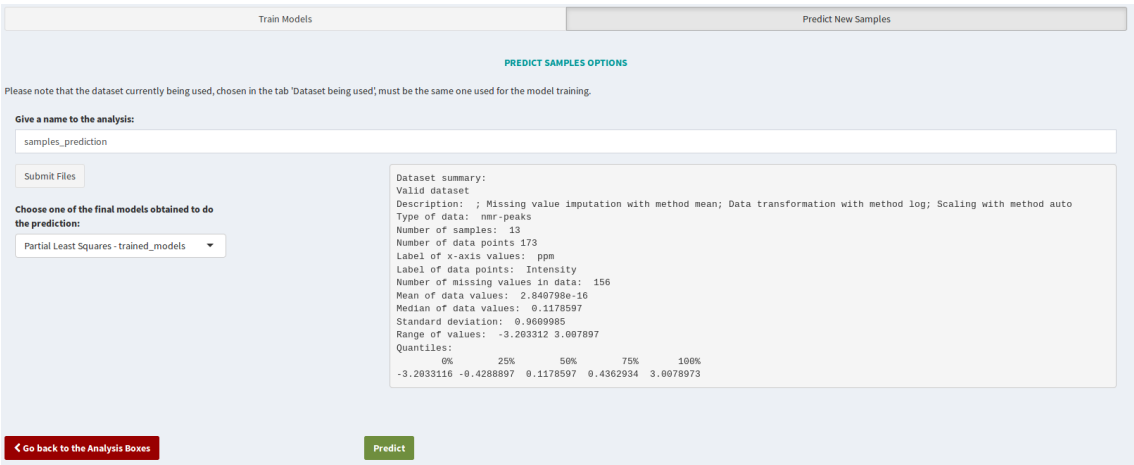



3. After this, the user will be asked to treat the missing values, with the same options present in the pre-processing page, if the new data has missing values;




4. Lastly, the data will be further pre-processed similarly to the data used to train the model chosen: "Convert to factor", "Remove data", "Remove data by NAs", "Low-level fusion", "Flat pattern filters" and "Aggregate samples" are not available to process new samples.

After submitting the new samples, a brief summary will appear at the bottom right of the page:




The **options** available are:

- Analysis Name;
- Choose a model to perform the prediction: only the models trained making use of the dataset currently in use will be made available to choose.

### 2.8.5 Feature Selection

The available **options** to set are:

- Analysis Name;
- Choose the metadata variable to be predicted;
- Method: "Recursive Feature Elimination" or "Selection by Filter";

- Function for model fitting, prediction and variable importance/filtering: "Random Forests", "Linear Regression", "Bagged Trees", "Linear Discriminant Analysis (LDA)" and "Naive-Bayes";

For the **model validation options**, the user must set:

- Validation Method: "Resampling", "Cross-validation", "Repeated Cross-validation", "Leave One Out Cross-validation" and "Leave Group-out Cross-validation";
- Number of resampling iterations: if "Resampling" method is chosen;
- Number of validation folds: if any of the other methods is chosen;
- Number of repeats: if the selected method is "Repeated Cross-validation".

The user can also choose if he wants to manually set the number of features that will be tested in each group test. If the user chooses to do so, he must give the size of each group test, separated by a comma. If not, the groups' sizes will be generated by default.

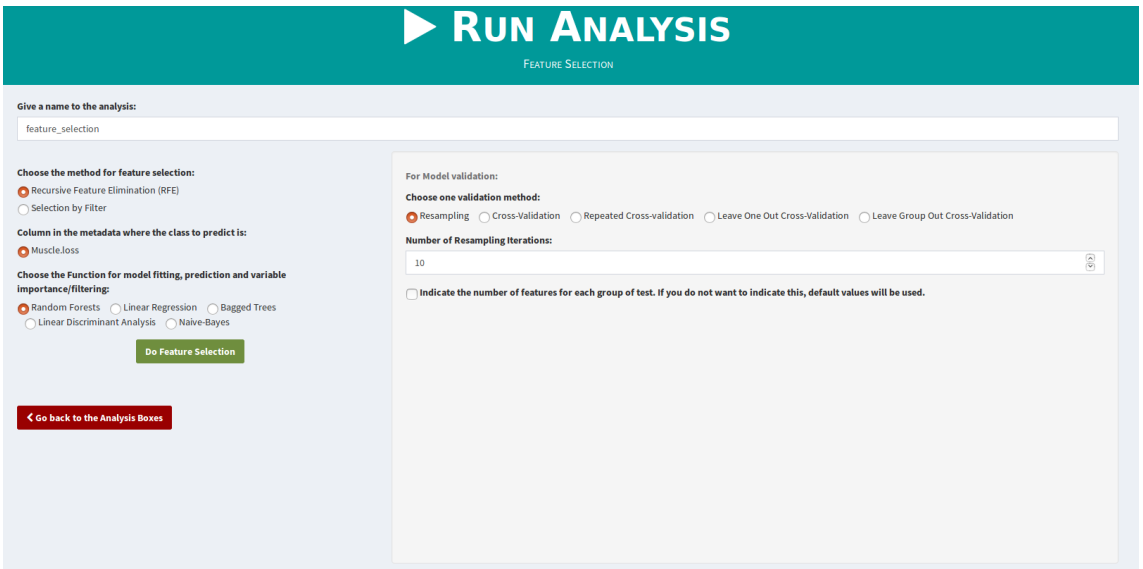



Figure 2.39: Feature Selection analysis page layout.

## 2.8.6 Metabolite Identification

The user can only perform identification of metabolites on data from LC-MS spectra or NMR Peaks. When entering the "Metabolite Identification" box, the available options will differ according to the type of data in question.

### LC-MS Spectra

The overall pipeline for identification of metabolites from this type of data starts starts with the detection of the existing peaks in the spectra, followed by discrimination of which peaks belong to the same source metabolite and, finally, each of these groups of peaks, which have a certain mass and were acquired under a certain chemical environment (ionization mode, for example), are compared to the peaks of each metabolite on a predefined database. This analysis is performed by using the *MAIT* R package.

The **options** that can be set are:

- Analysis Name;
- Column of the metadata that may help in the identification.

All the other parameters are already set by default and the user cannot change them, like the peak tolerance and mass tolerance ones, which are set to 0.005 and 0.5, respectively.

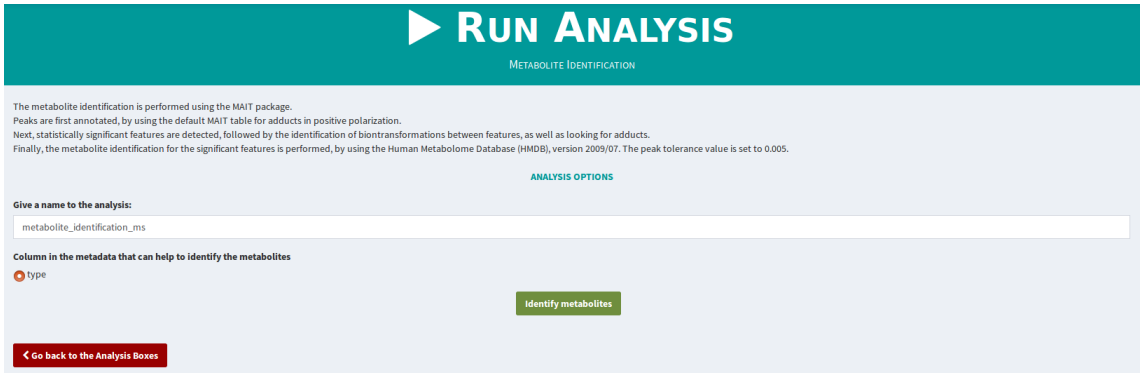



Figure 2.40: Overall layout of the LC-MS metabolite identfication results page.

### NMR Peaks

The overall pipeline for the NMR metabolite identification starts with the clustering of the peaks in the dataset according to a correlation. After clustering, the peaks are separated in the respective clusters based on a minimum correlation that each peak inside a cluster must have with the others on the same cluster. The value of this correlation can be set by the user or calculated, where the optimal value is the one that leads to the larger number of clusters.

Each of these clusters is considered a potential metabolite, as it can be assumed that peaks coming from the same molecule show similar behaviour across all samples and, therefore, correlate strongly with each other.

After setting the library of the reference metabolites, each cluster is compared with each reference metabolite, using the Jaccard index to score the match. This index is used to compare the similarity between sets, as it is defined by the division of the size of intersection by the size of the union of the sets: $J(A,B) = |A \cap B| \div |A \cup B|$.

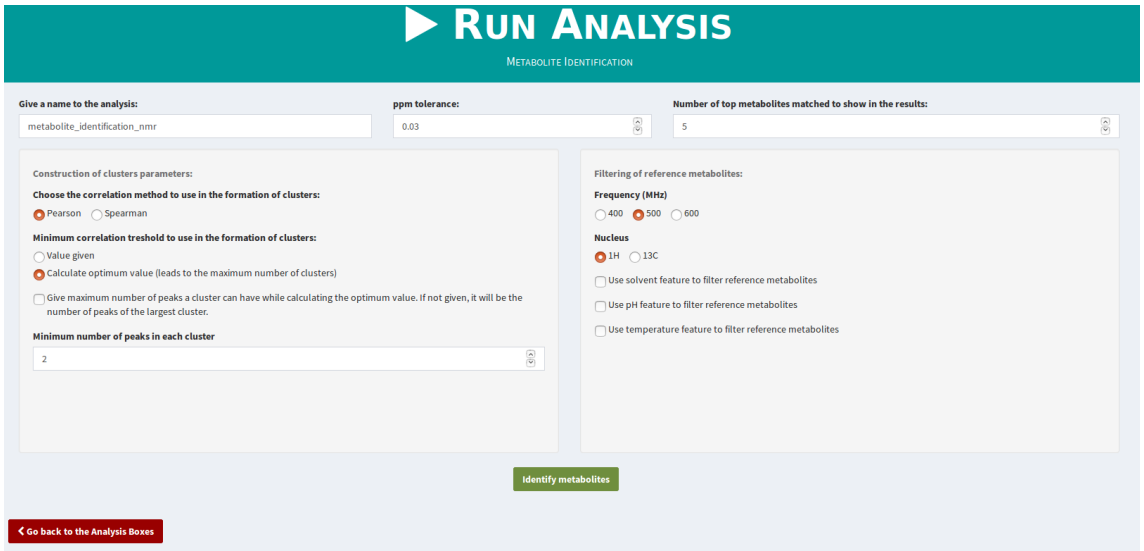



Figure 2.41: NMR metabolite identification analysis page layout.

The **options** that must be set are:

- Analysis Name;
- ppm tolerance when matching between cluster and reference peaks;
- Number of top metabolites matched to each cluster to show;

For the **construction of clusters**, the options to set are:

- Correlation Method: "Pearson" or "Spearman";
- Minimum number of peaks that the clusters must have, defaults to 40;
- If the minimum correlation value in the formation of clusters must be given by the user or calculated;
- If the above correlation value is calculated by the website, the user can give the maximum number of peaks a cluster can have while searching for this optimum value or let the website set it to the number of peaks of the biggest reference metabolite.

Finally, to **filter the reference metabolites**, the options made available are:

- Frequency: can either be 400, 500 or 600;
- Nucleus: can either be "1H" or "13C";
- Solvent, optional: "100% DMSO", "5% DMSO", "Acetone + DMSO + Tetramethylurea", "C", "CCl4", "CD3OD", "CDCl3", "Cyclohexane", "D2O", "DMSO-d6", "DMSO-d6 + HCl", "Neat", "TMS", or "Water";
- pH interval or value, optional: the user can choose the minimum and maximum values, or one single value of pH, by setting both values as the same value;
- Temperature, optional: can either be 25$^{o}$C or 50$^{o}$C.

While setting the filtering parameters, the website checks if the combination of the chosen parameters lead or not to no reference metabolites. If so, the warning message "There are no reference metabolites with all the features selected" appears under the box to alert the user and disables the button to do the identification.

## 2.8.7 Regression Analysis

### Linear Regression Analysis

As this type of analysis can only be performed on datasets with more than one metadata variables, this analysis won't be available for datasets that do not fill this requirement.

The available **options** to set are:

- Analysis Name;
- Metadata variables to use: to create the groups of samples based on the different values of the selected variables, that will be compared between each other for each data variable (one data variable - one linear regression)
- Write the formula specifying the model.

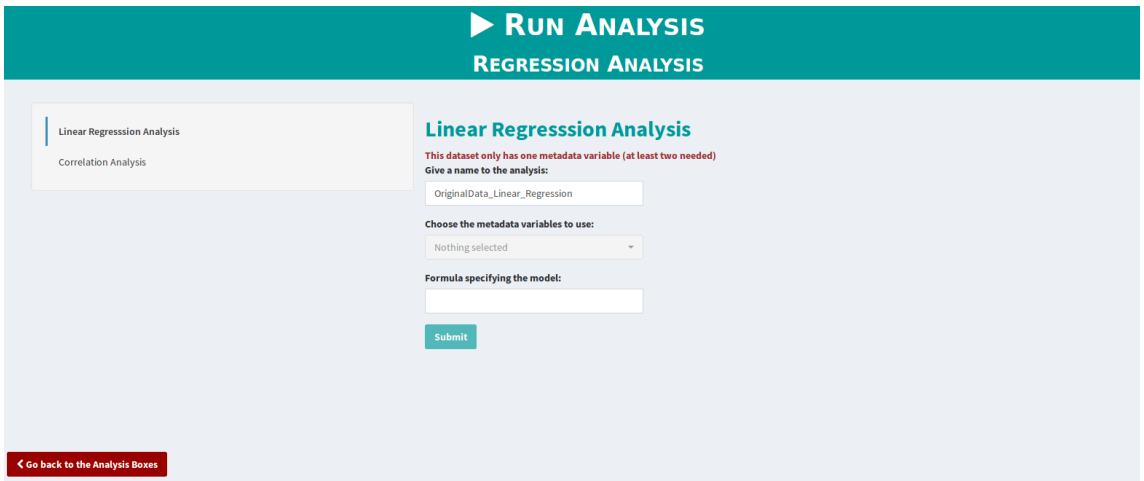



Figure 2.42: Linear Regression analysis page layout.

### Correlation Analysis

The available **options** to set are:

- Analysis Name;
- Correlation Method: "Pearson", "Spearman" or "Kendall";
- If the correlation is to be calculated between samples or variables;
- Color palette for the heatmap and if the reverse colors of the palette should be used.

If the user also chooses to perform a correlations test to the dataset, the following option must be given:

- Alternative hypothesis: "Two-sided", "Greater (positive association)" or "Less (negative association)".

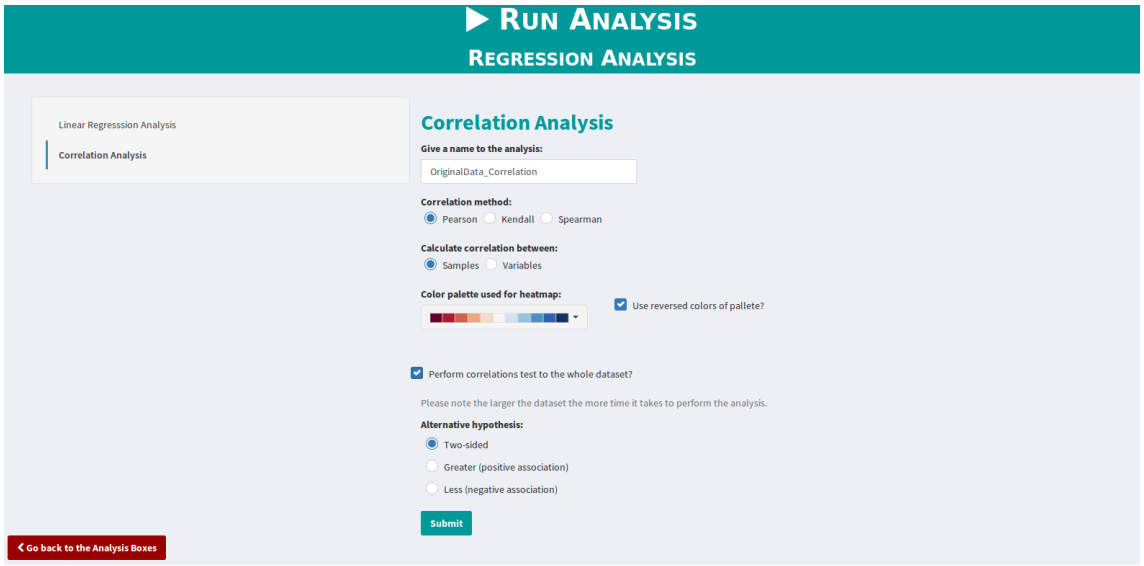



Figure 2.43: Correlation analysis page layout.

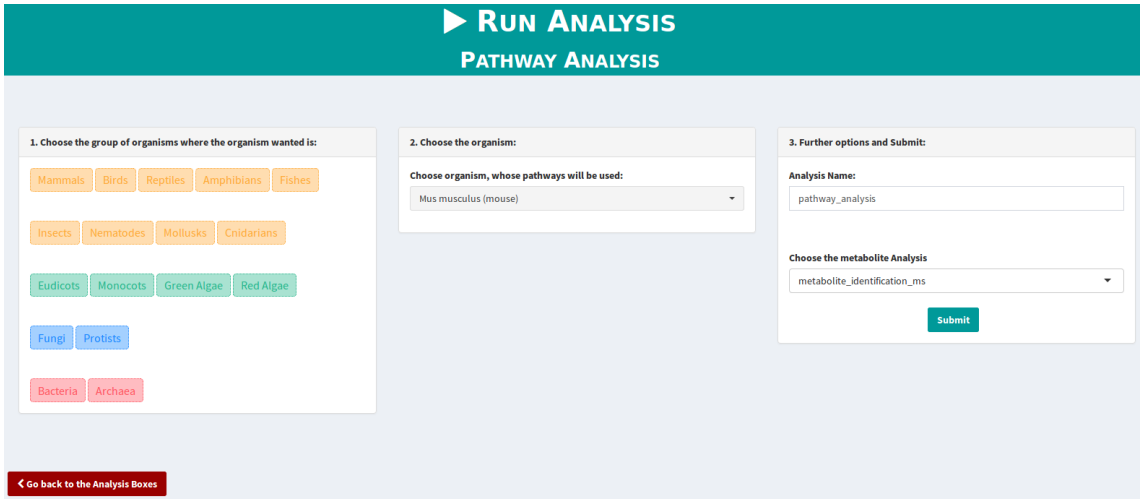
### 2.8.8 Pathway Analysis

The available **options** to set in order to perform this analysis are divided into 3 main boxes:

- Box 1: Choose the group of organisms where the organism wanted is. The available groups are the following: "Mammals", "Birds", "Reptiles", "Amphibians", "Fishes", "Insects", "Nematodes", "Mollusks", "Cnidarians", "Eudicots", "Monocots", "Green Algae", "Red Algae", "Fungi", "Protists", "Bacteria" and "Archaea". The pathways of the chosen organism will be the ones used in the analysis;
- Box 2: Choose the organism. A select input with the organism from the group of organisms chosen is made available.
- Box 3: Further options and Submit. Here, you have to give a name to the analysis and one of two options:

  *If the data is of concentrations type*: say if the metabolites are represented by KEGG codes or HMDB ones;

  *If the data is of NMR peaks or LC-MS type*: say the metabolite identification analysis with the metabolites identified that you want to use in the analysis.


Run Analysis
Pathway Analysis
1. Choose the group of organisms where the organism wanted is:
Mammals
Birds
Reptiles
Amphibians
Fishes
Insects
Nematodes
Mollusks
Cnidarians
Eudicots
Monocots
Green Algae
Red Algae
Fungi
Protists
Bacteria
Archaea
2. Choose the organism:
Choose organism, whose pathways will be used:
Mus musculus (mouse)
3. Further options and Submit:
Analysis Name:
pathway_analysis
Choose the metabolite Analysis
metabolite_identification_ms
Submit
Go back to the Analysis Boxes

## 2.9 Visualization of Results

Each time an analysis is finished, the user is redirected to the respective results page.

All the obtained results related to the data available for analysis are accessible through the sidebar panel, in the tab called "Analysis Results". This tab has subtabs that correspond to each type of analysis made. These subtypes have the links to the respective analysis results' pages, represented by the names given by the user.

Overall, for each results page, the users can access the options used in that analysis, by clicking in a circular button placed ate the top left corner of the results page, alongside with the results.

### 2.9.1 Univariate Analysis

#### T-Test

For results of this type, **options used** that can be consulted are:

- Analysis Name;
- Name of the dataset used;
- Variable used;
- P-value treshold chosen.

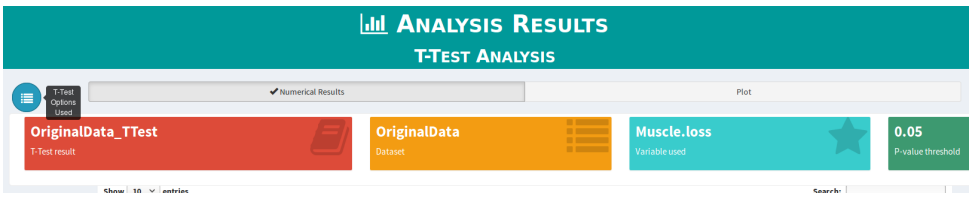
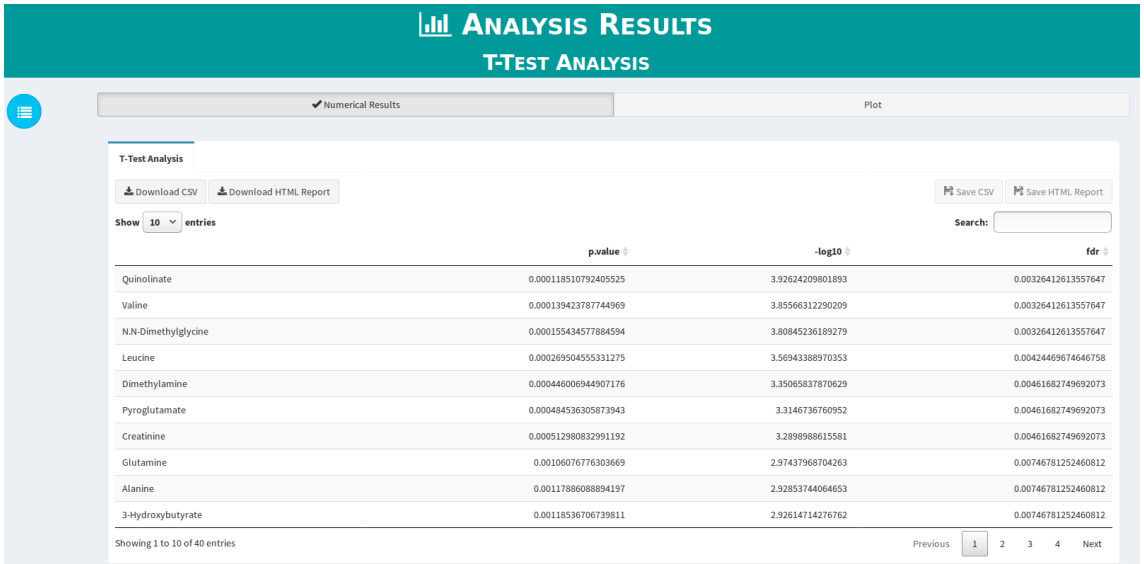



Figure 2.44: Layout of the dropdown menu of the options for T-Test analysis.

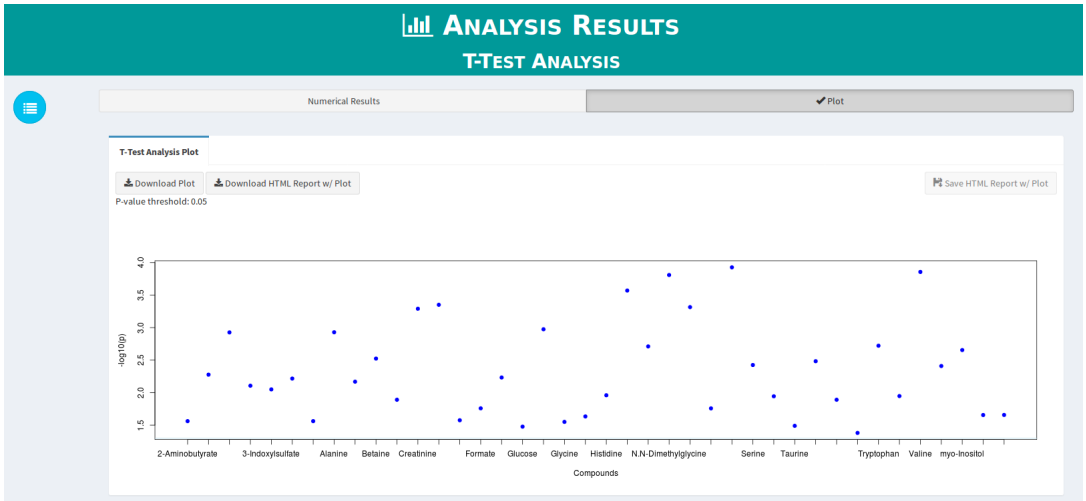
As regards to the actual **results**, at the right of the options button, there are two buttons, which allow the user to see the two different types of results obtained, shown below these buttons.

- *Numerical results*: consists on a table with the p-value, logarithm of p-value and corrected p-value (FDR method);




| | p.value | -log10 | fdr |
|---|---|---|---|
| Quinolinate | 0.000118510792405525 | 3.92624209801893 | 0.00326412613557647 |
| Valine | 0.000139423787744969 | 3.85566312290209 | 0.00326412613557647 |
| N,N-Dimethylglycine | 0.000155434577884594 | 3.80845236189279 | 0.00326412613557647 |
| Leucine | 0.000269504555331275 | 3.56943388970353 | 0.00424469674646758 |
| Dimethylamine | 0.000446006944907176 | 3.35065837870629 | 0.00461682749692073 |
| Pyroglutamate | 0.000484536305873943 | 3.3146736760952 | 0.00461682749692073 |
| Creatinine | 0.000512980832991192 | 3.2898988615581 | 0.00461682749692073 |
| Glutamine | 0.001060767763036 69 | 2.97437968704263 | 0.00746781252460812 |
| Alanine | 0.00117886088894197 | 2.92853744064653 | 0.00746781252460812 |
| 3-Hydroxybutyrate | 0.00118536706739811 | 2.92614714276762 | 0.00746781252460812 |

- *Plot*: The negative base 10 logarithm of the p-value is represented on the y axis and the variables on the x axis.




## One-Way ANOVA

For results of this type, **options used** that can be consulted are:

- Analysis Name;
- Name of the dataset used;
- Variable used;
- If tukeyHSD was performed or not.

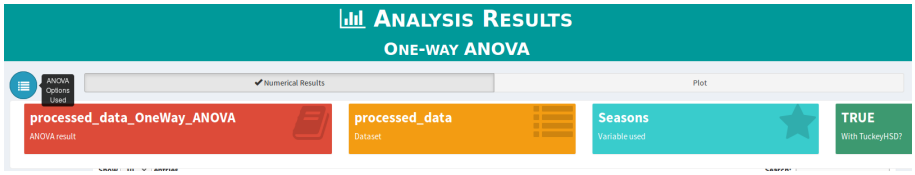
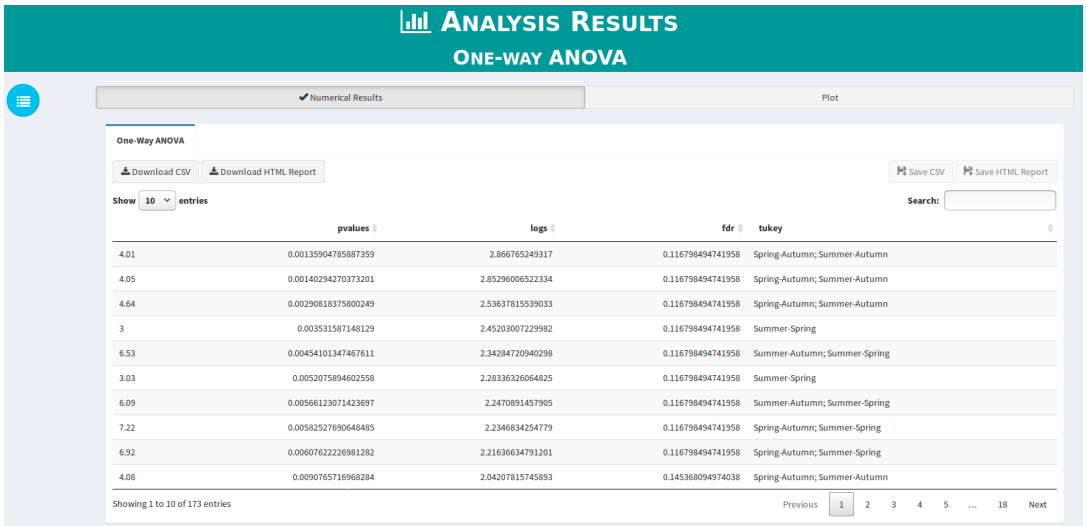



Figure 2.45: Layout of the dropdown menu of the options for one-way ANOVA analysis.

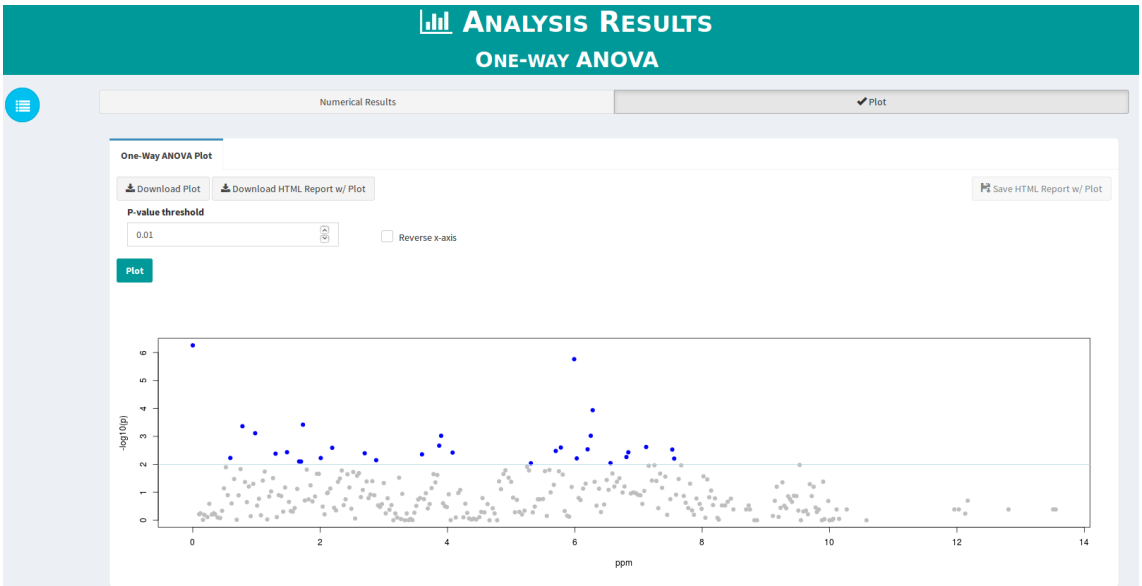
As regards to the actual **results**, at the right of the options button, there are two buttons, which allow the user to see the two different types of results obtained, shown below these buttons.

- *Numerical results*: consists on a table with the p-value, logarithm of p-value, corrected p-value (FDR method) and the result of tukeyHSD, if it was performed, for each variable tested;




| | pvalues | logs | fdr | tukey |
|---|---|---|---|---|
| 4.01 | 0.00135904785887359 | 2.86676524931 7 | 0.116798494741958 | Spring-Autumn; Summer-Autumn |
| 4.05 | 0.00140294270373201 | 2.85296006522334 | 0.116798494741958 | Spring-Autumn; Summer-Autumn |
| 4.64 | 0.00290818375800249 | 2.53637815539033 | 0.116798494741958 | Spring-Autumn; Summer-Autumn |
| 3 | 0.003531587148129 | 2.45203007229982 | 0.116798494741958 | Summer-Spring |
| 6.53 | 0.00454101347467611 | 2.34284720940298 | 0.116798494741958 | Summer-Autumn; Summer-Spring |
| 3.03 | 0.0052075894602558 | 2.28336326064825 | 0.116798494741958 | Summer-Spring |
| 6.09 | 0.00566123071423697 | 2.2470891457905 | 0.116798494741958 | Summer-Autumn; Summer-Spring |
| 7.22 | 0.00582527690648485 | 2.2346834254779 | 0.116798494741958 | Spring-Autumn; Summer-Spring |
| 6.92 | 0.00607622226981282 | 2.21636634791201 | 0.116798494741958 | Spring-Autumn; Summer-Spring |
| 4.08 | 0.0090765716968284 | 2.04207815745893 | 0.145368094974038 | Spring-Autumn; Summer-Autumn |

- *Plot*: The negative base 10 logarithm of the p-value is represented on the y axis and the variables on the x axis. The plot can be personalized through the following available changes:
  P-value Treshold: defaults to 0.01;
  Reverse the x-axis: only available for datasets whose type is not concentrations.




### Multi-factor ANOVA

For results of this type, **options used** that can be consulted are:

- Analysis Name;
- Name of the dataset used;
- Formula used.

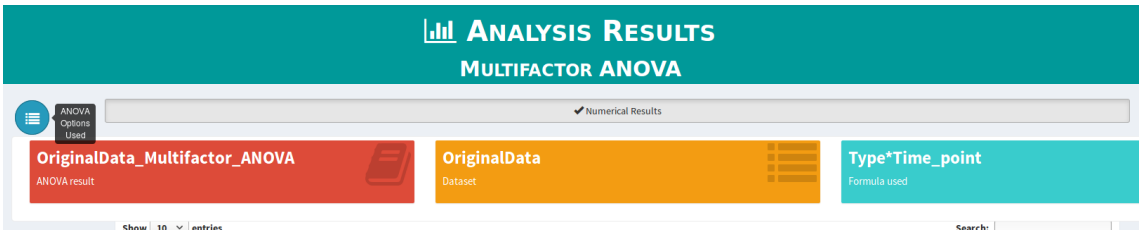



Figure 2.46: Layout of the dropdown menu of the options for Multi-factor ANOVA.

As regards to the actual **results**, the numerical results are available, in the form of a table, where each line corresponds to the different results obtained for each data variable on the dataset. The following information is given in the table's columns:

- Variables' combination (*Var*): if formula variableA*variableB is chosen, for each metabolite, there will be results for variableA, variableB, variableA and variableB, and Residuals;
- Degrees of Freedom (*DF*);
- Sum of Squares (*Sum sq*);
- Mean Squares (*Mean Sq*);
- F-Value (*F value*);
- P-Value (*Pr(>F)*);
- Explained Variability (*Var Exp*);

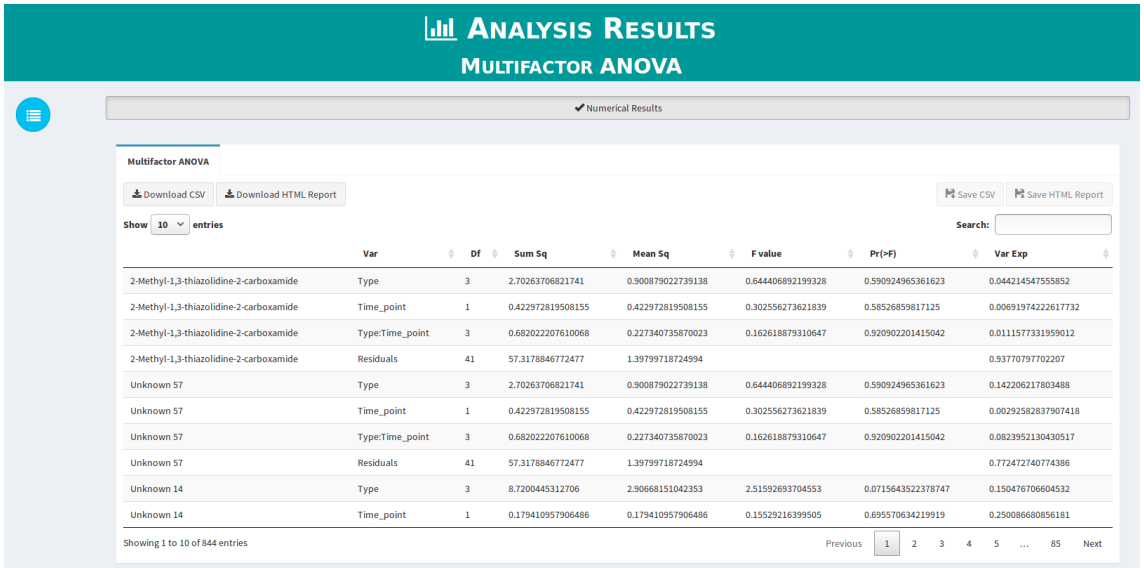



Figure 2.47: Layout of the results for Multi-factor ANOVA.

### Kruskal-Wallis Test

For results of this type, **options used** that can be consulted are:

- Analysis Name;
- Name of the dataset used;
- Variable used;
- P-value treshold chosen.

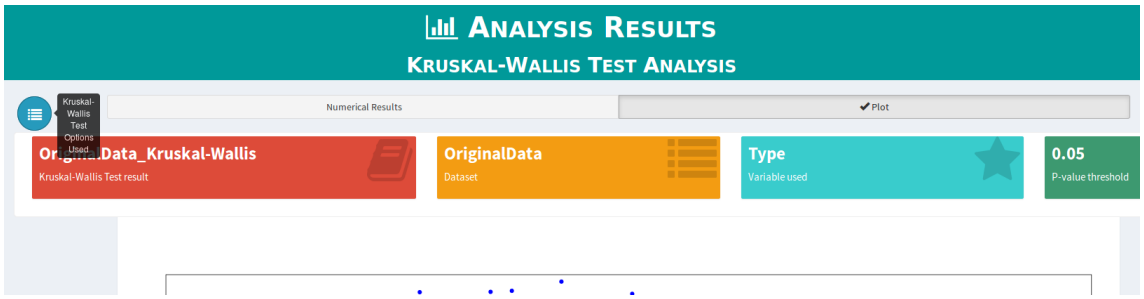
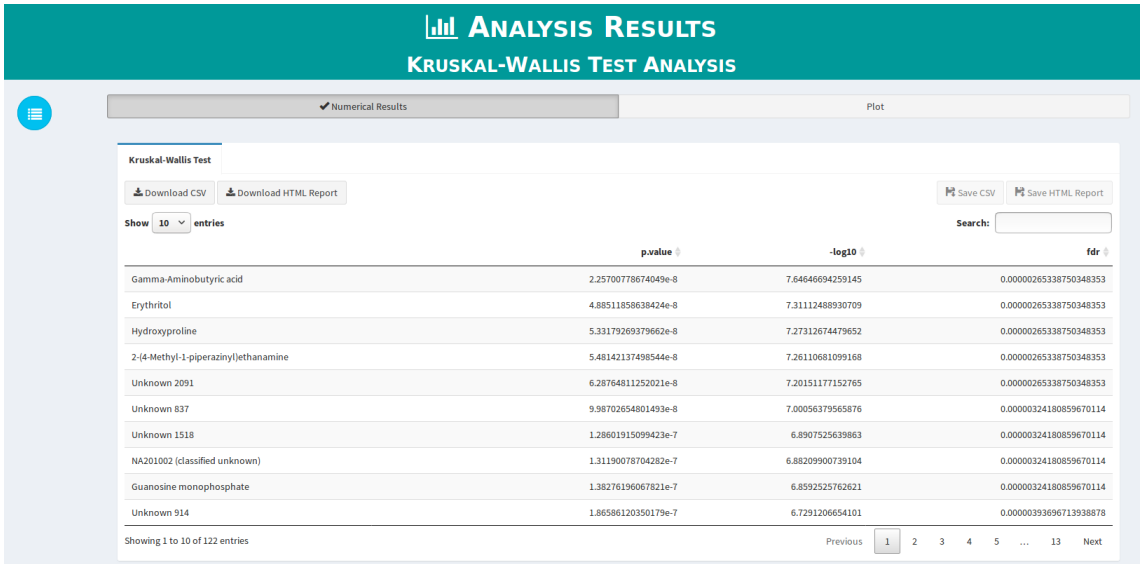


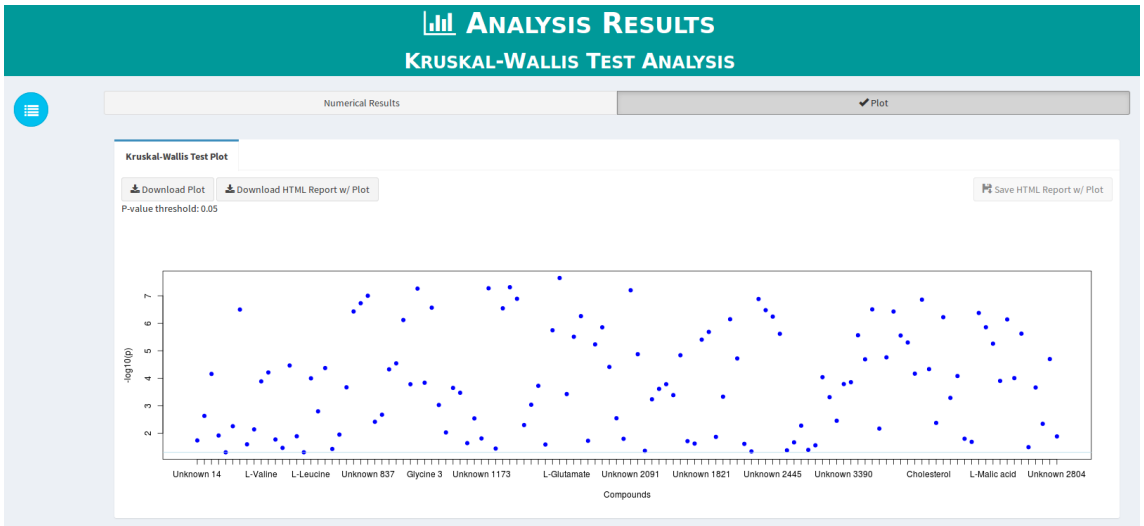

Figure 2.48: Layout of the dropdown menu of the options for Kruskal-Wallis Test.

As regards to the actual **results**, at the right of the options button, there are two buttons, which allow the user to see the two different types of results obtained, shown below these buttons.

- *Numerical results*: consists on a table with the p-value, logarithm of p-value and corrected p-value (FDR method);




- *Plot*: The negative base 10 logarithm of the p-value is represented on the y axis and the variables on the x axis.




### Kolmogorov-Smirnov Test

For results of this type, **options used** that can be consulted are:

- Analysis Name;
- Name of the dataset used;
- Variable used;
- P-value treshold chosen.

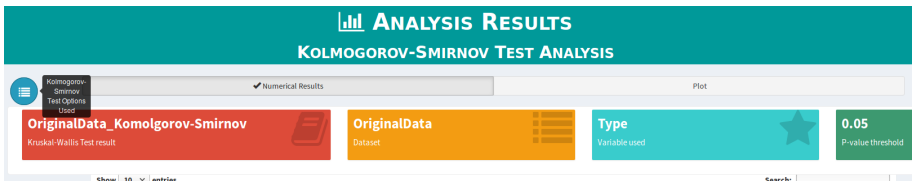
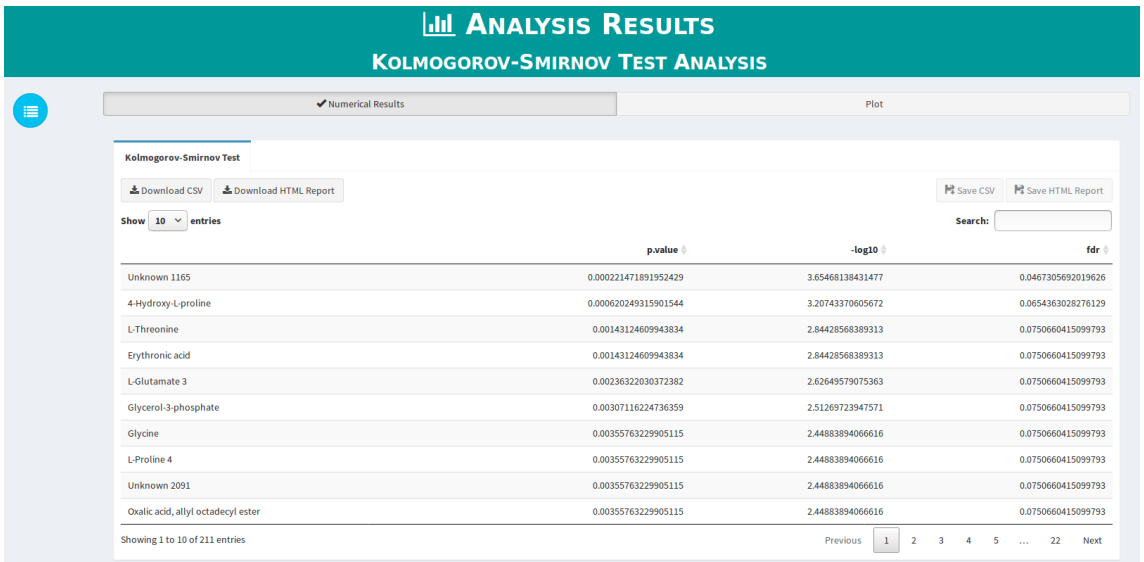


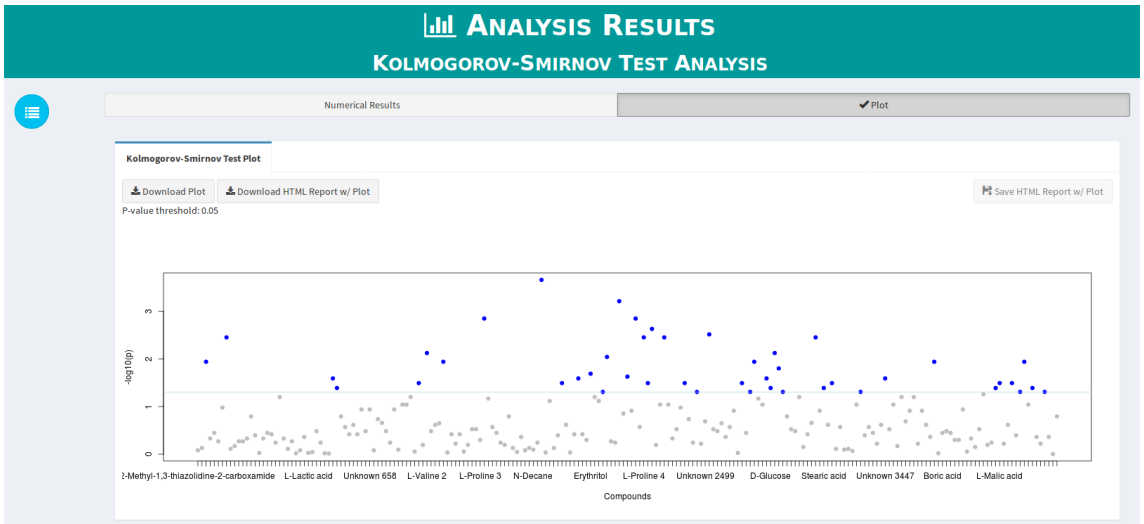

Figure 2.49: Layout of the dropdown menu of the options for Kolmogorov-Smirnov Test.

As regards to the actual **results**, at the right of the options button, there are two buttons, which allow the user to see the two different types of results obtained, shown below these buttons.

- *Numerical results*: consists on a table with the p-value, logarithm of p-value and corrected p-value (FDR method);




- *Plot*: The negative base 10 logarithm of the p-value is represented on the y axis and the variables on the x axis.




### Fold Change Analysis

For results of this type, **options used** that can be consulted are:

- Analysis Name;
- Name of the dataset used;
- Variable used;
- Metadata variable class chosen as the reference value.

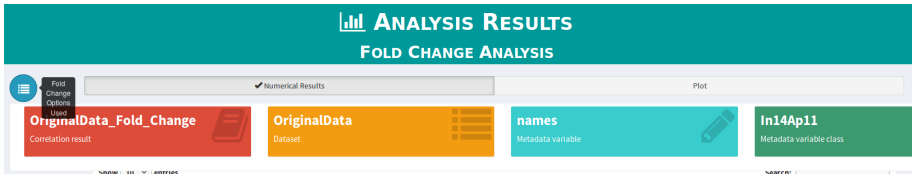
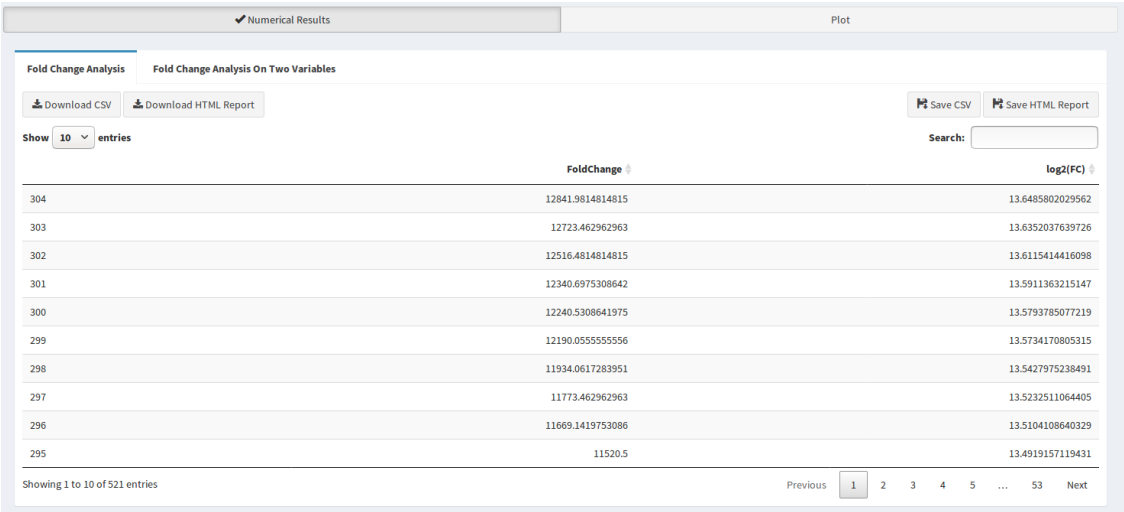



Figure 2.50: Layout of the dropdown menu of the options for Fold change Analysis.

As regards to the actual **results**, at the right of the options button, there are two buttons, which allow the user to see the two different types of results obtained, shown below these buttons.

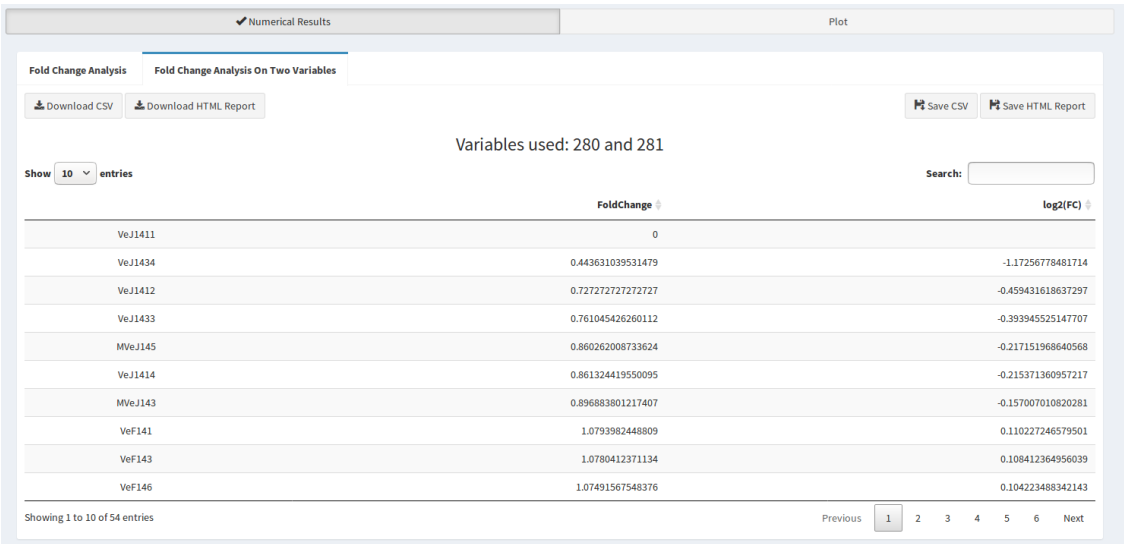
For the *Numerical Results*, there is a tabset panel with two tab panels:

- *Fold Change Analysis*: Table with the results of the fold change for all variables. Each line is represented by a data variable, and has information on the fold change value and the base two logarithm of the respective fold change value;



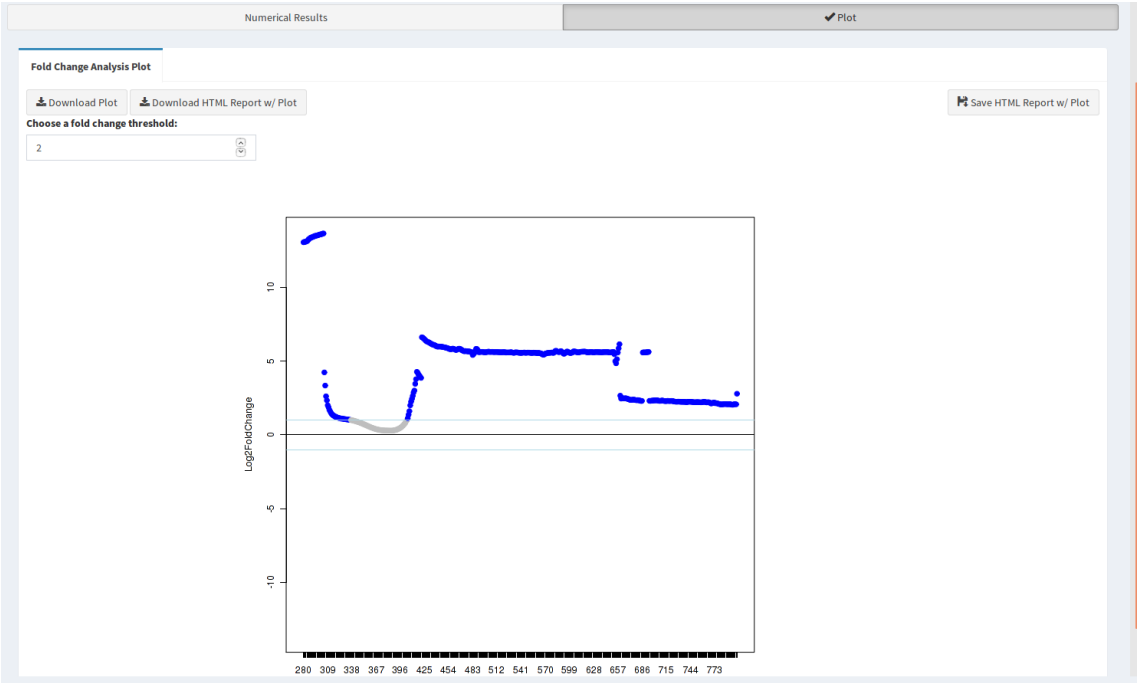

| | FoldChange | log2(FC) |
|---|---|---|
| 304 | 12841.9814814815 | 13.6485802029562 |
| 303 | 12723.462962963 | 13.6352037639726 |
| 302 | 12516.4814814815 | 13.6115414416098 |
| 301 | 12340.6975308642 | 13.5911363215147 |
| 300 | 12240.5308641975 | 13.5793785077219 |
| 299 | 12190.0555555556 | 13.5734170805315 |
| 298 | 11934.0617283951 | 13.5427975238491 |
| 297 | 11773.462962963 | 13.5232511064405 |
| 296 | 11669.1419753086 | 13.5104108640329 |
| 295 | 11520.5 | 13.4919157119431 |

- *Fold Change Analysis on Two Variables*: This tabpanel is only present when this type of analysis is chosen. It contains a table with the values of the fold change and the base two logarithm of the respective fold change value for each group in the metadata variable chosen.




| | FoldChange | log2(FC) |
|---|---|---|
| VeJ1411 | 0 | |
| VeJ1434 | 0.443631039531479 | -1.17256778481714 |
| VeJ1412 | 0.727272727272727 | -0.459431618637297 |
| VeJ1433 | 0.761045426260112 | -0.393945525147707 |
| MVeJ145 | 0.860262008733624 | -0.217151968640568 |
| VeJ1414 | 0.861324419550095 | -0.215371360957217 |
| MVeJ143 | 0.896883801217407 | -0.157007010820281 |
| VeF141 | 1.0793982448809 | 0.110227246579501 |
| VeF143 | 1.0780412371134 | 0.108412364956039 |
| VeF146 | 1.07491567548376 | 0.104223488342143 |

For the **Plot Results**, the base 2 logarithm of the fold change value is represented on the y axis and the variables on the x axis. The point below the threshold value are colored in grey and the other ones in blue. This threshold can be chosen in the numerical input present at the top of the plot.




### Results/Reports available to download/save

All tables present in the Univariate Results can be downloaded or saved (if logged in) in the CSV format.

For each of the T-test, one-way and multi-factor ANOVA, Kolmogorov-Smirnov, Kruskal-Wallis and Fold change analyses there are HTML reports. With exception for multi-factor ANOVA, which does not have any results in form of a plot, the users can choose to download or save (if logged in) a report with or without the plot result.

For fold change analysis reports, they may or may not contain the results on the fold change analysis on two variables, according to if this type analysis was done or not.

## 2.9.2 PCA

The layout of the results for both normal and robust PCA is the same.

For results of this type, **options used** that can be consulted are:

- Analysis Name;
- Name of the dataset used;
- If the dataset was scaled and/or centered

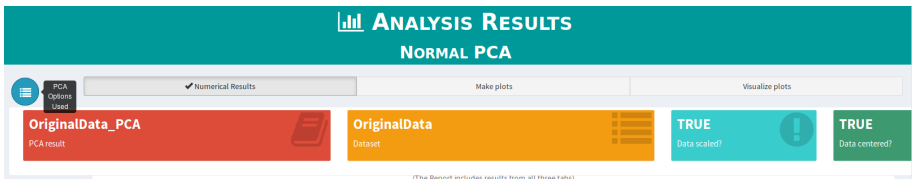
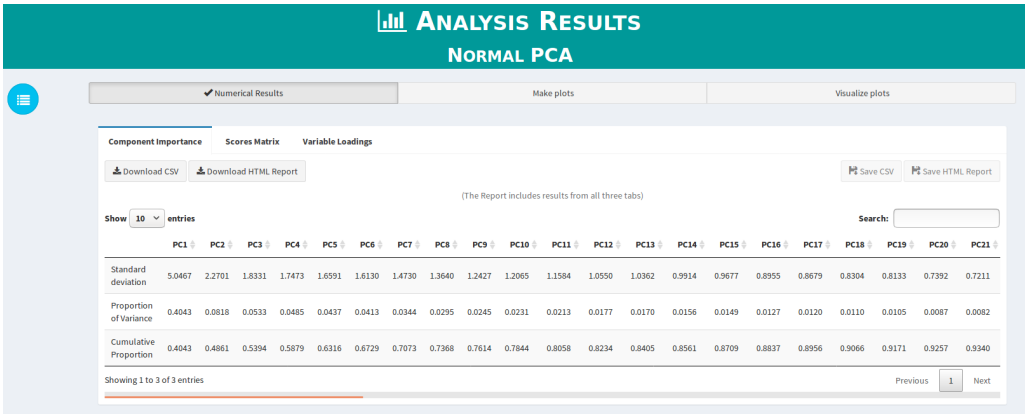



Figure 2.51: Layout of the dropdown menu of the options for PCA Analysis.

As regards to the actual **results**, at the right of the options button, there are three buttons, which allow the user to see the numerical results, set the options to make the plots and visualize the plots:

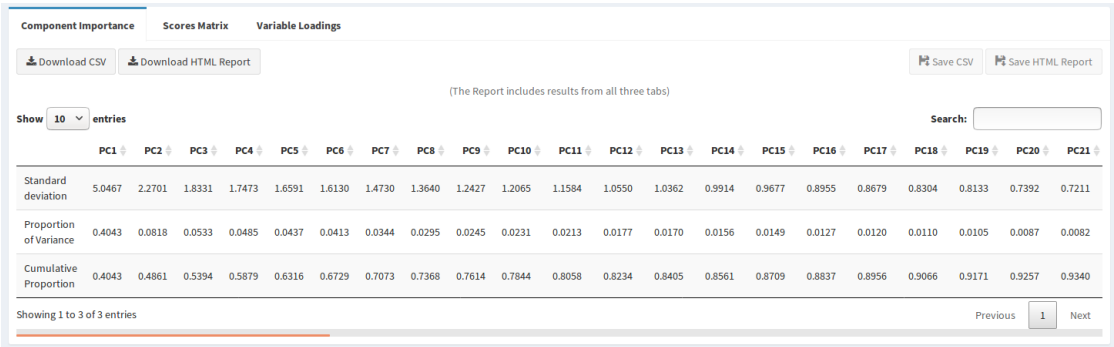



Each one of these sections is detailed below.

### *Numerical Results*

In the numerical results, there is a tabset panel with tabs with the following results:

- *Component Importance*: It contains a table with information on the standard deviation, proportion of variance and and cumulative proportion of the importance of each component;



Component Importance | Scores Matrix | Variable Loadings

Download CSV | Download HTML Report | Save CSV | Save HTML Report

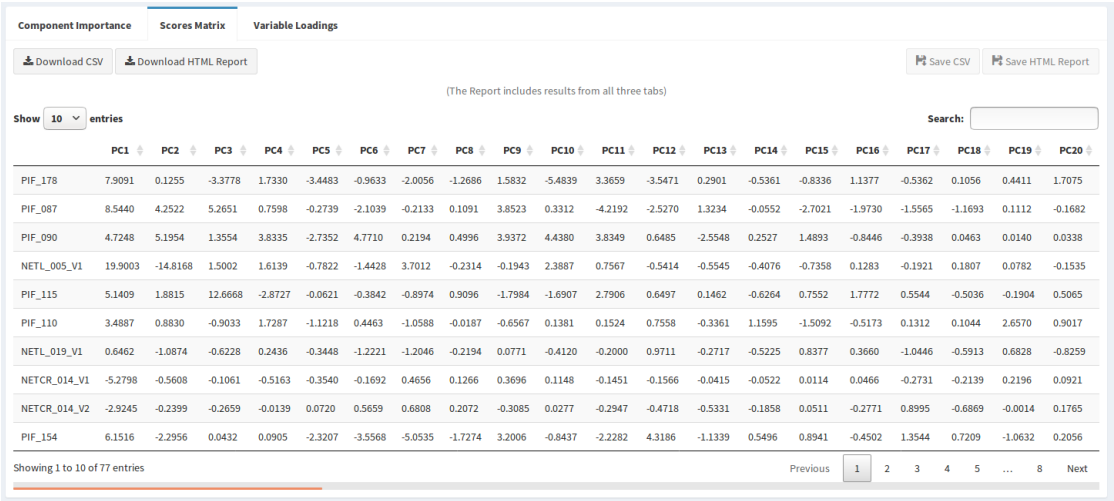
(The Report includes results from all three tabs)

Show 10 entries | Search:

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 | PC11 | PC12 | PC13 | PC14 | PC15 | PC16 | PC17 | PC18 | PC19 | PC20 | PC21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Standard deviation | 5.0467 | 2.2701 | 1.8331 | 1.7473 | 1.6591 | 1.6130 | 1.4730 | 1.3640 | 1.2427 | 1.2065 | 1.1584 | 1.0550 | 1.0362 | 0.9914 | 0.9677 | 0.8955 | 0.8679 | 0.8304 | 0.8133 | 0.7392 | 0.7211 |
| Proportion of Variance | 0.4043 | 0.0818 | 0.0533 | 0.0485 | 0.0437 | 0.0413 | 0.0344 | 0.0295 | 0.0245 | 0.0231 | 0.0213 | 0.0177 | 0.0170 | 0.0156 | 0.0149 | 0.0127 | 0.0120 | 0.0110 | 0.0105 | 0.0087 | 0.0082 |
| Cumulative Proportion | 0.4043 | 0.4861 | 0.5394 | 0.5879 | 0.6316 | 0.6729 | 0.7073 | 0.7368 | 0.7614 | 0.7844 | 0.8058 | 0.8234 | 0.8405 | 0.8561 | 0.8709 | 0.8837 | 0.8956 | 0.9066 | 0.9171 | 0.9257 | 0.9340 |

Showing 1 to 3 of 3 entries | Previous | 1 | Next

- *Scores Matrix*: Table with the scores of each sample for each component. The samples are represented in the lines and the components in the columns;



Component Importance | Scores Matrix | Variable Loadings

Download CSV | Download HTML Report | Save CSV | Save HTML Report

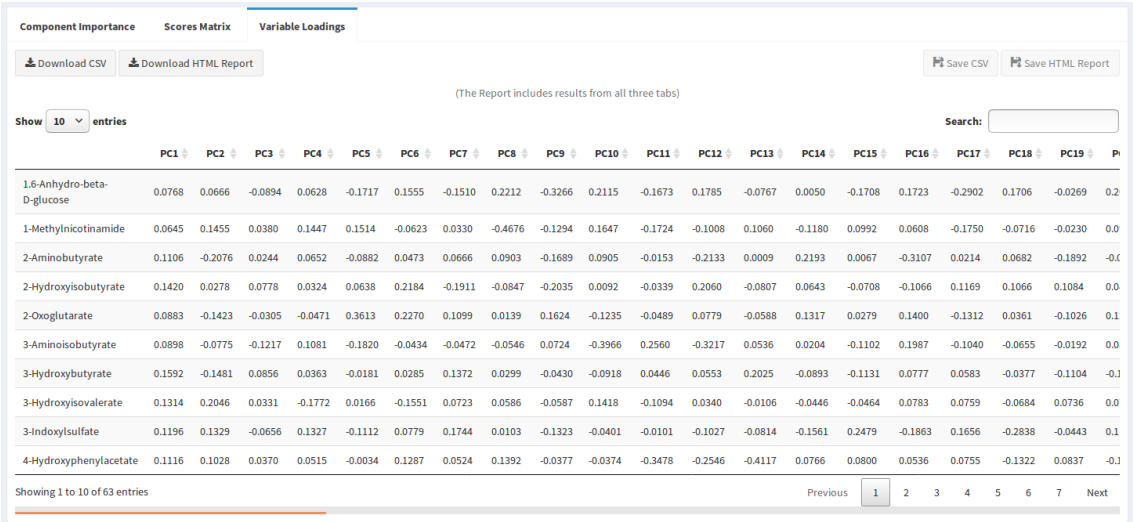
(The Report includes results from all three tabs)

Show 10 entries | Search:

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 | PC11 | PC12 | PC13 | PC14 | PC15 | PC16 | PC17 | PC18 | PC19 | PC20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PIF_178 | 7.9091 | 0.1255 | -3.3778 | 1.7330 | -3.4483 | -0.9633 | -2.0056 | -1.2686 | 1.5832 | -5.4839 | 3.3659 | -3.5471 | 0.2901 | -0.5361 | -0.8336 | 1.1377 | -0.5362 | 0.1056 | 0.4411 | 1.7075 |
| PIF_087 | 8.5440 | 4.2522 | 5.2651 | 0.7598 | -0.2739 | -2.1039 | -0.2133 | 0.1091 | 3.8523 | 0.3312 | -4.2192 | -2.5270 | 1.3234 | -0.0552 | -2.7021 | -1.9730 | -1.5565 | -1.1693 | 0.1112 | -0.1682 |
| PIF_090 | 4.7248 | 5.1954 | 1.3554 | 3.8335 | -2.7352 | 4.7710 | 0.2194 | 0.4996 | 3.9372 | 4.4380 | 3.8349 | 0.6485 | -2.5548 | 0.2527 | 1.4893 | -0.8446 | -0.3938 | 0.0463 | 0.0140 | 0.0338 |
| NETL_005_V1 | 19.9003 | -14.8168 | 1.5002 | 1.6139 | -0.7822 | -1.4428 | 3.7012 | -0.2314 | -0.1943 | 2.3887 | 0.7567 | -0.5414 | -0.5545 | -0.4076 | -0.7358 | 0.1283 | -0.1921 | 0.1807 | 0.0782 | -0.1535 |
| PIF_115 | 5.1409 | 1.8815 | 12.6668 | -2.8727 | -0.0621 | -0.3842 | -0.8974 | 0.9096 | -1.7984 | -1.6907 | 2.7906 | 0.6497 | 0.1462 | -0.6264 | 0.7552 | 1.7772 | 0.5544 | -0.5036 | -0.1904 | 0.5065 |
| PIF_110 | 3.4887 | 0.8830 | -0.9033 | 1.7287 | -1.1218 | 0.4463 | -1.0588 | -0.0187 | -0.6567 | 0.1381 | 0.1524 | 0.7558 | -0.3361 | 1.1595 | -1.5092 | -0.5173 | 0.1312 | 0.1044 | 2.6570 | 0.9017 |
| NETL_019_V1 | 0.6462 | -1.0874 | -0.6228 | 0.2436 | -0.3448 | -1.2221 | -1.2046 | -0.2194 | 0.0771 | -0.4120 | -0.2000 | 0.9711 | -0.2717 | -0.5225 | 0.8377 | 0.3660 | -1.0446 | -0.5913 | 0.6828 | -0.8259 |
| NETCR_014_V1 | -5.2798 | -0.5608 | -0.1061 | -0.5163 | -0.3540 | -0.1692 | 0.4656 | 0.1266 | 0.3696 | 0.1148 | -0.1451 | -0.1566 | -0.0415 | -0.0522 | 0.0114 | 0.0466 | -0.2731 | -0.2139 | 0.2196 | 0.0921 |
| NETCR_014_V2 | -2.9245 | -0.2399 | -0.2659 | -0.0139 | 0.0720 | 0.5659 | 0.6808 | 0.2072 | -0.3085 | 0.0277 | -0.2947 | -0.4718 | -0.5331 | -0.1858 | 0.0511 | -0.2771 | 0.8995 | -0.6869 | -0.0014 | 0.1765 |
| PIF_154 | 6.1516 | -2.2956 | 0.0432 | 0.0905 | -2.3207 | -3.5568 | -5.0535 | -1.7274 | 3.2006 | -0.8437 | -2.2282 | 4.3186 | -1.1339 | 0.5496 | 0.8941 | -0.4502 | 1.3544 | 0.7209 | -1.0632 | 0.2056 |

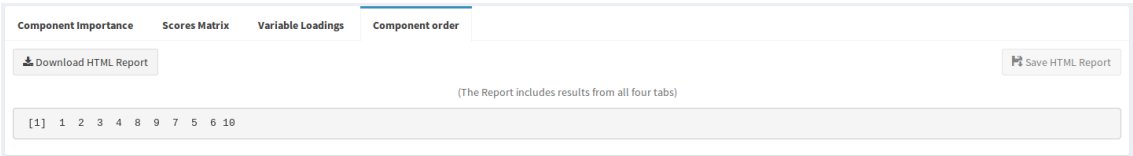
Showing 1 to 10 of 77 entries | Previous | 1 | 2 | 3 | 4 | 5 | ... | 8 | Next

- *Variable Loadings*: Table with the loadings value of each variable for each component. The variables are represented in the lines and the components in the columns.




| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 | PC11 | PC12 | PC13 | PC14 | PC15 | PC16 | PC17 | PC18 | PC19 | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.6-Anhydro-beta-D-glucose | 0.0768 | 0.0666 | -0.0894 | 0.0628 | -0.1717 | 0.1555 | -0.1510 | 0.2212 | -0.3266 | 0.2115 | -0.1673 | 0.1785 | -0.0767 | 0.0050 | -0.1708 | 0.1723 | -0.2902 | 0.1706 | -0.0269 | 0.2 |
| 1-MethylNicotinamide | 0.0645 | 0.1455 | 0.0380 | 0.1447 | 0.1514 | -0.0623 | 0.0330 | -0.4676 | -0.1294 | 0.1647 | -0.1724 | -0.1008 | 0.1060 | -0.1180 | 0.0992 | 0.0608 | -0.1750 | -0.0716 | -0.0230 | 0.0 |
| 2-Aminobutyrate | 0.1106 | -0.2076 | 0.0244 | 0.0652 | -0.0882 | 0.0473 | 0.0666 | 0.0903 | -0.1689 | 0.0905 | -0.0153 | -0.2133 | 0.0009 | 0.2193 | 0.0067 | -0.3107 | 0.0214 | 0.0682 | -0.1892 | -0. |
| 2-Hydroxyisobutyrate | 0.1420 | 0.0278 | 0.0778 | 0.0324 | 0.0638 | 0.2184 | -0.1911 | -0.0847 | -0.2035 | 0.0092 | -0.0339 | 0.2060 | -0.0807 | 0.0643 | -0.0708 | -0.1066 | 0.1169 | 0.1066 | 0.1084 | 0.0 |
| 2-Oxoglutarate | 0.0883 | -0.1423 | -0.0305 | -0.0471 | 0.3613 | 0.2270 | 0.1099 | 0.0139 | 0.1624 | -0.1235 | -0.0489 | 0.0779 | -0.0588 | 0.1317 | 0.0279 | 0.1400 | -0.1312 | 0.0361 | -0.1026 | 0.1 |
| 3-Aminoisobutyrate | 0.0898 | -0.0775 | -0.1217 | 0.1081 | -0.1820 | -0.0434 | -0.0472 | -0.0546 | 0.0724 | -0.3966 | 0.2560 | -0.3217 | 0.0536 | 0.0204 | -0.1102 | 0.1987 | -0.1040 | -0.0655 | -0.0192 | 0.0 |
| 3-Hydroxybutyrate | 0.1592 | -0.1481 | 0.0856 | 0.0363 | -0.0181 | 0.0285 | 0.1372 | 0.0299 | -0.0430 | -0.0918 | 0.0446 | 0.0553 | 0.2025 | -0.0893 | -0.1131 | 0.0777 | 0.0583 | -0.0377 | -0.1104 | -0. |
| 3-Hydroxyisovalerate | 0.1314 | 0.2046 | 0.0331 | -0.1772 | 0.0166 | -0.1551 | 0.0723 | 0.0586 | -0.0587 | 0.1418 | -0.1094 | 0.0340 | -0.0106 | -0.0446 | -0.0464 | 0.0783 | 0.0759 | -0.0684 | 0.0736 | 0.0 |
| 3-IndoxylSulfate | 0.1196 | 0.1329 | -0.0656 | 0.1327 | -0.1112 | 0.0779 | 0.1744 | 0.0103 | -0.1323 | -0.0401 | -0.0101 | -0.1027 | -0.0814 | -0.1561 | 0.2479 | -0.1863 | 0.1656 | -0.2838 | -0.0443 | 0.1 |
| 4-Hydroxyphenylacetate | 0.1116 | 0.1028 | 0.0370 | 0.0515 | -0.0034 | 0.1287 | 0.0524 | 0.1392 | -0.0377 | -0.0374 | -0.3478 | -0.2546 | -0.4117 | 0.0766 | 0.0800 | 0.0536 | 0.0755 | -0.1322 | 0.0837 | -0. |

- *Component Order*: only present when the results come from a robust PCA, it contains the order of the components



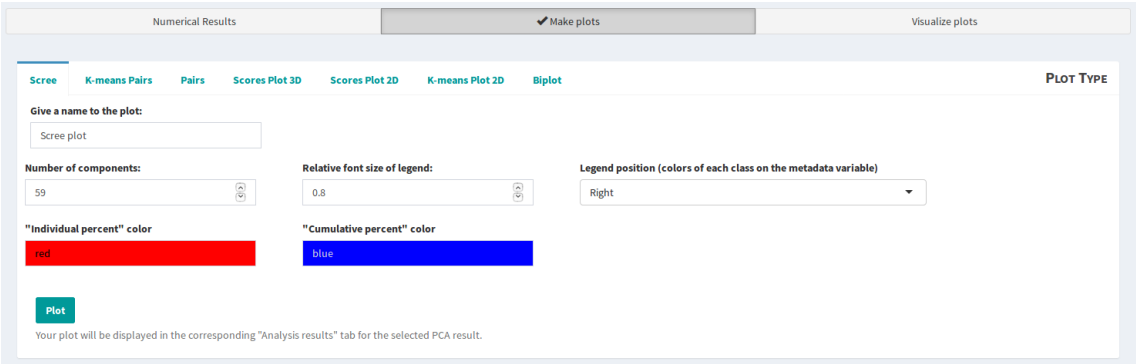

### *Make plots* and *Visualize plots*

The user is able to obtain more than one different plot for each type of plot. After setting the options for a plot and click in the button "Plot" so that the website can construct the plot, in the section *Make plots*, in the section *Visualize plots* the users are able to see the plots constructed, by choosing the one to see in the input located below the download buttons. After choosing the plot to display, the plot appears below.

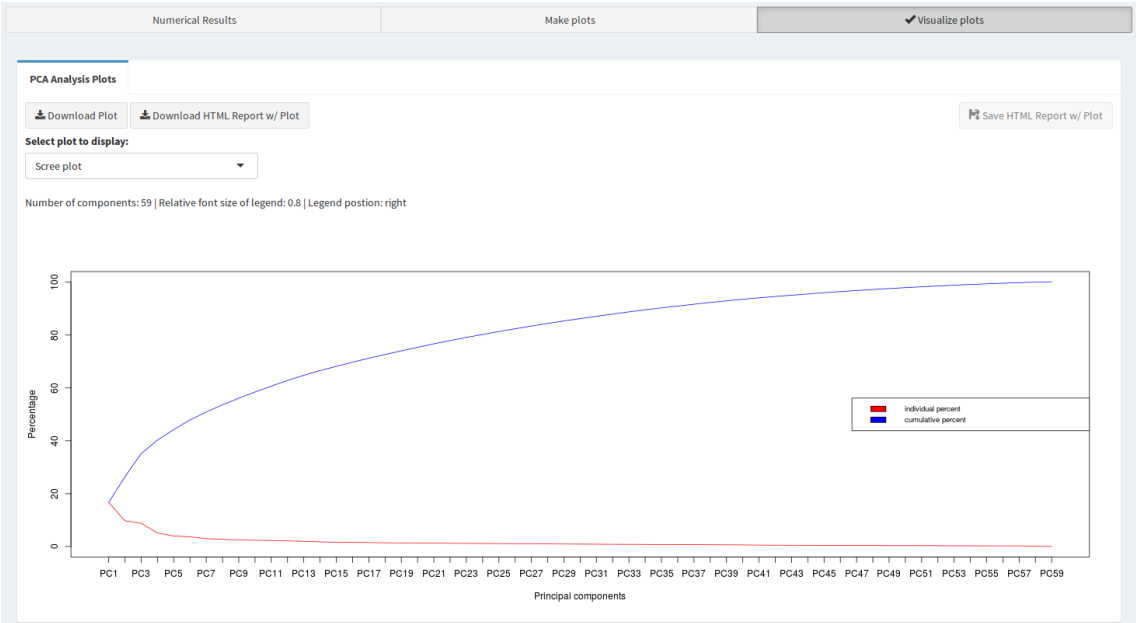
For each type of plot, the options to set are:

***Scree*** - Shows the individual and cumulative percentages of the explained variance of each principal component:

- Give a name to the plot;
- Number of components to show on the xx-axis;
- Relative font size of legend: if 0.8, it will be 80% of the normal size;
- Legend position in the plot: "Bottom right", "Bottom", "Bottom left", "Left", "Top Left", "Top", "Top right", "Right", or "Center";
- Color of the line that represents the individual percent;
- Color of the line that represents the cumulative percent.




(a)

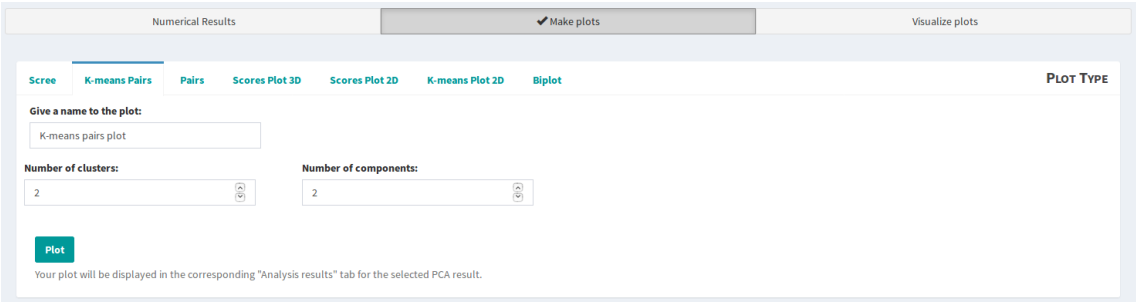



(b)

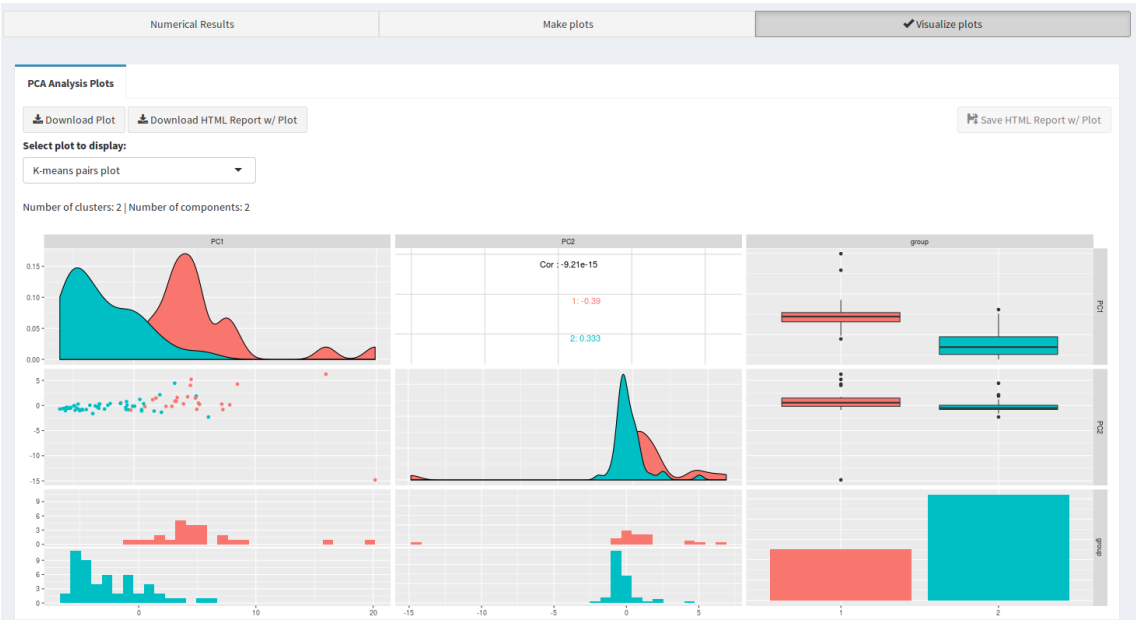
Figure 2.52: Layouts of the (a) screeplot options in the *Make plot* section and (b) *Visualize plots* section when a scree plot is selected, on the PCA analysis results page.

***K-means Pairs*** - Shows the pairs plot of the scores of the defined principal components, using the K-means results for coloring the points according to the cluster they belong:

- Name of the plot;
- Number of clusters for the K-means clustering;
- Number of components to show on the plot.




(a)

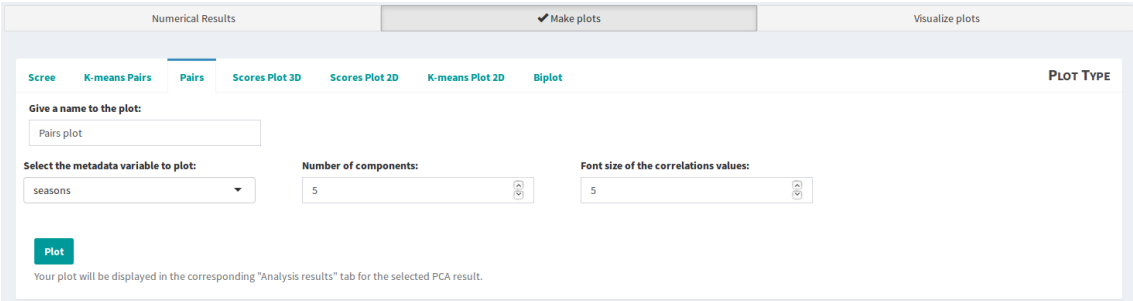



(b)

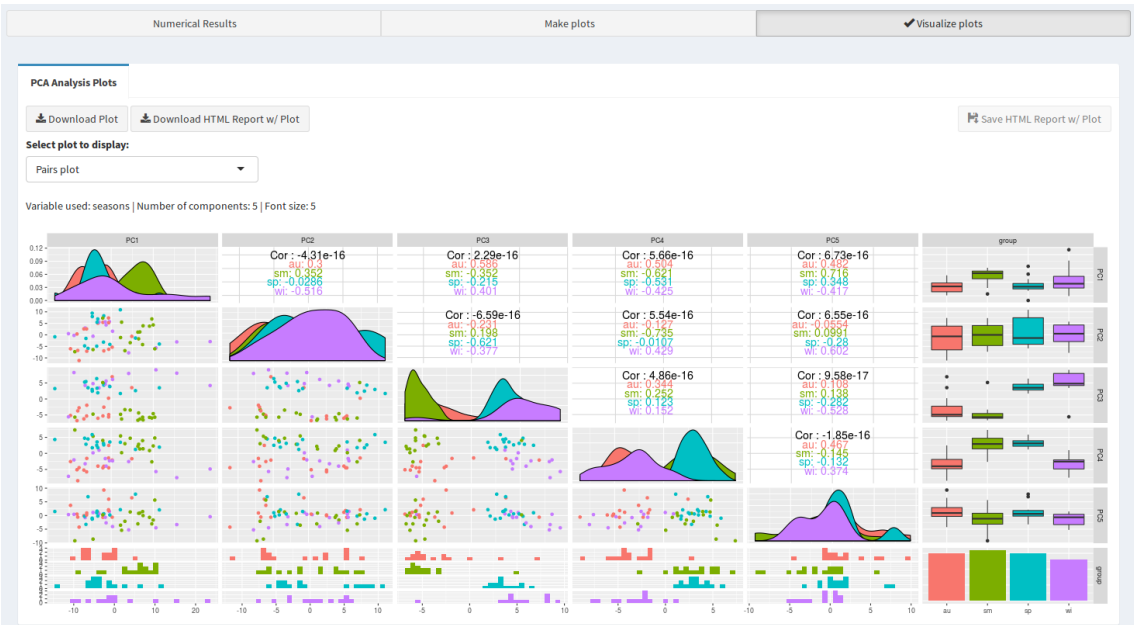
Figure 2.53: Layouts of the (a) k-means pairs plot options in the *Make plot* section and (b) *Visualize plots* section when a k-means pairs plot is selected, on the PCA analysis results page.

***Pairs*** - Shows the pairs plot of the scores of the defined principal components, for a chosen variable

- Name of the plot;
- Metadata variable to plot: the plot will be also coloured according to the classes of the metadata variable chosen;
- Number of components to show on the plot;
- Font size of the correlations values.




(a)

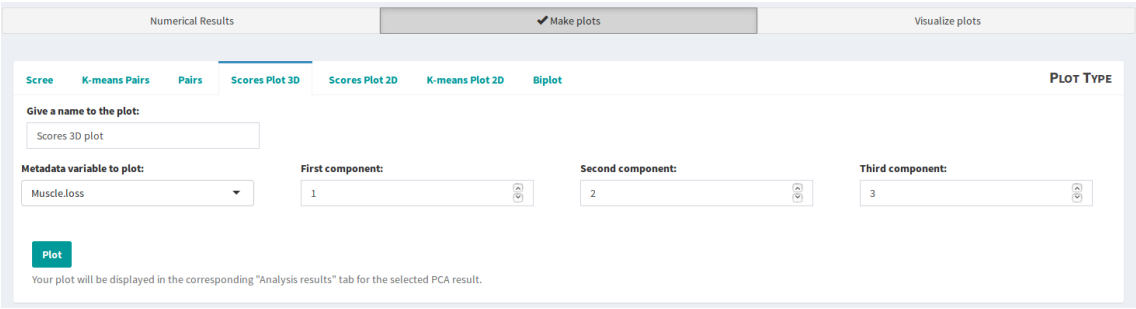



(b)

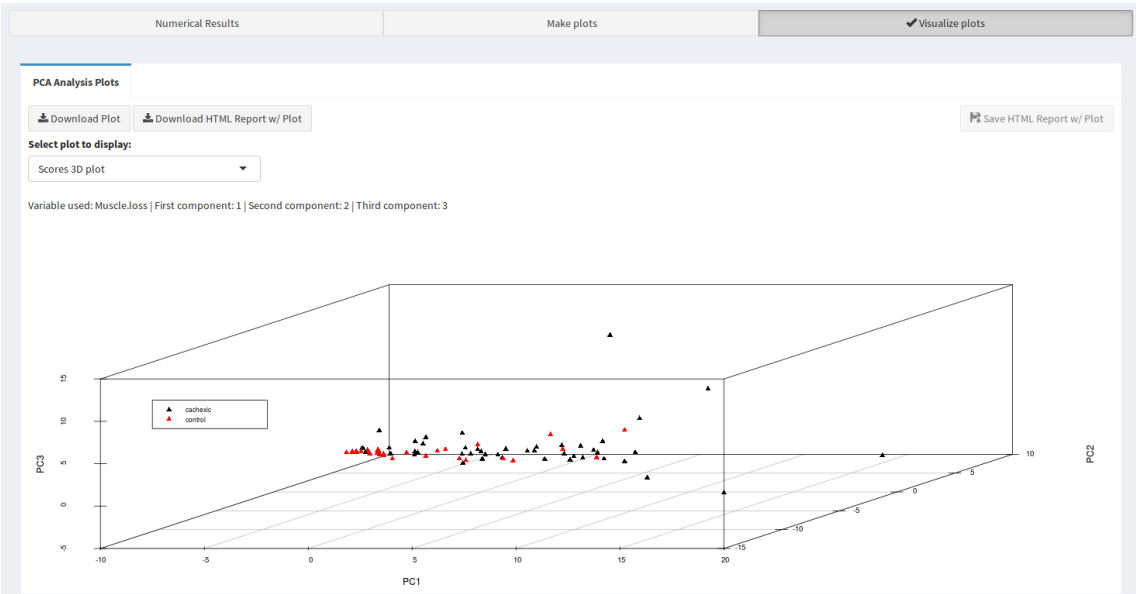
Figure 2.54: Layouts of the (a) pairs plot options in the *Make plot* section and (b) *Visualize plots* section when a pairs plot is selected, on the PCA analysis results page.

***Scores Plot 3D*** - Shows the scores of three different principal components

- Name of the plot;
- Metadata varianle to plot: the plot will be coloured according to the classes of the metadata variable chosen;
- Give the three components to plot.




(a)

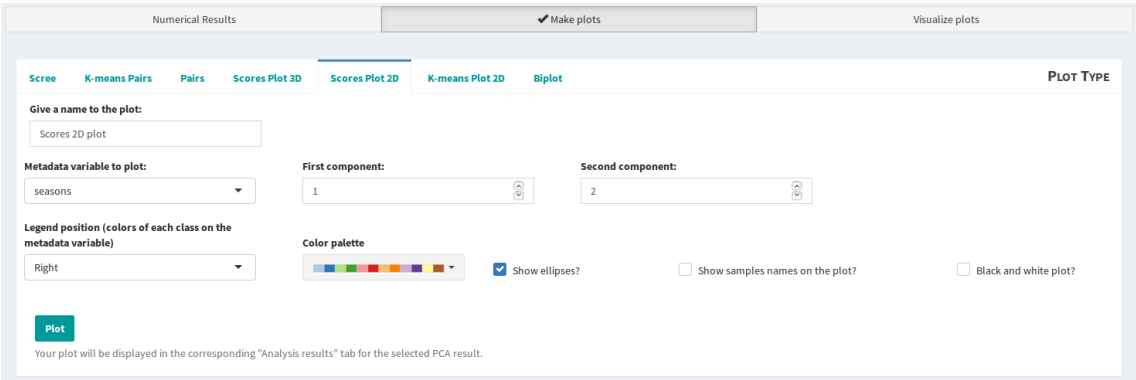



(b)

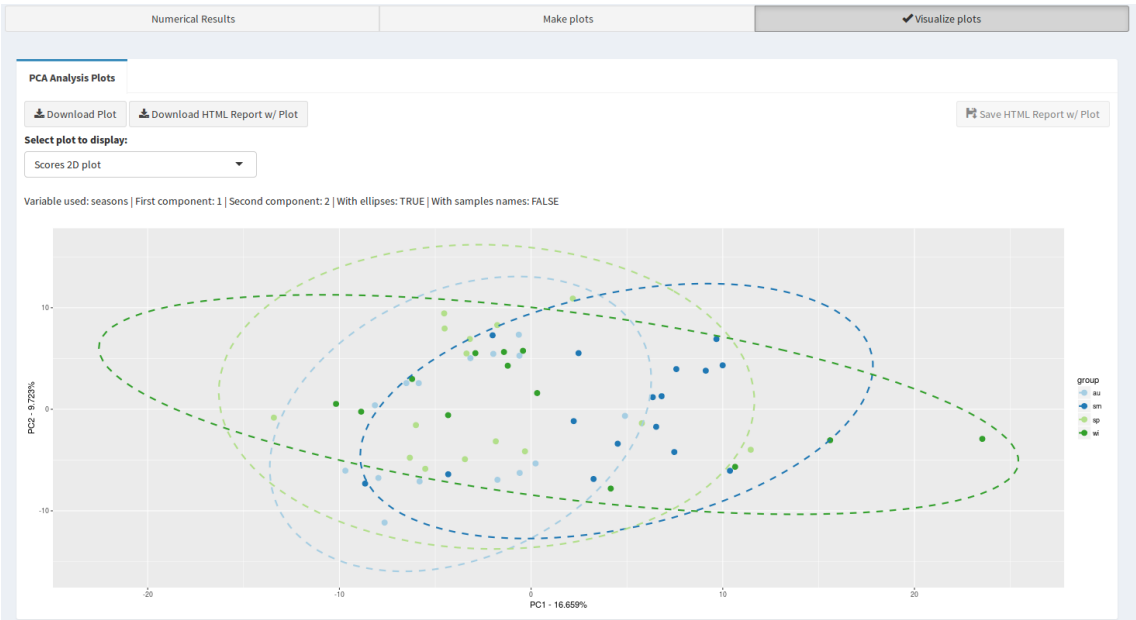
Figure 2.55: Layouts of the (a) scores plot 3D options in the *Make plot* section and (b) *Visualize plots* section when a scores plot 3D is selected, on the PCA analysis results page.

***Scores Plot 2D*** - Shows the scores of two different principal components

- Name of the plot;
- Metadata variable to plot: the plot will be also coloured according to the classes of the metadata variable chosen;
- Give the two components to plot;
- Legend position in the plot: "Bottom right", "Bottom", "Bottom left", "Left", "Top Left", "Top", "Top right", "Right", or "Center";
- Color palette: to color the data points according to the classes of the metadata variables chosen;
- If ellipses should be drawn on each class of the metadata's variable chosen;
- If samples' names should be shown in the plot (each point corresponds to a sample);
- If the plot should be black and white instead of colored. If this options is chosen, no ellipses will be drawn, as they can only be distinguished by the color.




(a)

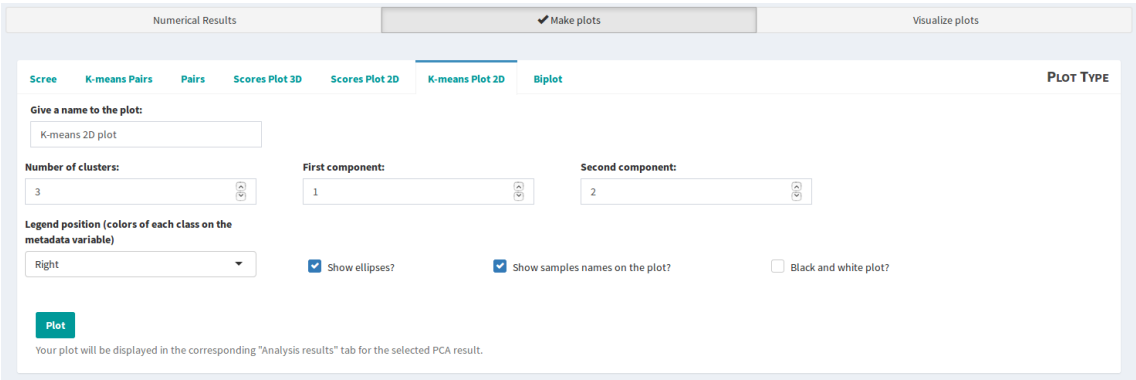



(b)

Figure 2.56: Layouts of the (a) scores plot 2D options in the *Make plot* section and (b) *Visualize plots* section when a scores plot 2D is selected, on the PCA analysis results page.

***K-means Plot 2D*** - Shows the scores of two different principal components, using the K-means results for coloring the points according to the cluster they belong:

- Name of the plot;
- Number of clusters for the K-means clustering;
- Give the two plots to show;
- Legend position in the plot: "Bottom right", "Bottom", "Bottom left", "Left", "Top Left", "Top", "Top right", "Right", or "Center";
- Color palette: to color the data points according to the clusters of the k-means;
- If ellipses should be drawn on each class of the metadata's variable chosen;
- If samples' names should be shown in the plot (each point corresponds to a sample);
- If the plot should be black and white instead of colored. If this options is chosen, no ellipses will be drawn, as they can only be distinguished by the color.



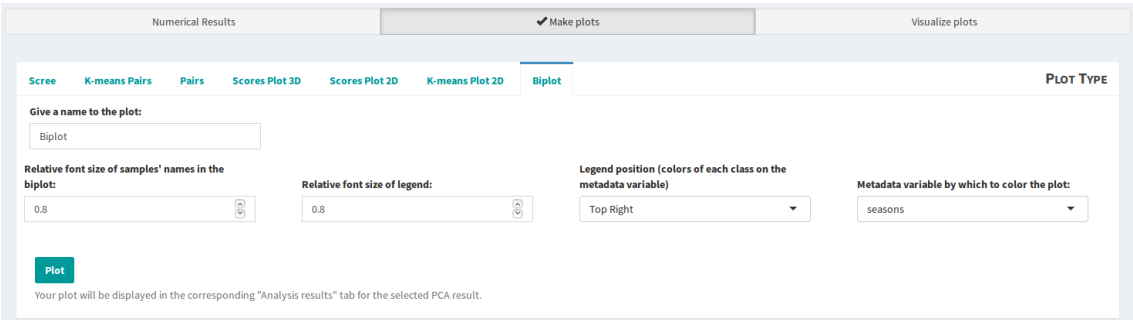

(a)


(b)

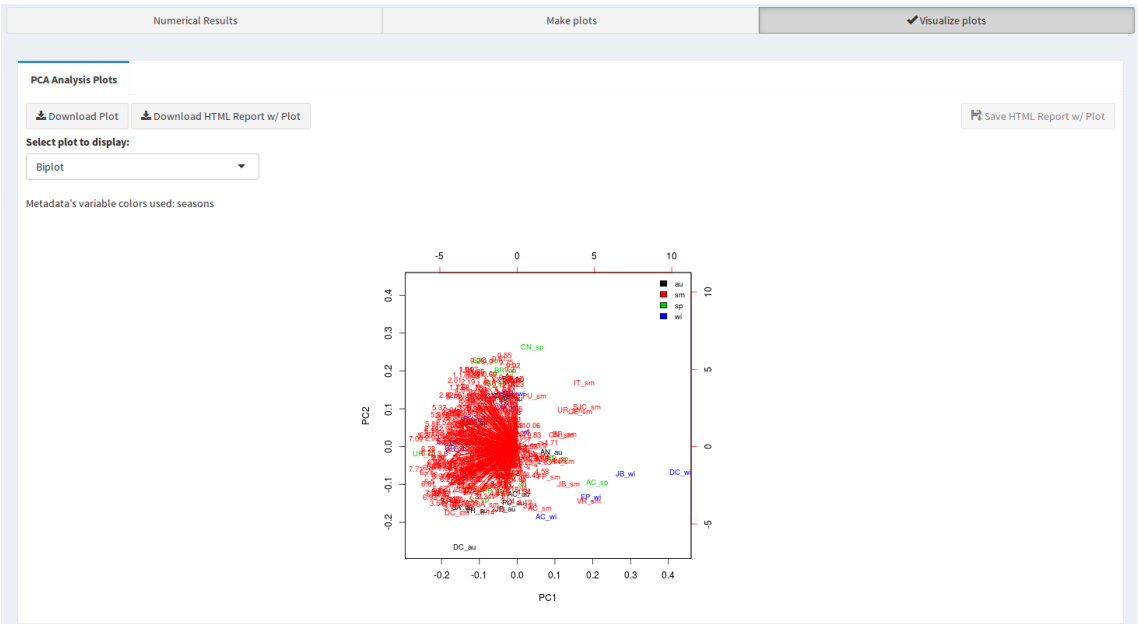
Figure 2.57: Layouts of the (a) K-means plot 2D options in the *Make plot* section and (b) *Visualize plots* section when a K-means plot 2D is selected, on the PCA analysis results page.

***Biplot*** - displays the samples as points, while the variables are displayed either as vectors, linear axes or nonlinear trajectories, considering the first and second PCs as axes

- Name of the plot;
- Relative font size of the samples names in the plot: if 0.8, it will be 80% of the normal size;
- Relative font size of legend: if 0.8, it will be 80% of the normal size;
- Legend position in the plot: "Bottom right", "Bottom", "Bottom left", "Left", "Top Left", "Top", "Top right", "Right", or "Center";
- Metadata variable to plot: the data points will be coloured according to the classes of the metadata variable chosen.




(a)




(b)

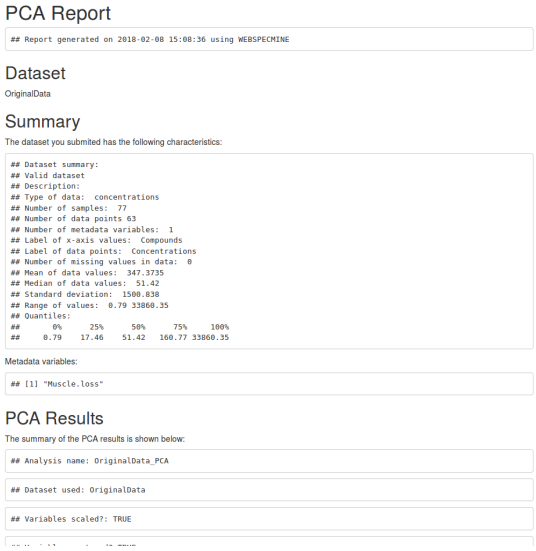
Figure 2.58: Layouts of the (a) Biplot options in the *Make plot* section and (b) *Visualize plots* section when a Biplot is selected, on the PCA analysis results page.

### Results/Reports available to download/save

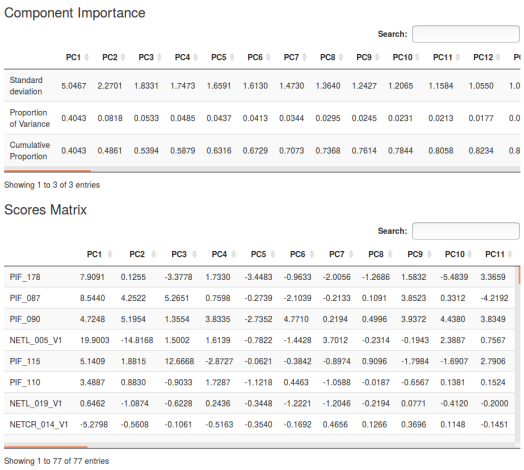
All tables in the PCA results can be downloaded or saved (if logged in) in the CSV format.

All the numerical results can be downloaded in the form of an HTML report.

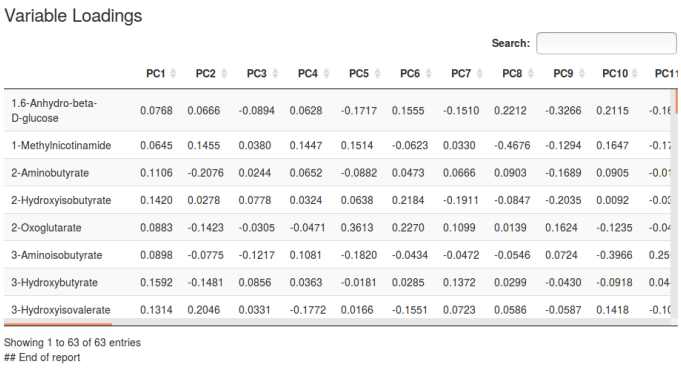

(a)




(b)




(c)




(d)

Figure 2.59: (a) A report on the numerical results on the PCA can be downloaded or saved (if logged in) through the buttons (marked with the red rectangles) present at the top of the tab panel, in the section "Numerical Results". An example of a report of this type is present at (b), (c), (d).

All plots generated can be downloaded or saved (if logged in) as image files.

Furthermore, an HTML report can be generated, containing the plots chosen from the ones generated at the time:

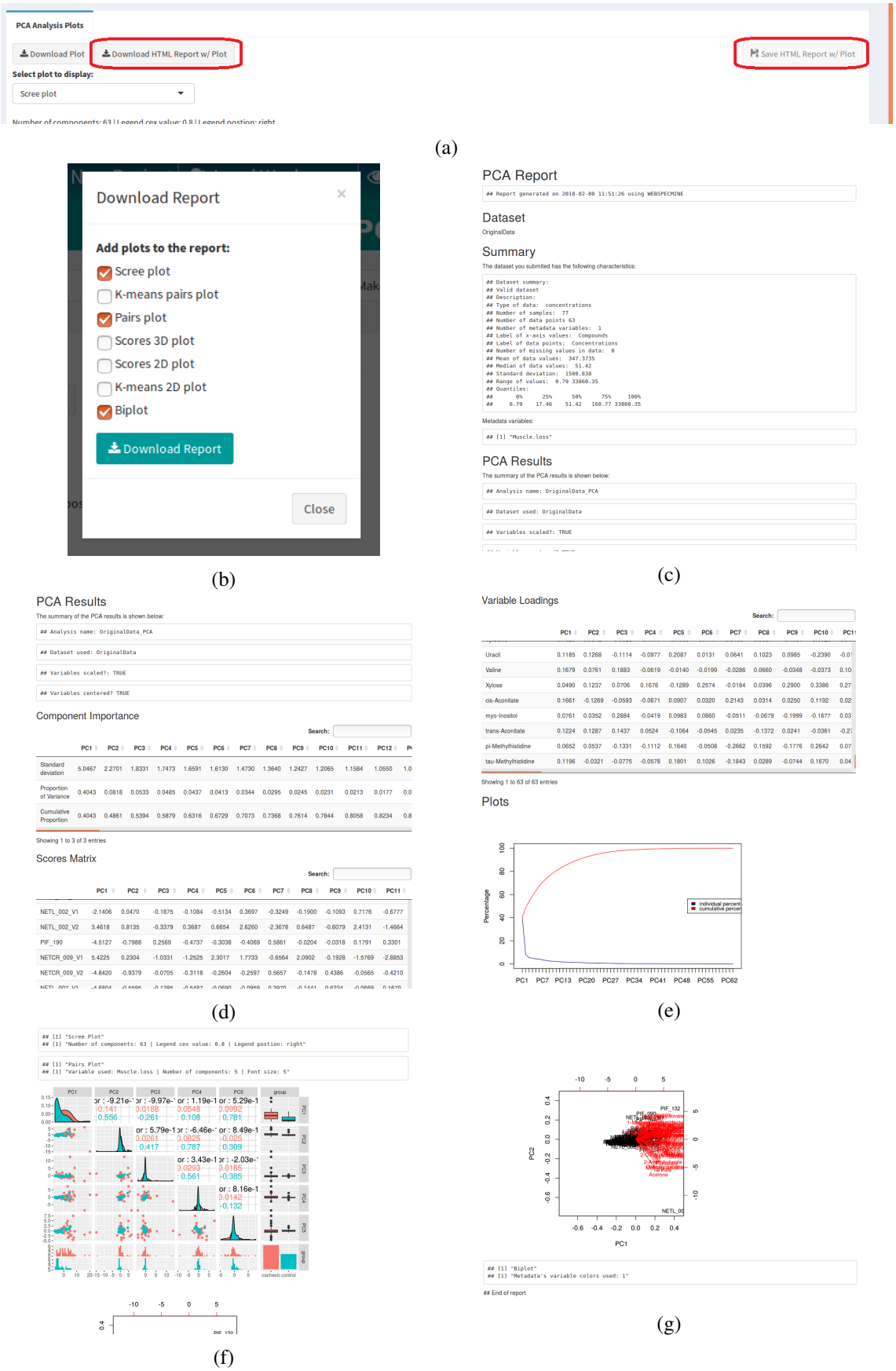



Figure 2.60: (a) A report on all the results on the PCA, inluding the plots, can be downloaded or saved (if logged in) through the buttons (marked with the red rectangles) present at the top of the tab panel, in the section "Visualize plots". (b) After cliking one of these buttons, a pop-up window appears so that the user can specify which plots he wants to insert in the report. An example of a report of this type is present at (c), (d), (e), (f) and (g).

### 2.9.3 Clustering Analysis

#### Hierarchical Clustering

For results of this type, **options used** that can be consulted are:

- Analysis Name;
- Name of the dataset used;
- The distance measure used;
- The agglomeration method;
- If clustering was on samples or variables;
- Metadata variable used to color the samples in the dendogram, in case the clustering was done on samples.

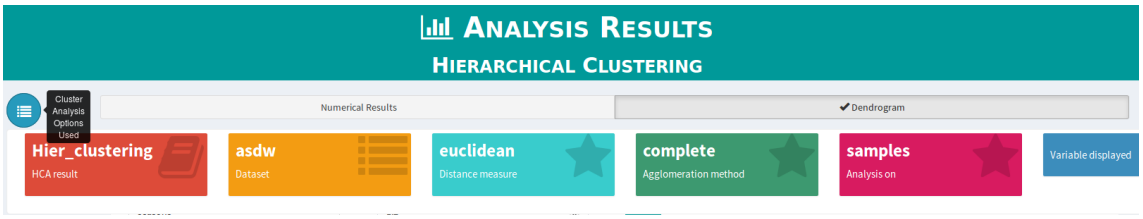



Figure 2.61: Layout of the dropdown menu of the options for Hierarchical Clustering Analysis.

As regards to the actual **results**, there are two buttons, which allow the user to access the numerical results and the dendogram plot.

- *Numerical Results*: distance values (heights) between the formed clusters (or samples, in case the distance is still between two samples, that form a cluster). The values are ordered from lower to higher. The visual representation of this distances is observable in the dendogram plot.

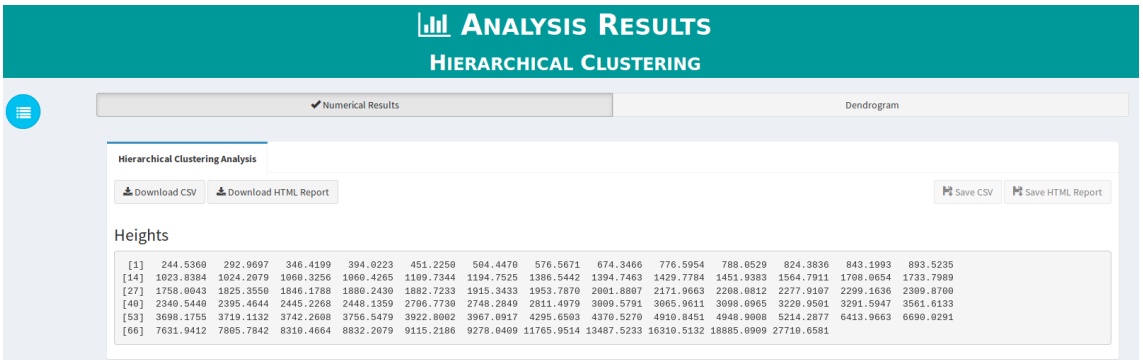


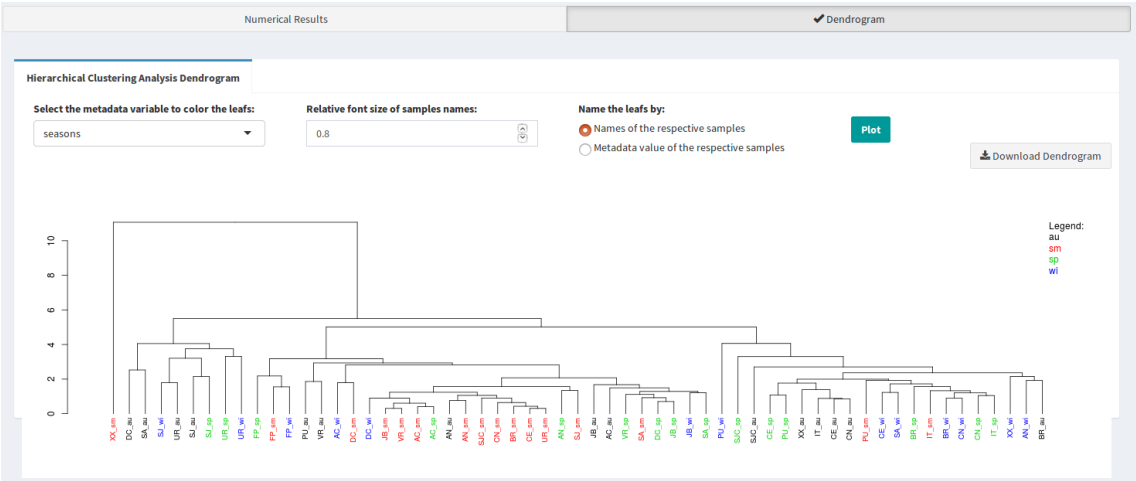

Figure 2.62: Layout of the dendogram section of the results for Hierarchical Clustering Analysis.

- *Dendogram Plot*: the dendogram plot is plotted in such a way so that no branches cross. Also, when the clustering is on the samples, the following options can be set to personalize the plot:

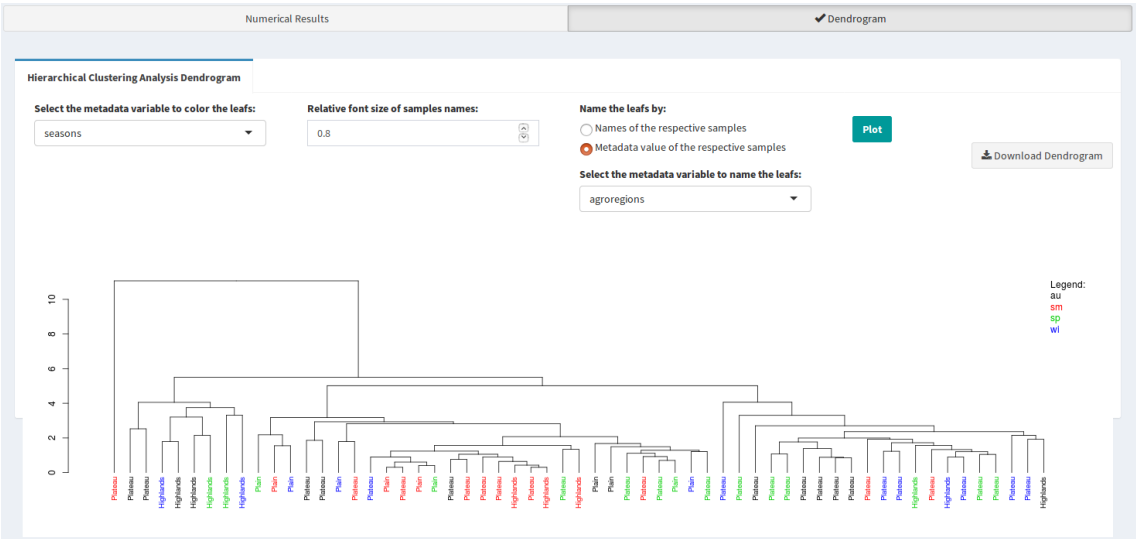
  Select the metadata variables to color the leafs by: the names of the samples are coloured according to the classes they belong to on the metadata variable chosen;

  Relative font size of the samples' names.

  Name the leafs by the names of the respective samples:




Or name the leafs by the metadata value of the respective samples. When this option is selected, a select input appears to select the metadata variable:




### K-means Clustering

For results of this type, **options used** that can be consulted are:

- Analysis Name;
- Name of the dataset used;
- If clustering was on samples or variables;
- Number of predefined clusters chosen.

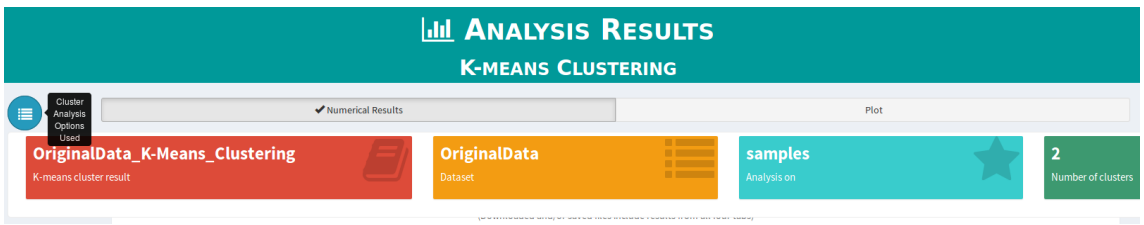



Figure 2.63: Layout of the dropdown menu of the options for K-means Clustering Analysis.

As regards to the actual **results**, there are two buttons, which allow the user to access the numerical results and the result's plot:

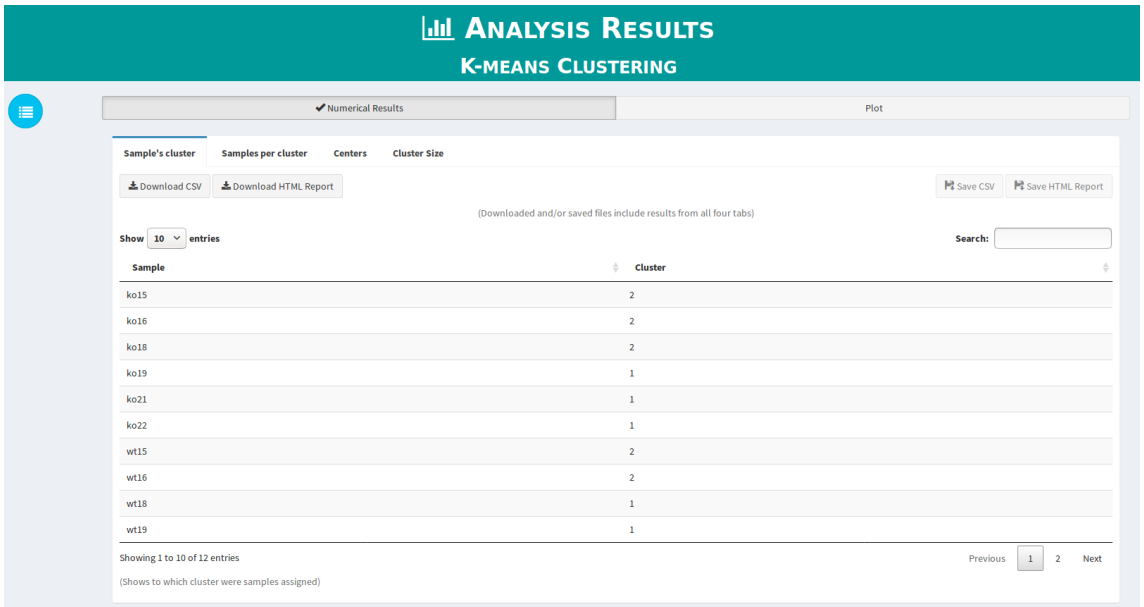


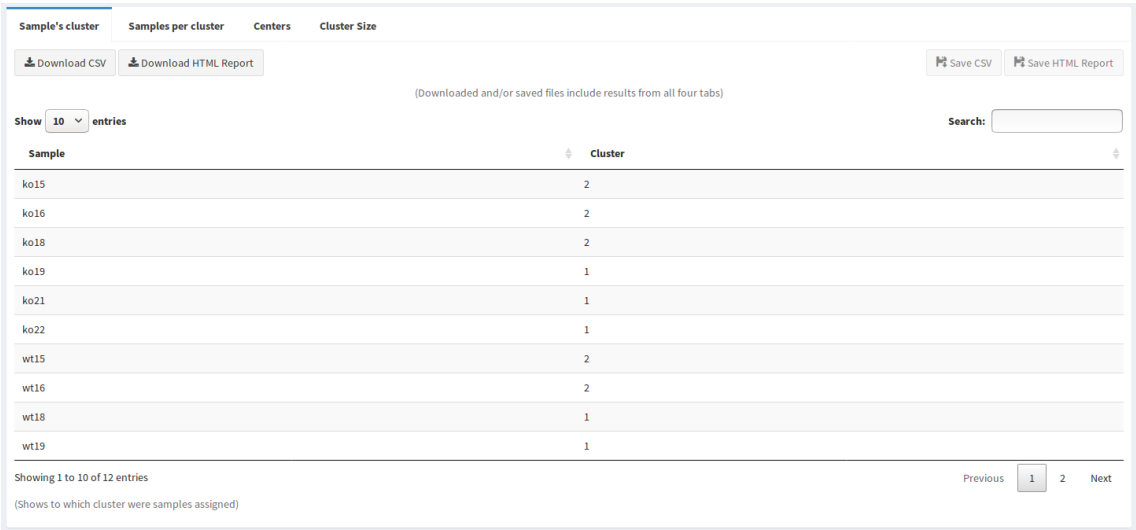

Figure 2.64: Layout of the results section for K-means Clustering Analysis.

Each one of these sections is detailed below.

### Numerical Results

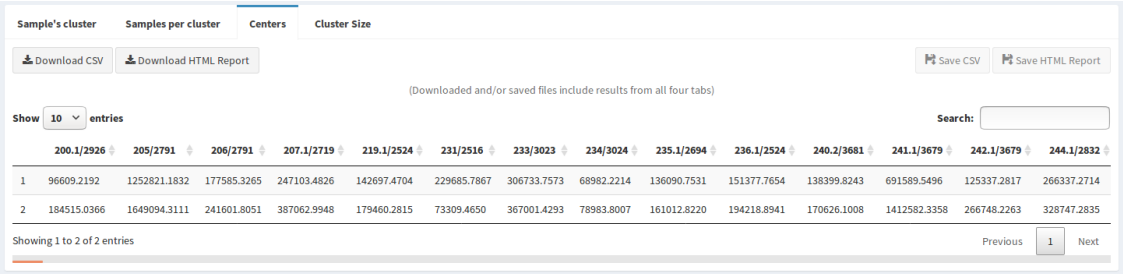
In the numerical results, there is a tabset panel with tabs with the following results:

- *Sample's cluster*: It contains a table with the cluster to which it belongs each sample;

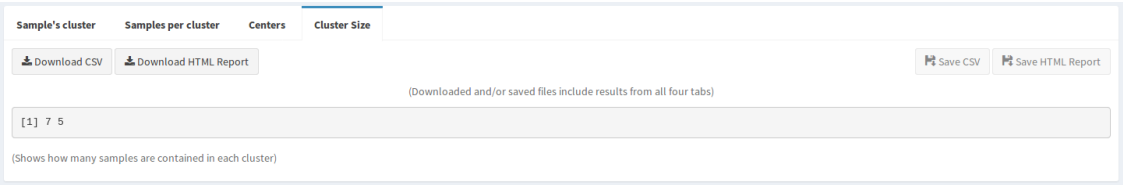


- *Samples per cluster*: It contains a table with the samples that belong to each cluster formed;


- *Centers*: It contains a table with the center values of each variable in each cluster;

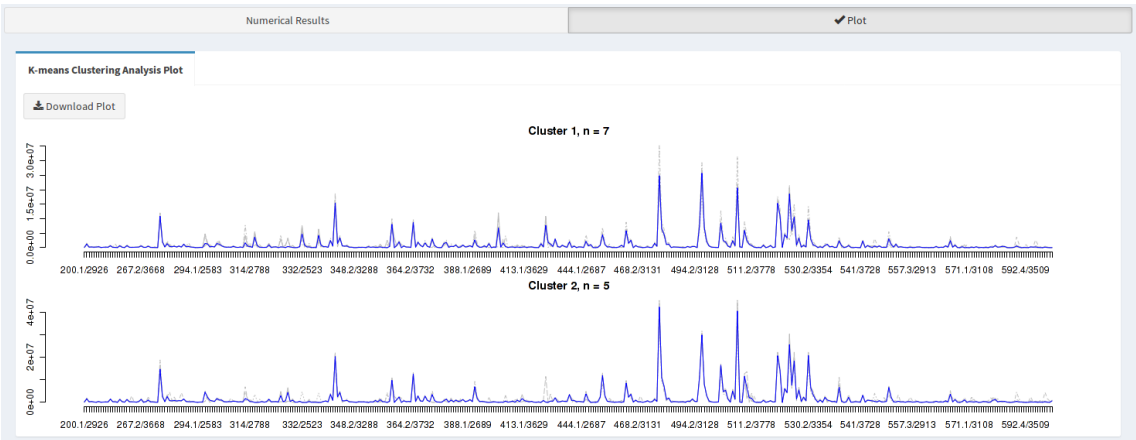



- *Cluster Size*: Contains the sizes of each cluster, i.e., the number of samples present in each cluster formed.




### Plot Results

For the plot results, a plot for each cluster is shown, with all the data values for each data variable present across the samples of the cluster, colored in blue, and the means of those data values for each data variable, colored in grey:



### 2.9.4 Machine Learning

#### Model Training

For results of this type, **options used** that can be consulted are:

- Name of the dataset used;
- Name of the metadata variable predicted;
- Validation method;
- Validation metric;

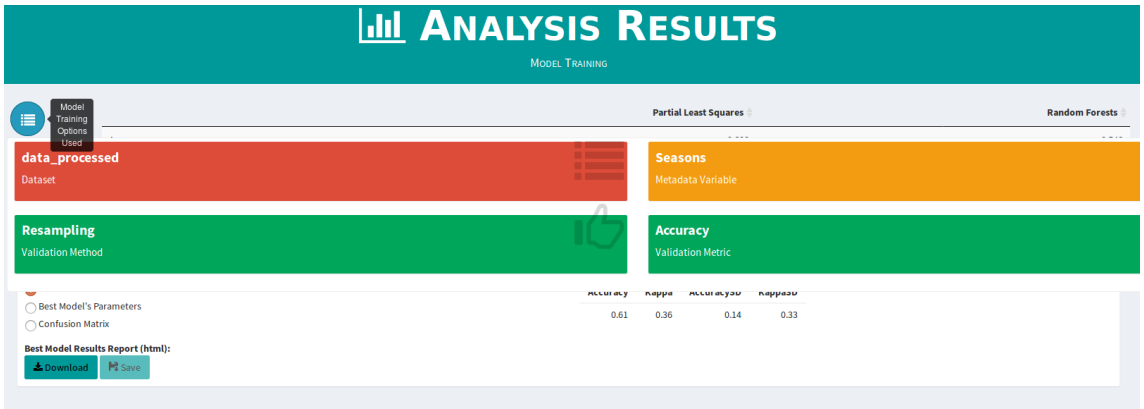



Figure 2.65: Layout of the dropdown menu of the options for Train Models Analysis.

As regards to the actual **results**, if more than one model was trained, there will be a summary table with the accuracies of all models trained in the analysis, so users can have a quick overview of each model, at the right of the options button.

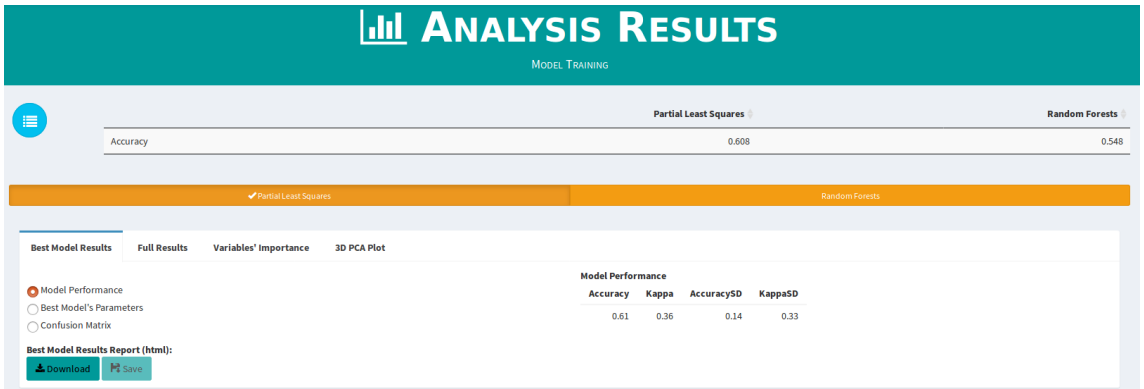
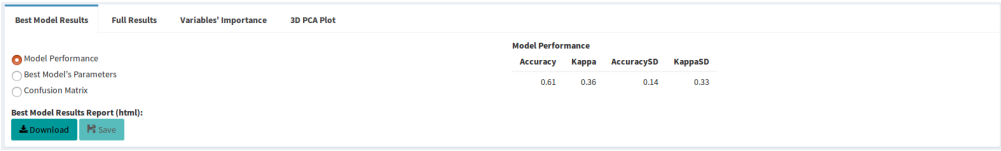


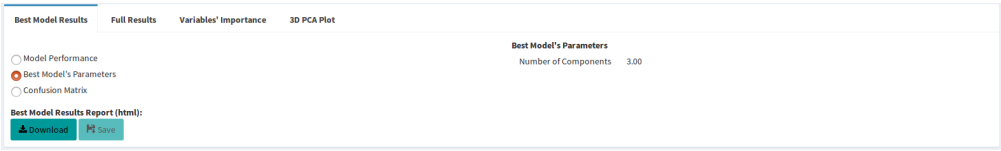

Figure 2.66: Overall layout of the Train Models Analysis results page.

Below this, there are one or more buttons, each representing a trained model. By clicking in one of the buttons, all the results regarding the respective model are shown below. By default, the results that are shown when the user is redirected to this page are the ones of the first model. The results are shown in the form of tabbed panels:

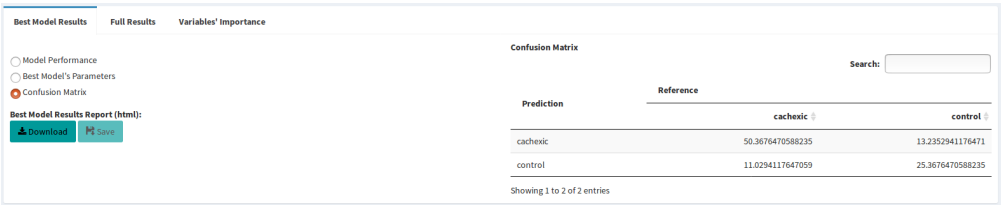
- *Best Model Results*: contains information on the best model obtained, suchs as the performance, parameters and confusion matrix;




(a)



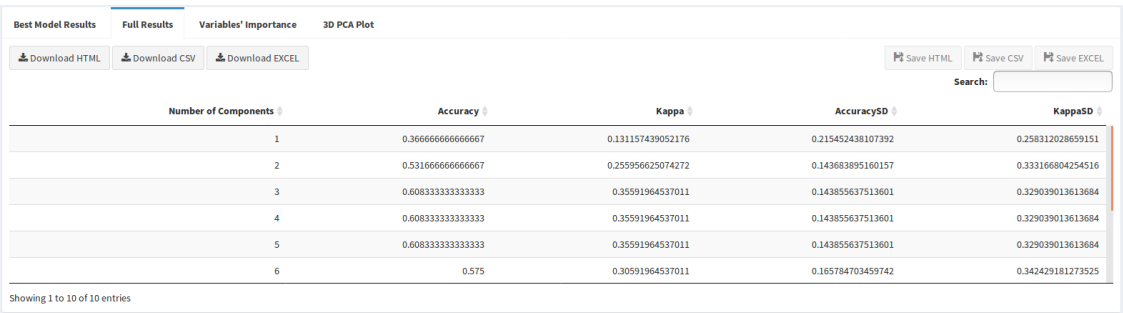

(b)



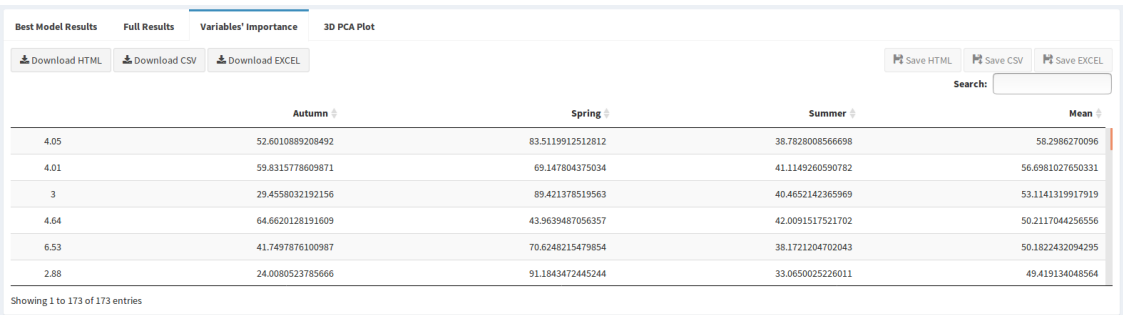

(c)

Figure 2.67: Sections regarding (a) model performance, (b) model's parameters and (c) confusion matrix, with an example for a pls model.

- *Full Results*: provides a table with all the results for the trained model, with the values of accuracy, kappa and respective standard deviations for each combination of parameter values tested;

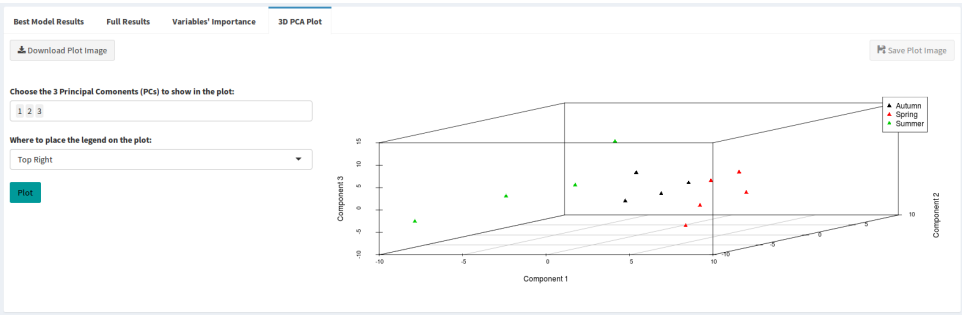



- *Variables' Importance*: provides a table with the importance of each variable used in the model training;

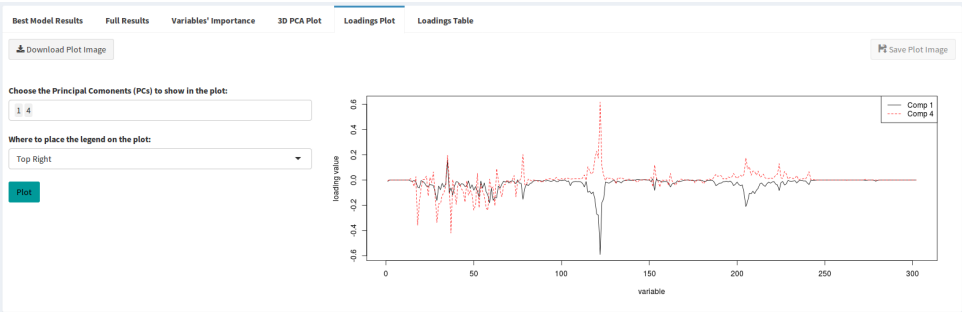



- *3D PCA Plot*: if the selected model is a PLS model whose best model has 3 or more components, this additional tab appears. Here, you can choose in a select input three of

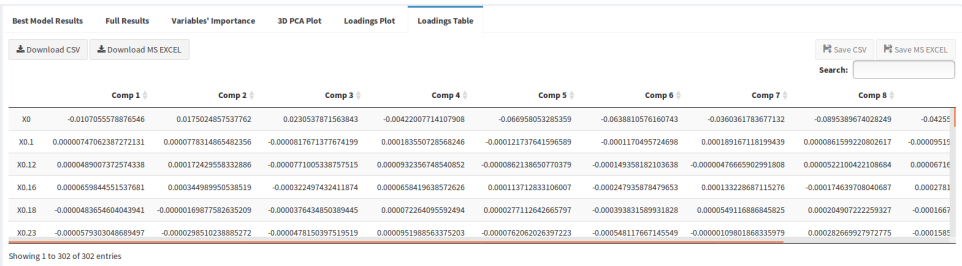
the components formed to appear on the plot. The data points are colored according to the metadata variable predicted.




- *Loadings Plot*: if the selected model is a PLS model, this additional tab appears. Here, you can choose in a select input the component(s) whose loadings you want to see in the plot and the place of the legend in the plot.




- *Loadings Table*: again, if the selected model is a PLS model, this additional tab appears. Here, you can see the numeric values of the variables' loadings for each component.



### New Samples Prediction

For results of this type, **options used** that can be consulted are:

- Name of the dataset used to train the model used to predict the new samples;
- Name of the metadata variable predicted;
- Characteristics of the model used for prediction: name of the analysis from which the model comes from, name of the model, and values of the model's parameters.

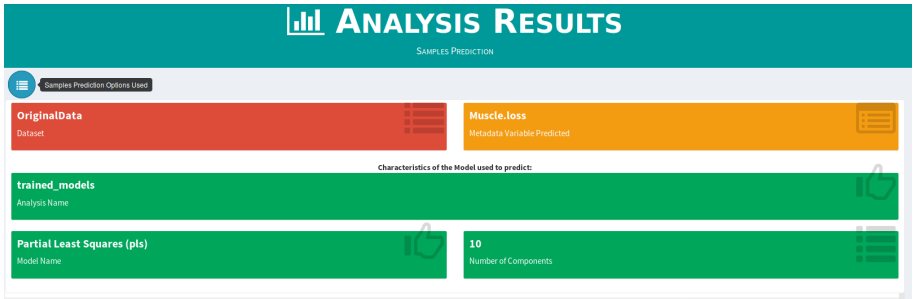



Figure 2.68: Layout of the dropdown menu of the options for Sample Prediction Analysis.

As regards to the actual **results**, they are presented in the form of a table, with the predicted class for each of the new samples submitted.

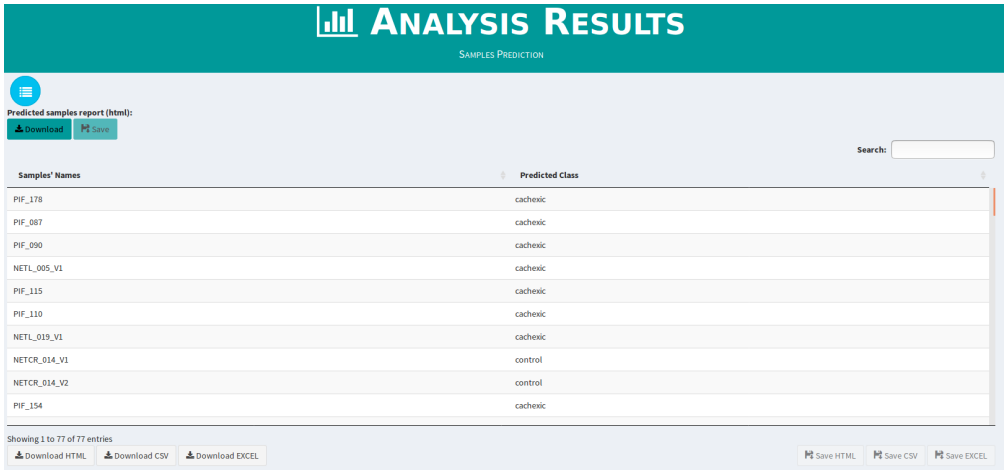



Figure 2.69: Layout of the results section for Sample Prediction Analysis.

### Results/Reports available to download/save

All tables in these results can be downloaded or saved (if logged in) in the CSV, MS EXCEL or HTML format.

For model training results, a report of the results for each model trained is present in the tabs *Best Model Results* and the respective options used in the analysis is available to download or save (if logged in), at the bottom of this tab.

An example of a report of this type is as follows:

## Best Random Forests (rf) model results for OriginalData dataset

```
## Report generated on 2018-03-05 16:29:37 using WEBSPECMINE
```

### Analysis Name

trained_models

### Options Used

- **Dataset** OriginalData
- **Metadata variable** Muscle.loss
- **Validation Method** Resampling
- **Validation Metric** Accuracy

### Model Performance

| Accuracy | Kappa | AccuracySD | KappaSD |
|---|---|---|---|
| 0.7560752 | 0.4982331 | 0.1204973 | 0.2347551 |

### Best Model's Parameters

| | FALSE |
|---|---|
| Number of randomly selected Predictors | 42 |

### Confusion Matrix

Search:

| Prediction | Reference | |
|---|---|---|
| | cachexic | control |
| cachexic | 50.3676470588235 | 13.2352941176471 |
| control | 11.0294117647059 | 25.3676470588235 |

Showing 1 to 2 of 2 entries
## End of report

For samples prediction, a report with the results table, the options used to perform the analysis and the characteristics of the model used in the prediction is available to download or save (if logged in) at the top of this page. An example of a report of this type is as follows (a and b):

## Predicted Samples Results Report

```
## Report generated on 2018-02-08 10:44:44 using WEBSPECMINE
```

### Analysis Names

samples_prediction2

### Options used

- **Metadata variable predicted:** Muscle.loss

### Characteristics of the model used to predict

- **Analysis Name:** trained_models
- **Dataset used to train the model:** OriginalData
- **Model Name:** Random Forests (rf)
- **Number of randomly selected Predictors**: 35

### Table with the predictions results

| Samples' Names | Predicted Class |
|---|---|
| PIF_178 | cachexic |
| PIF_087 | cachexic |
| PIF_090 | cachexic |
| NETL_005_V1 | cachexic |
| PIF_115 | cachexic |
| PIF_110 | cachexic |
| NETL_019_V1 | cachexic |
| NETCR_014_V1 | cachexic |
| NETCR_014_V2 | cachexic |
| PIF_154 | cachexic |
| NETL_022_V1 | cachexic |
| NETL_022_V2 | cachexic |
| NETL_008_V1 | cachexic |
| PIF_146 | cachexic |
| PIF_164 | control |
| NETL_013_V1 | control |
| PIF_188 | control |
| PIF_195 | control |
| NETCR_015_V1 | control |
| PIF_102 | control |
| NETL_010_V1 | control |
| NETL_010_V2 | control |
| NETL_001_V1 | control |
| NETCR_015_V2 | control |
| NETCR_005_V1 | control |
| PIF_111 | control |
| PIF_171 | control |
| NETCR_008_V1 | control |
| NETCR_008_V2 | control |
| NETL_017_V1 | control |
| NETL_017_V2 | control |
| NETL_002_V1 | control |
| NETL_002_V2 | control |
| PIF_190 | control |
| NETCR_009_V1 | control |
| NETCR_009_V2 | control |
| NETL_007_V1 | control |
| PIF_112 | control |
| NETCR_019_V2 | control |
| NETL_012_V1 | control |
| NETL_012_V2 | control |
| NETL_003_V1 | control |
| NETL_003_V2 | control |

## End of report

(a) (b)

### 2.9.5 Feature Selection

For results of this type, it is possible to see the following **options chosen**:

- Name of the dataset used;
- Name of the metadata variable to be predicted;
- Selection method;
- Function for model fitting, prediction and variable importance/filtering;
- Validation method.

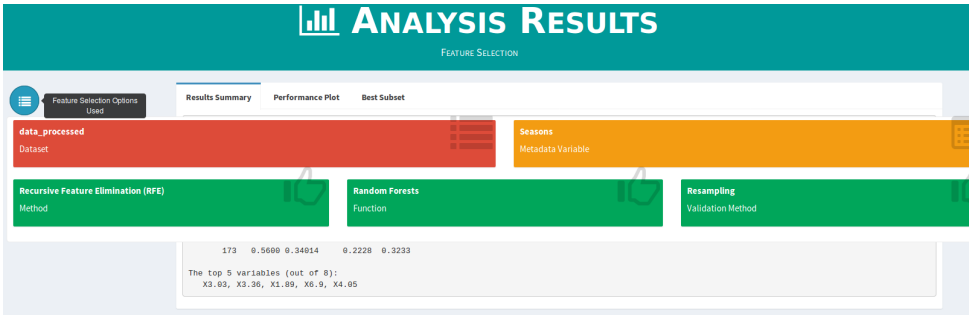
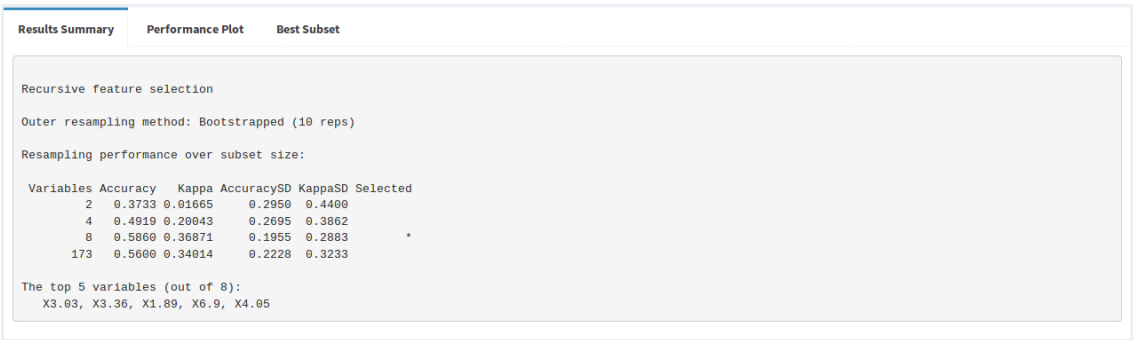


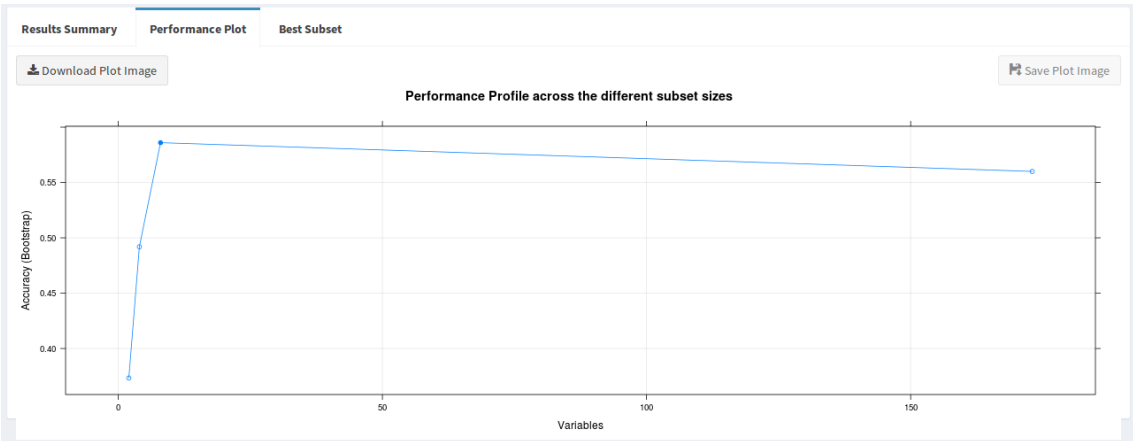

Figure 2.71: Layout of the dropdown menu of the options for Feature Selection Analysis.

The actual **results** are disposed in a tabset panel:

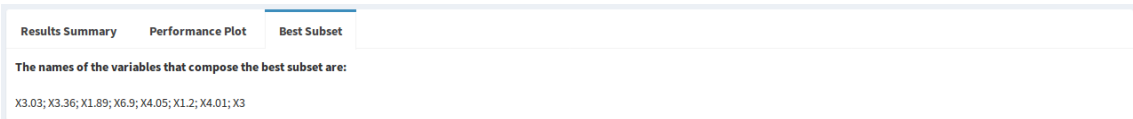
- *Results Summary*: contains a brief summary of the results;




- *Performance Plot*: if recursive feature selection was used, this tab appears. It contains a performance plot that shows the accuracy across the different subset sizes;



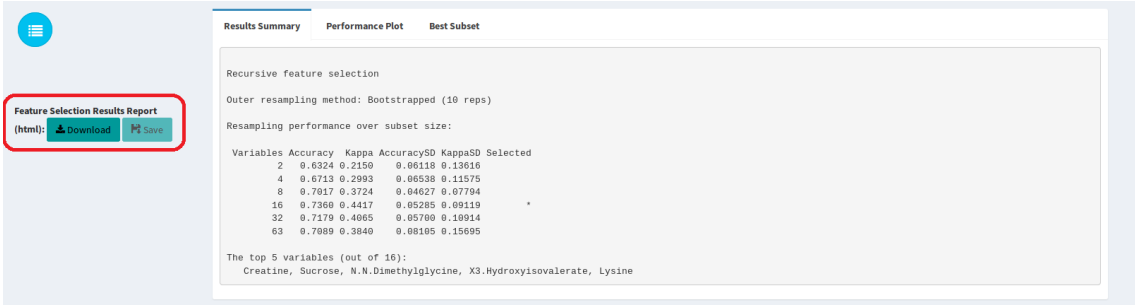

- *Best Subset*: contains the names of the variables that make up the best subset.



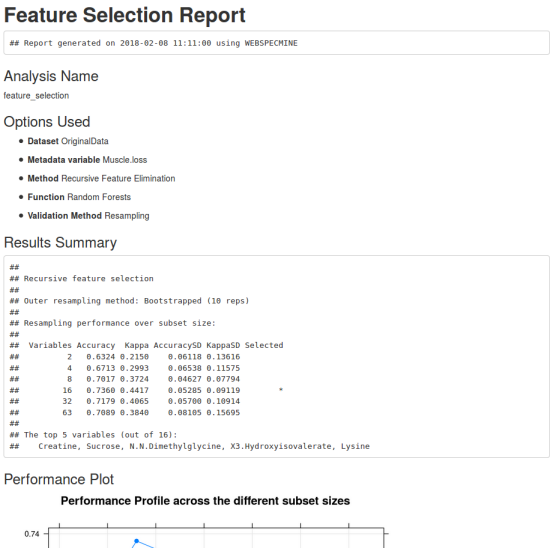
### Results/Reports available to download/save

In the *Performance Plot* tab, the plot image can be saved (if the user is logged in) or downloaded in PDF format.

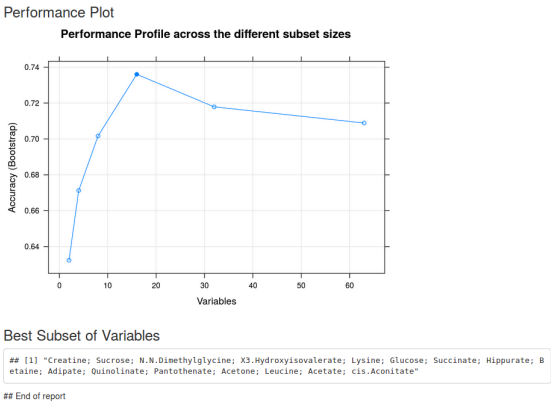
A report can also be saved (if the user is logged in) or downloaded, containing all the results shown in the tab panels and the options chosen to run the analysis.




(a)




(b)




(c)

Figure 2.72: (a) A report on the results on the feature selection can be downloaded or saved (if logged in) through the buttons present at the left of the page, below the options button. An example of a report of this type is present at (b) and (c).

### 2.9.6 Metabolite Identification

#### LC-MS Data

For results of this type, it is possible to see the following **options chosen** and used in the respective identification:

- Name of the dataset used for the identification;
- Name of the metadata variable used to help in the identification.

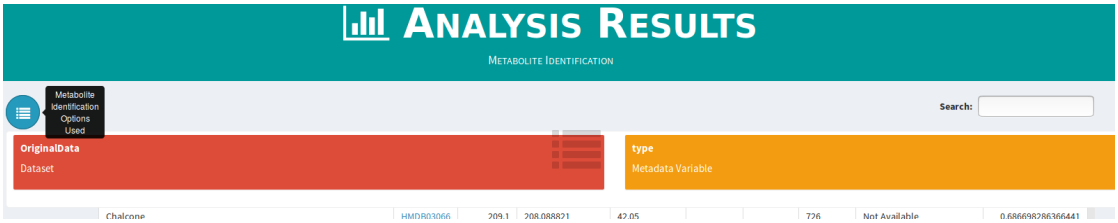
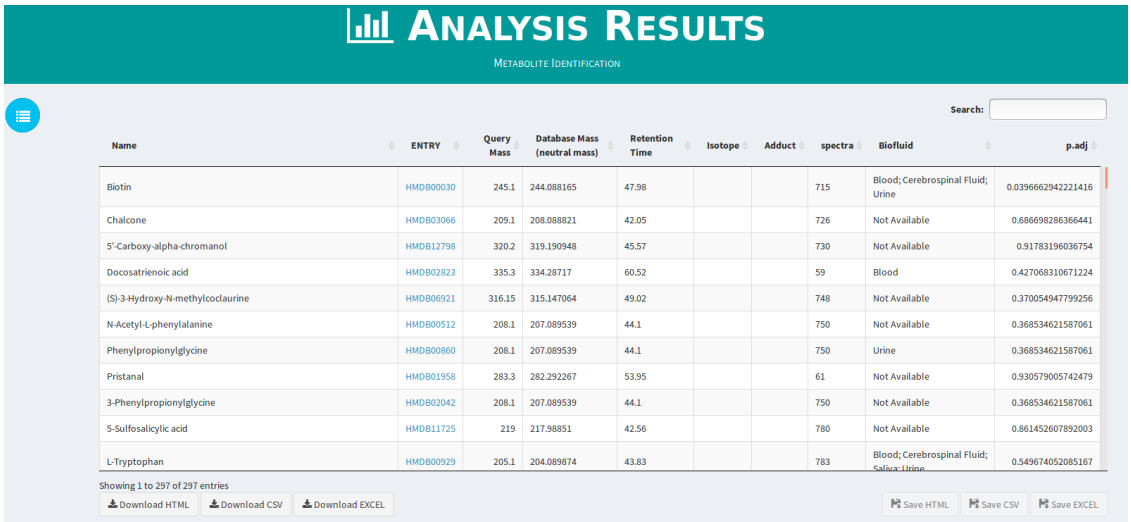



Figure 2.73: Layout of the dropdown menu of the options for LC-MS metabolite identification.

The actual **results** are available in the form of a results table, where each line corresponds to an identified metabolite. For each of these metabolites, there is information on:

- HMDB entry number: with a link to the HMDB webpage of the respective metabolite;
- Query and Theoretical masses;
- Retention time;
- Isotopes;
- Adducts;
- Spectra;
- Biofluids where was found;
- Adjusted p-value.




| Name | ENTRY | Query Mass | Database Mass (neutral mass) | Retention Time | Isotope | Adduct | spectra | Biofluid | p.adj |
|---|---|---|---|---|---|---|---|---|---|
| Biotin | HMDB00030 | 245.1 | 244.088165 | 47.98 | | | 715 | Blood; Cerebrospinal Fluid; Urine | 0.0396662942221416 |
| Chalcone | HMDB03066 | 209.1 | 208.088821 | 42.05 | | | 726 | Not Available | 0.686698286366441 |
| 5'-Carboxy-alpha-chromanol | HMDB12798 | 320.2 | 319.190948 | 45.57 | | | 730 | Not Available | 0.91783196036754 |
| Docosatrienoic acid | HMDB02823 | 335.3 | 334.28717 | 60.52 | | | 59 | Blood | 0.427068310671224 |
| (S)-3-Hydroxy-N-methylcoclaurine | HMDB06921 | 316.15 | 315.147064 | 49.02 | | | 748 | Not Available | 0.370054947799256 |
| N-Acetyl-L-phenylalanine | HMDB00512 | 208.1 | 207.089539 | 44.1 | | | 750 | Not Available | 0.368534621587061 |
| Phenylpropionylglycine | HMDB00860 | 208.1 | 207.089539 | 44.1 | | | 750 | Urine | 0.368534621587061 |
| Pristanal | HMDB01958 | 283.3 | 282.292267 | 53.95 | | | 61 | Not Available | 0.930579005742479 |
| 3-Phenylpropionylglycine | HMDB02042 | 208.1 | 207.089539 | 44.1 | | | 750 | Not Available | 0.368534621587061 |
| 5-Sulfosalicylic acid | HMDB11725 | 219 | 217.98851 | 42.56 | | | 780 | Not Available | 0.861452607892003 |
| L-Tryptophan | HMDB00929 | 205.1 | 204.089874 | 43.83 | | | 783 | Blood; Cerebrospinal Fluid; Saliva; Urine | 0.549674052085167 |

Figure 2.74: Layout of the LC-MS metabolite identification results page.

#### NMR Peaks

For the results of this page, again, the available **options used** in the respective identification are:

- Dataset used for the identification;
- PPM tolerance used;
- Number of top metabolites matched to show per cluster;

- Parameters for the construction of clusters: correlation method, correlation value (if the value was provided by the user), minimum number of peaks chosen for a cluster, maximum number of peaks in a cluster (if provided by the user for the calculation of the optimum correlation value);
- Parameters to filter the reference metabolites: frequency, nucleus, and, if used, solvent, pH and temperature.

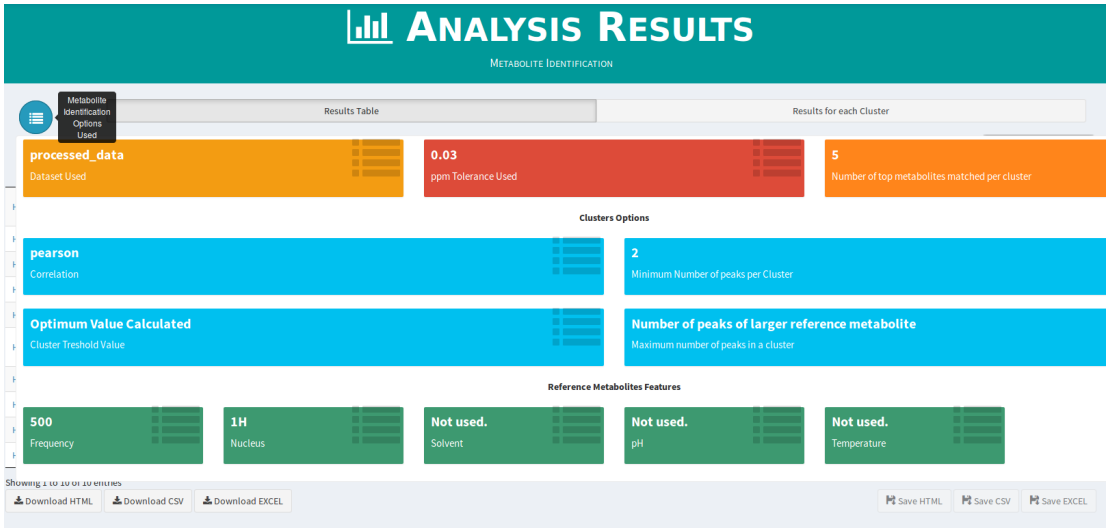
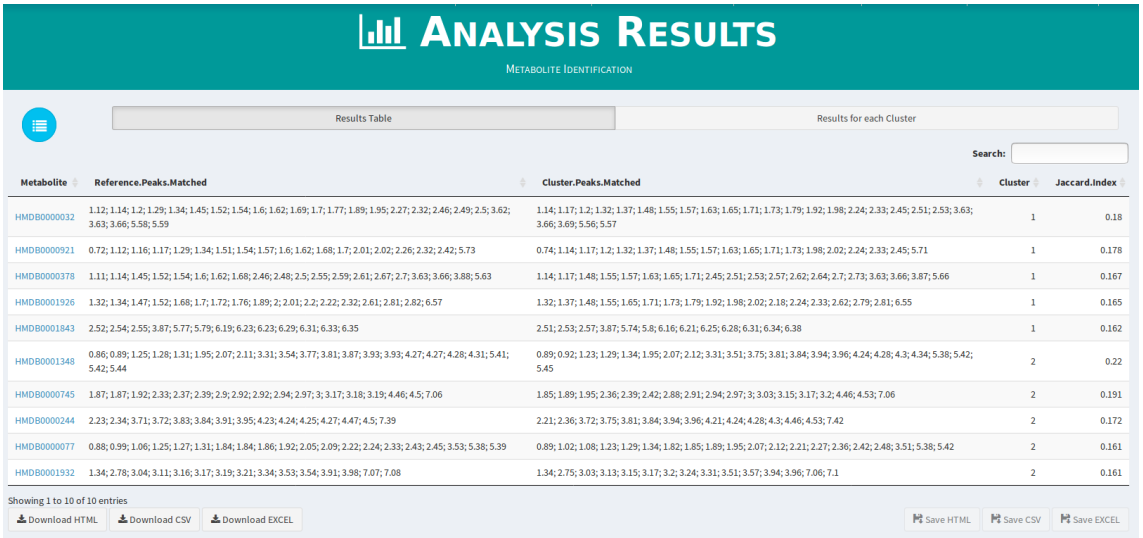



Figure 2.75: Layout of the dropdown menu of the options for NMR metabolite identification.

At the right of the options button, there are two buttons, which allow the user to see the different types of results obtained, shown below these buttons:




| Metabolite | Reference.Peaks.Matched | Cluster.Peaks.Matched | Cluster | Jaccard.Index |
|---|---|---|---|---|
| HMDB0000032 | 1.12; 1.14; 1.2; 1.29; 1.34; 1.45; 1.52; 1.54; 1.6; 1.62; 1.69; 1.7; 1.77; 1.89; 1.95; 2.27; 2.32; 2.46; 2.49; 2.5; 3.62; 3.63; 3.66; 5.58; 5.59 | 1.14; 1.17; 1.2; 1.32; 1.37; 1.48; 1.55; 1.57; 1.63; 1.65; 1.71; 1.73; 1.79; 1.92; 1.98; 2.24; 2.33; 2.45; 2.51; 2.53; 3.63; 3.66; 3.69; 5.56; 5.57 | 1 | 0.18 |
| HMDB0000921 | 0.72; 1.12; 1.16; 1.17; 1.29; 1.34; 1.51; 1.54; 1.57; 1.6; 1.62; 1.68; 1.7; 2.01; 2.02; 2.26; 2.32; 2.42; 5.73 | 0.74; 1.14; 1.17; 1.2; 1.32; 1.37; 1.48; 1.55; 1.57; 1.63; 1.65; 1.71; 1.73; 1.98; 2.02; 2.24; 2.33; 2.45; 5.71 | 1 | 0.178 |
| HMDB0000378 | 1.11; 1.14; 1.45; 1.52; 1.54; 1.6; 1.62; 1.68; 2.46; 2.48; 2.5; 2.55; 2.59; 2.61; 2.67; 2.7; 3.63; 3.66; 3.88; 5.63 | 1.14; 1.17; 1.48; 1.55; 1.57; 1.63; 1.65; 1.71; 2.45; 2.51; 2.53; 2.57; 2.62; 2.64; 2.7; 2.73; 3.63; 3.66; 3.87; 5.66 | 1 | 0.167 |
| HMDB0001926 | 1.32; 1.34; 1.47; 1.52; 1.68; 1.7; 1.72; 1.76; 1.89; 2; 2.01; 2.2; 2.22; 2.32; 2.61; 2.81; 2.82; 6.57 | 1.32; 1.37; 1.48; 1.55; 1.65; 1.71; 1.73; 1.79; 1.92; 1.98; 2.02; 2.18; 2.24; 2.33; 2.62; 2.79; 2.81; 6.55 | 1 | 0.165 |
| HMDB0001843 | 2.52; 2.54; 2.55; 3.87; 5.77; 5.79; 6.19; 6.23; 6.23; 6.29; 6.31; 6.33; 6.35 | 2.51; 2.53; 2.57; 3.87; 5.74; 5.8; 6.16; 6.21; 6.25; 6.28; 6.31; 6.34; 6.38 | 1 | 0.162 |
| HMDB0001348 | 0.86; 0.89; 1.25; 1.28; 1.31; 1.95; 2.07; 2.11; 3.31; 3.54; 3.77; 3.81; 3.87; 3.93; 3.93; 4.27; 4.27; 4.28; 4.31; 5.41; 5.42; 5.44 | 0.89; 0.92; 1.23; 1.29; 1.34; 1.95; 2.07; 2.12; 3.31; 3.51; 3.75; 3.81; 3.84; 3.94; 3.96; 4.24; 4.28; 4.3; 4.34; 5.38; 5.42; 5.45 | 2 | 0.22 |
| HMDB0000745 | 1.87; 1.87; 1.92; 2.33; 2.37; 2.39; 2.9; 2.92; 2.92; 2.94; 2.97; 3; 3.17; 3.18; 3.19; 4.46; 4.5; 7.06 | 1.85; 1.89; 1.95; 2.36; 2.39; 2.42; 2.88; 2.91; 2.94; 2.97; 3; 3.03; 3.15; 3.17; 3.2; 4.46; 4.53; 7.06 | 2 | 0.191 |
| HMDB0000244 | 2.23; 2.34; 3.71; 3.72; 3.83; 3.84; 3.91; 3.95; 4.23; 4.24; 4.25; 4.27; 4.47; 4.5; 7.39 | 2.21; 2.36; 3.72; 3.75; 3.81; 3.84; 3.94; 3.96; 4.21; 4.24; 4.28; 4.3; 4.46; 4.53; 7.42 | 2 | 0.172 |
| HMDB0000077 | 0.88; 0.99; 1.06; 1.25; 1.27; 1.31; 1.84; 1.84; 1.86; 1.92; 2.05; 2.09; 2.22; 2.24; 2.33; 2.43; 2.45; 3.53; 5.38; 5.39 | 0.89; 1.02; 1.08; 1.23; 1.29; 1.34; 1.82; 1.85; 1.89; 1.95; 2.07; 2.12; 2.21; 2.27; 2.36; 2.42; 2.48; 3.51; 5.38; 5.42 | 2 | 0.161 |
| HMDB0001932 | 1.34; 2.78; 3.04; 3.11; 3.16; 3.17; 3.19; 3.21; 3.34; 3.53; 3.54; 3.91; 3.98; 7.07; 7.08 | 1.34; 2.75; 3.03; 3.13; 3.15; 3.17; 3.2; 3.24; 3.31; 3.51; 3.57; 3.94; 3.96; 7.06; 7.1 | 2 | 0.161 |

The **"Results table"** option leads to a table where each line corresponds to an identified metabolite. All metabolites identified in each cluster are here present, which can lead to repetitions if the same metabolite matched different clusters. The information here provided for each identified metabolite includes:

- HMDB entry number: with a link to the HMDB webpage of the respective metabolite;
- Cluster and Reference peaks that were matched;

- The number of the cluster;
- Jaccard Index Score.

Results Table | Results for each Cluster

Search:

| Metabolite | Reference.Peaks.Matched | Cluster.Peaks.Matched | Cluster | Jaccard.Index |
|---|---|---|---|---|
| HMDB0000921 | 0.72; 0.87; 0.91; 0.93; 1.01; 1.03; 1.04; 1.1; 1.12; 1.25; 1.27; 1.29; 1.34; 1.38; 1.54; 1.57; 1.6; 1.62; 1.68; 1.83; 1.85; 2.01; 2.02; 2.03; 2.05; 2.26; 2.28; 2.32; 2.36; 2.39; 2.42; 2.45 | 0.74; 0.89; 0.92; 0.96; 0.99; 1.02; 1.05; 1.08; 1.14; 1.23; 1.29; 1.32; 1.34; 1.37; 1.55; 1.57; 1.63; 1.65; 1.71; 1.82; 1.85; 1.98; 2.02; 2.04; 2.07; 2.24; 2.27; 2.31; 2.33; 2.39; 2.42; 2.45 | 1 | 0.333 |
| HMDB0000077 | 0.88; 0.98; 0.99; 1; 1.02; 1.06; 1.12; 1.25; 1.27; 1.29; 1.31; 1.52; 1.54; 1.64; 1.65; 1.68; 1.84; 1.84; 1.95; 2.05; 2.07; 2.09; 2.09; 2.12; 2.22; 2.22; 2.24; 2.3; 2.31; 2.43; 2.45; 2.49; 5.38 | 0.89; 0.96; 0.99; 1.02; 1.05; 1.08; 1.14; 1.23; 1.29; 1.32; 1.34; 1.55; 1.57; 1.63; 1.65; 1.71; 1.82; 1.85; 1.98; 2.02; 2.04; 2.07; 2.12; 2.15; 2.21; 2.24; 2.27; 2.31; 2.33; 2.42; 2.45; 2.51; 5.38 | 1 | 0.284 |
| HMDB0000871 | 0.71; 0.86; 0.91; 1.01; 1.05; 1.06; 1.07; 1.11; 1.22; 1.27; 1.29; 1.33; 1.34; 1.53; 1.54; 1.69; 1.79; 1.81; 1.82; 1.99; 2; 2.01; 2.04; 2.09; 2.24; 2.27; 2.28; 2.29; 2.3; 2.37; 2.4; 2.43 | 0.74; 0.89; 0.92; 0.99; 1.02; 1.05; 1.08; 1.14; 1.23; 1.29; 1.32; 1.34; 1.37; 1.55; 1.57; 1.71; 1.79; 1.82; 1.85; 1.98; 2.02; 2.04; 2.07; 2.12; 2.21; 2.24; 2.27; 2.31; 2.33; 2.39; 2.42; 2.45 | 1 | 0.276 |
| HMDB0000610 | 0.87; 0.9; 0.93; 0.96; 0.99; 1.03; 1.06; 1.11; 1.2; 1.26; 1.31; 1.34; 1.35; 1.53; 1.54; 1.6; 1.62; 1.8; 1.81; 1.82; 1.95; 2; 2.01; 2.04; 2.09; 2.12; 2.26; 2.27; 2.29; 2.31; 2.77; 2.78; 2.82; 4.59; 4.61; 5.37 | 0.89; 0.92; 0.96; 0.99; 1.02; 1.05; 1.08; 1.14; 1.23; 1.29; 1.32; 1.34; 1.37; 1.55; 1.57; 1.63; 1.65; 1.79; 1.82; 1.85; 1.98; 2.02; 2.04; 2.07; 2.12; 2.15; 2.24; 2.27; 2.31; 2.33; 2.79; 2.81; 2.85; 4.59; 4.64; 5.38 | 1 | 0.273 |
| HMDB0001830 | 0.96; 0.97; 0.99; 1.02; 1.05; 1.11; 1.2; 1.26; 1.3; 1.31; 1.53; 1.54; 1.6; 1.64; 1.68; 1.76; 1.86; 2.02; 2.03; 2.04; 2.09; 2.13; 2.16; 2.18; 2.21; 2.27; 2.28; 2.3; 2.37; 2.4; 2.43; 2.53; 2.54; 2.56 | 0.96; 0.99; 1.02; 1.05; 1.08; 1.14; 1.23; 1.29; 1.32; 1.34; 1.55; 1.57; 1.63; 1.65; 1.71; 1.79; 1.85; 2.02; 2.04; 2.07; 2.12; 2.15; 2.18; 2.21; 2.24; 2.27; 2.31; 2.33; 2.39; 2.42; 2.45; 2.51; 2.53; 2.57 | 1 | 0.264 |
| HMDB0000174 | 1.21; 3.44; 3.45; 3.45; 3.63; 3.64; 3.65; 3.66; 3.75; 3.76; 3.78; 3.81; 3.87; 4.18; 4.21; 4.55; 4.57; 5.21; 5.21 | 1.2; 3.42; 3.46; 3.48; 3.6; 3.63; 3.66; 3.69; 3.75; 3.78; 3.81; 3.84; 3.9; 4.21; 4.24; 4.53; 4.56; 5.18; 5.24 | 2 | 0.209 |
| HMDB0000055 | 3.28; 3.37; 3.41; 3.43; 3.45; 3.48; 3.51; 3.54; 3.58; 3.6; 3.63; 3.66; 3.74; 3.75; 3.81; 3.82; 3.88; 3.98; 4.51; 4.67; 4.68; 5.23 | 3.27; 3.39; 3.42; 3.46; 3.48; 3.51; 3.54; 3.57; 3.6; 3.63; 3.66; 3.69; 3.75; 3.78; 3.81; 3.84; 3.9; 4.01; 4.53; 4.66; 4.7; 5.24 | 2 | 0.202 |
| HMDB0000098 | 3.2; 3.21; 3.29; 3.4; 3.42; 3.44; 3.5; 3.51; 3.52; 3.59; 3.6; 3.6; 3.63; 3.67; 3.9; 4.56; 4.57; 5.18 | 3.2; 3.24; 3.27; 3.39; 3.42; 3.46; 3.48; 3.51; 3.54; 3.57; 3.6; 3.63; 3.66; 3.69; 3.9; 4.53; 4.56; 5.18 | 2 | 0.194 |

Showing 1 to 10 of 10 entries

Download HTML | Download CSV | Download EXCEL | Save HTML | Save CSV | Save EXCEL

The **"Results for each Cluster"** option leads to more detailed information on the matches obtained in each cluster. A select input with all the clusters obtained is available, so that the user can see the results of the chosen cluster.

These results are organized in three boxes:

- *Scores*: contains the scores of the top matches. Each match is represented by the HMDB entry, with a link to the HMDB webpage of the respective metabolite;
- *Cluster Peaks*: contains the ppm peaks of the cluster;
- *Top Metabolites*: for each match, it contains the reference ppm peaks and the ppm peaks that were matched between the cluster and reference metabolite.

Results Table | Results for each Cluster

Cluster1

Scores

| | Var.2 |
|---|---|
| HMDB0000921 | 0.333 |
| HMDB0000077 | 0.284 |
| HMDB0000871 | 0.276 |
| HMDB0000610 | 0.273 |

Cluster Peaks

| ppm | Intensity |
|---|---|
| 0.74 | -1.66533453693773e-16 |
| 0.89 | 3.71501874402158e-16 |
| 0.92 | 2.64728810454114e-16 |
| 0.96 | -5.0444591079066e-17 |

Top Metabolites

Select a metabolite to see the result

HMDB0000921

Showing results for Metabolite HMDB0000921 from Cluster1

Reference Peaks

| ppm |
|---|
| 0.72 |
| 0.87 |
| 0.88 |
| 0.91 |
| 0.93 |

Matched Peaks

| Matched Cluster Peaks | Matched Reference Peaks |
|---|---|
| 0.74 | 0.72 |
| 0.89 | 0.87 |
| 0.92 | 0.91 |
| 0.96 | 0.93 |
| 0.99 | 1.01 |

However, in case no metabolites matched a certain cluster, only two boxes, side by side, will appear. One with the cluster peaks and the other with the message "No metabolites matched this cluster".

So that the user does not need to enter in each cluster to know if it got macthes or not, the clusters with no matches are followed by the message "No matches" in the select input.

### Available results to save/download

The results tables are available for save (if the user is logged in) or download, in the CSV, MS EXCEL or HTML formats.

## 2.9.7 Regression Analysis

### Linear Regression Analysis

For the results of this page, the available **options used** are:

- Analysis Name;
- Name of the dataset used;
- Metadata variables used;
- Formula used.

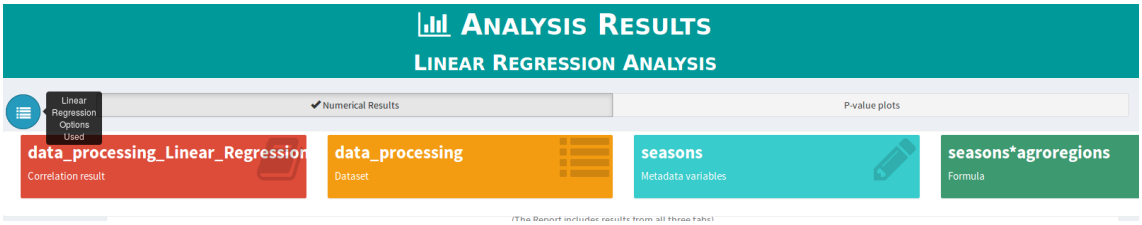
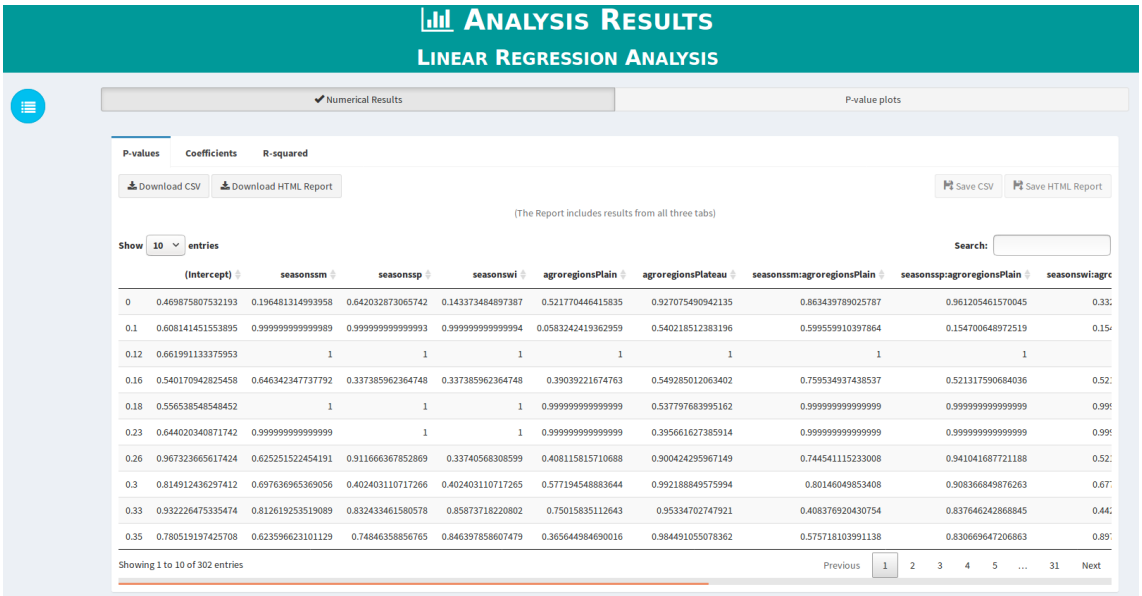



Figure 2.76: Layout of the dropdown menu of the options for Linear Regression Analysis.

At the right of the options button, there are two buttons, which allow the user to see the numerical results and the result's plot:




As regards to the **numerical results**, a tabset panel with the following results can be seen:

- *P-values*: It contains a table with the p-values of the comparisons between the classes of the metadata variables chosen, making use of the formula specified, for each linear regression on a data variable;

P-values Coefficients R-squared

Download CSV Download HTML Report Save CSV Save HTML Report

(The Report includes results from all three tabs)

Show 10 entries Search:

| | (Intercept) | seasonssm | seasonssp | seasonswi | agroregionsPlain | agroregionsPlateau | seasonssm:agroregionsPlain | seasonssp:agroregionsPlain | seasonswi:agro |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.469875807532193 | 0.196481314993958 | 0.642032873065742 | 0.143373484897387 | 0.521770446415835 | 0.927075490942135 | 0.863439789025787 | 0.961205461570045 | 0.33 |
| 0.1 | 0.608141451553895 | 0.999999999999989 | 0.999999999999993 | 0.999999999999994 | 0.0583242419362959 | 0.540218512383196 | 0.599559910397864 | 0.154700648972519 | 0.15 |
| 0.12 | 0.661991133375953 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 0.16 | 0.540170942825458 | 0.646342347737792 | 0.337385962364748 | 0.337385962364748 | 0.39039221674763 | 0.549285012063402 | 0.759534937438537 | 0.521317590684036 | 0.52 |
| 0.18 | 0.556538548548452 | 1 | 1 | 1 | 0.999999999999999 | 0.537797683995162 | 0.999999999999999 | 0.999999999999999 | 0.99 |
| 0.23 | 0.644020340871742 | 0.999999999999999 | 1 | 1 | 0.999999999999999 | 0.395661627385914 | 0.999999999999999 | 0.999999999999999 | 0.99 |
| 0.26 | 0.967323665617424 | 0.625251522454191 | 0.911666367852869 | 0.33740568308599 | 0.408115815710688 | 0.900424295967149 | 0.744541115233008 | 0.941041687721188 | 0.52 |
| 0.3 | 0.814912436297412 | 0.697636965369056 | 0.402403110717266 | 0.402403110717265 | 0.577194548883644 | 0.992188849575994 | 0.80146049853408 | 0.908366849876263 | 0.67 |
| 0.33 | 0.932226475335474 | 0.812619253519089 | 0.832433461580578 | 0.85873718220802 | 0.75015835112643 | 0.95334702747921 | 0.408376920430754 | 0.837646242868845 | 0.44 |
| 0.35 | 0.780519197425708 | 0.623596623101129 | 0.74846358856765 | 0.846397858607479 | 0.365644984690016 | 0.984491055078362 | 0.575718103991138 | 0.830669647206863 | 0.89 |

Showing 1 to 10 of 302 entries Previous 1 2 3 4 5 ... 31 Next

- *Coefficients*: It contains a table with the coefficient values of the comparisons between the classes of the metadata variables chosen, making use of the specified formula, for each linear regression on a data variable;

P-values Coefficients R-squared

Download CSV Download HTML Report Save CSV Save HTML Report

(The Report includes results from all three tabs)

Show 10 entries Search:

| | (Intercept) | seasonssm | seasonssp | seasonswi | agroregionsPlain | agroregionsPlateau | seasonssm:agroregionsPlain | seasonssp:agroregionsPlain |
|---|---|---|---|---|---|---|---|---|
| 0 | -0.320846479362002 | -0.815997779949908 | 0.291388039227598 | 0.926836425082108 | 0.449426841878778 | 0.0462022239516954 | 0.161558059533093 | 0.0456820681415225 |
| 0.1 | -0.298432228351458 | -1.1008044561259e-14 | -7.34279386844011e-15 | -6.49980643597303e-15 | 1.77390825286228 | 0.406710913022516 | -0.648308266666609 | -1.77390825286228 |
| 0.12 | -0.255980309884728 | 4.42504926493762e-16 | 2.71055847377609e-17 | 6.70638597060664e-17 | -1.13073364336417e-16 | 3.31280664613475e-16 | -2.08133329565391e-16 | 6.5003801918465e-16 |
| 0.16 | 0.367045402893399 | 0.388486532731107 | -0.815337353949 | -0.815337353949002 | -0.815337353949003 | -0.409084315753307 | -0.388486532731106 | 0.815337353949004 |
| 0.18 | -0.335805360871845 | 5.27412754618147e-16 | 3.36415509284067e-16 | -4.63035427877678e-16 | -9.68753202371161e-16 | 0.401284258160858 | 1.49054209061081e-15 | 9.35108373538052e-16 |
| 0.23 | -0.28086762061259 | -6.88411899256587e-16 | -1.98648726325143e-16 | -3.65121438582392e-16 | -8.94233725265687e-16 | 0.590264755586469 | 1.37786025660674e-15 | 1.28103190175595e-15 |
| 0.26 | 0.0232963675187002 | -0.393310279436462 | -0.0892255761555326 | 0.775334379257214 | -0.746541826960395 | 0.0811377809783613 | 0.393310279436462 | 0.0892255761555333 |
| 0.3 | 0.139014400755103 | -0.326448325978114 | -0.705657954504924 | -0.705657954504925 | 0.524170422112513 | 0.00662668903002442 | -0.316782268265183 | -0.144960793585703 |
| 0.33 | -0.0500076647234887 | 0.197178441487404 | -0.175984209838066 | -0.148027072411272 | 0.296208541607925 | -0.0392232031813984 | -1.03503153457661 | -0.255642711873867 |
| 0.35 | 0.159962324394074 | -0.398781096971919 | 0.260373543112647 | -0.157238355096007 | -0.824433128173602 | -0.0127183706873837 | 0.682405651166452 | -0.260373543112647 |

Showing 1 to 10 of 302 entries Previous 1 2 3 4 5 ... 31 Next

- *R-Squared*: It contains the r-squared and adjusted r-squared values for each linear regression, represented by the data variables.

P-values Coefficients R-squared

Download CSV Download HTML Report Save CSV Save HTML Report

(The Report includes results from all three tabs)

Show 10 entries Search:

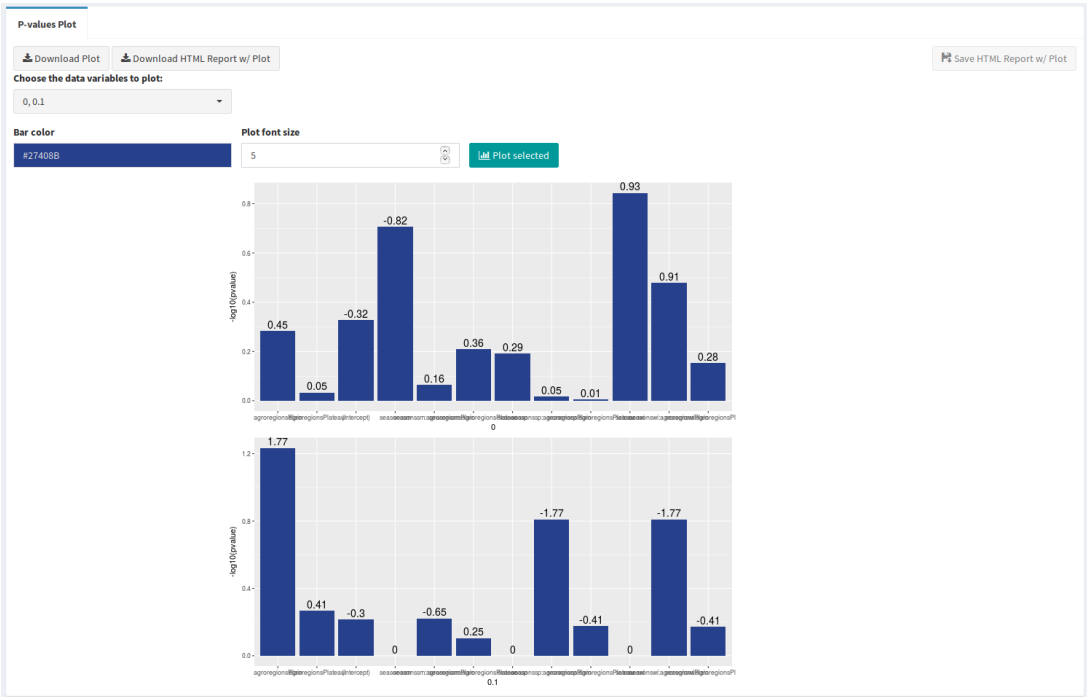
| | r.squared | adj.r.squared |
|---|---|---|
| 0 | 0.528549847924879 | 0.418210450630702 |
| 0.1 | 0.187427030697946 | -0.00274962169189652 |
| 0.12 | 0.176997760600666 | -0.0156197847906681 |
| 0.16 | 0.139855830152921 | -0.061454507470863 |
| 0.18 | 0.218384506668354 | 0.0354532209949903 |
| 0.23 | 0.113384730717933 | -0.0941209706034019 |
| 0.26 | 0.222123843193866 | 0.0400677213881755 |
| 0.3 | 0.152271191961187 | -0.0461334226861945 |
| 0.33 | 0.168381340375756 | -0.0262528140043865 |
| 0.35 | 0.207954606578056 | 0.022582280458026 |

Showing 1 to 10 of 302 entries Previous 1 2 3 4 5 ... 31 Next

As regards to the **P-value plots** results, the user is able to see a visual representation of the results seen in the *P-values* tab of the numerical results with a bar plot of the negative base 10 logaritm of the p-value of a linear regression on a data variable.

Three options are available to personalize the plot:

- Choose the data variables to plot;
- Choose the color of the bars;
- Size of the text in the plot bars;




### Correlation Analysis

For the results of this page, the available **options used** are:

- Analysis Name;
- Name of the dataset used;
- Correlation method used;
- If correlation was performed between samples or variables;
- The alternative hypothesis chosen, in case correlation tests were performed.

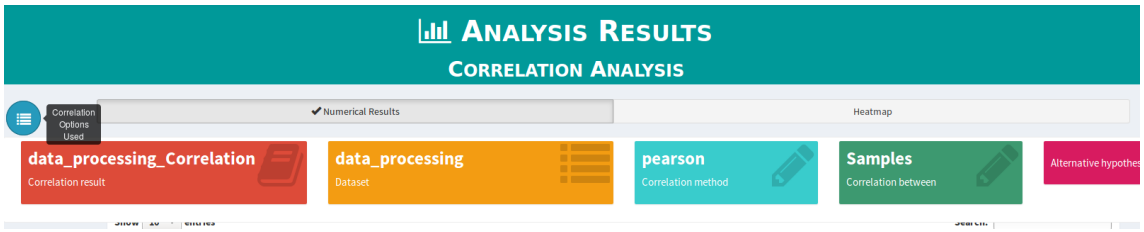
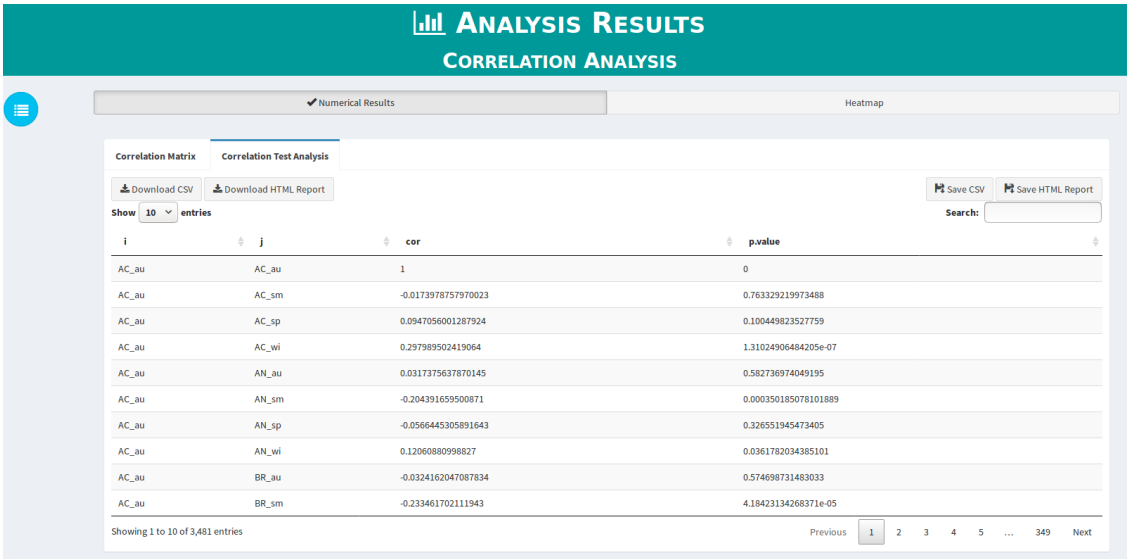


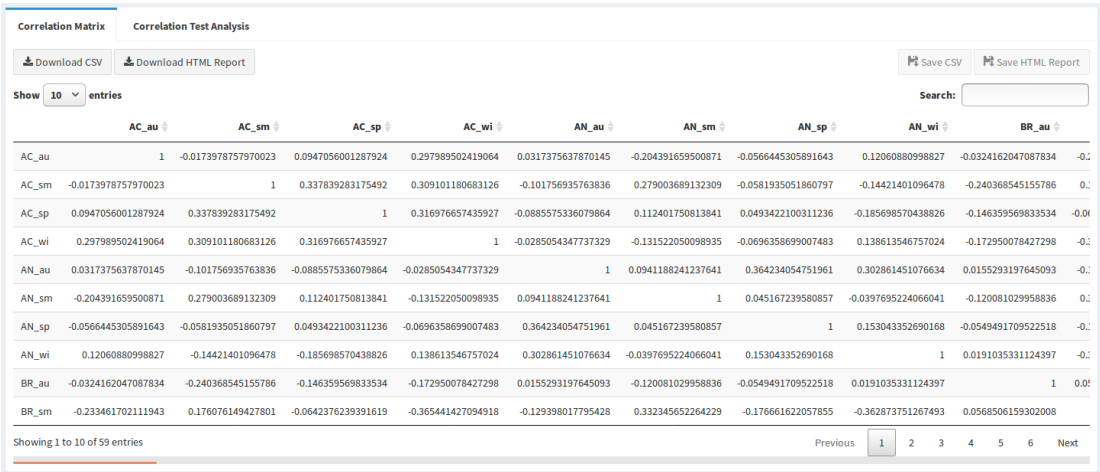

Figure 2.77: Layout of the dropdown menu of the options for Linear Regression Analysis.

At the right of the options button, there are two buttons, which allow the user to see the numerical results and the heatmap of the correlations:

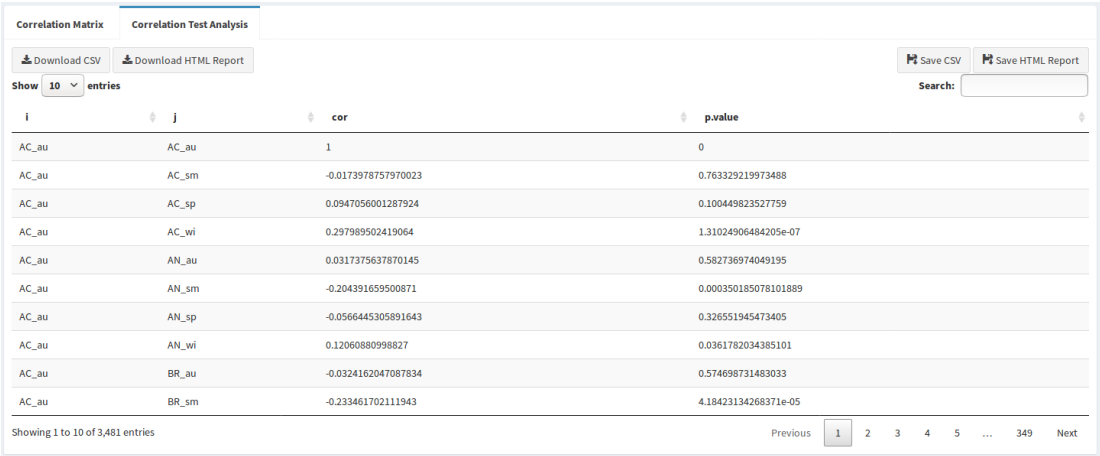



As regards to the **numerical results**, a tabset panel with the following results can be seen:

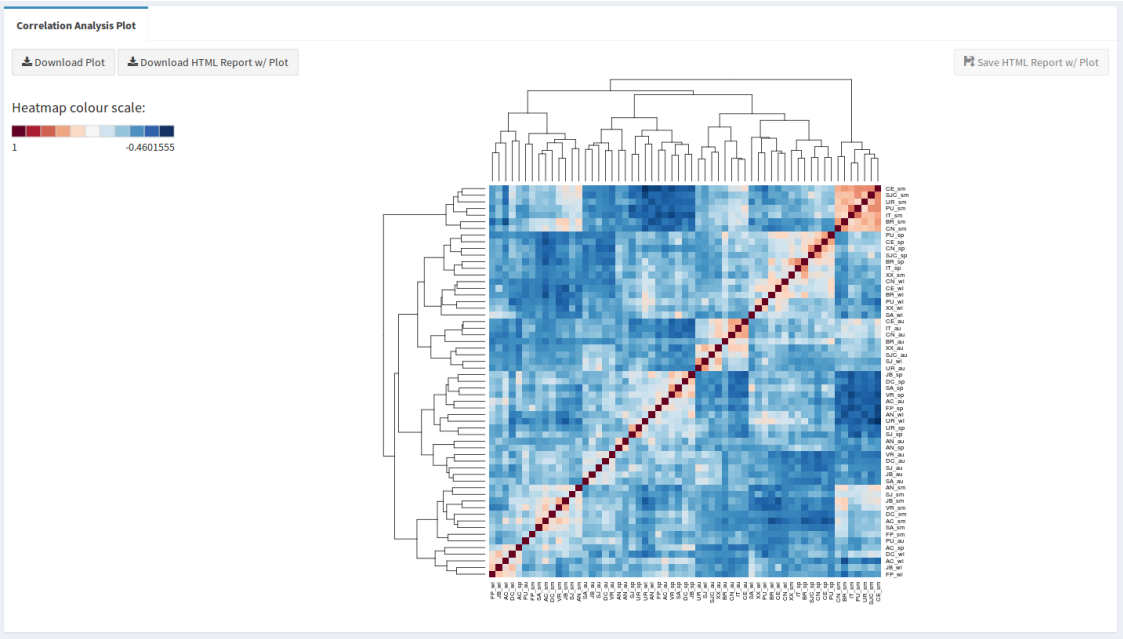
- *Correlation Matrix*: It contains a table with the correlation values between the different samples;




- *Correlation Test Analysis*: It contains a table with the correlation value and p-value between the different samples.



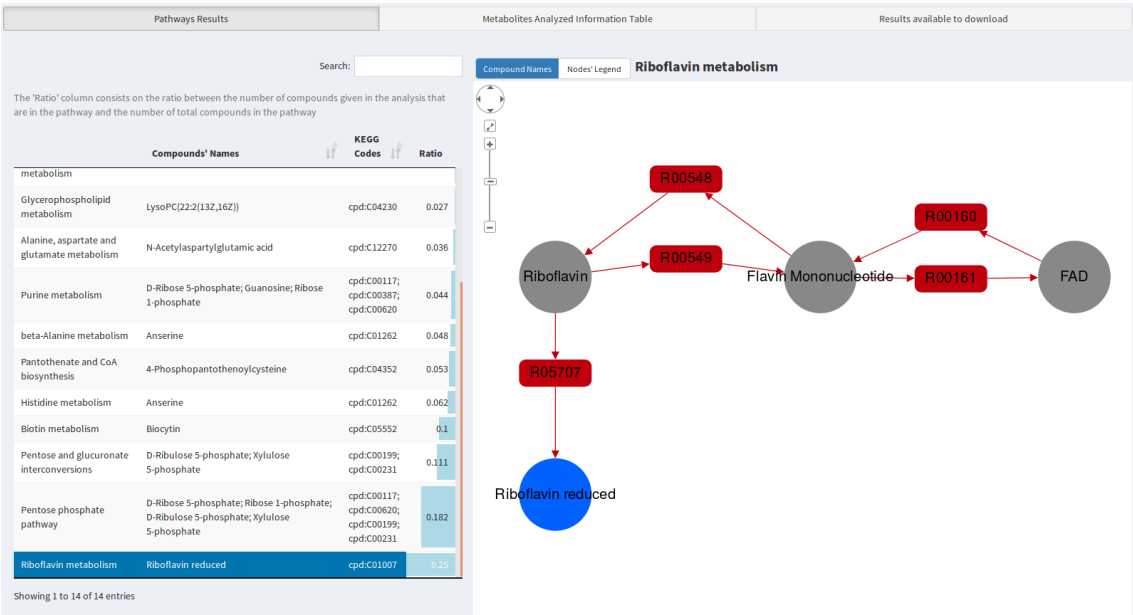
Furthermore, the user is able to see a **Heatmap** plot of the correlations between the samples, as a visual representation of the *Correlation Matrix* in the numerical results:



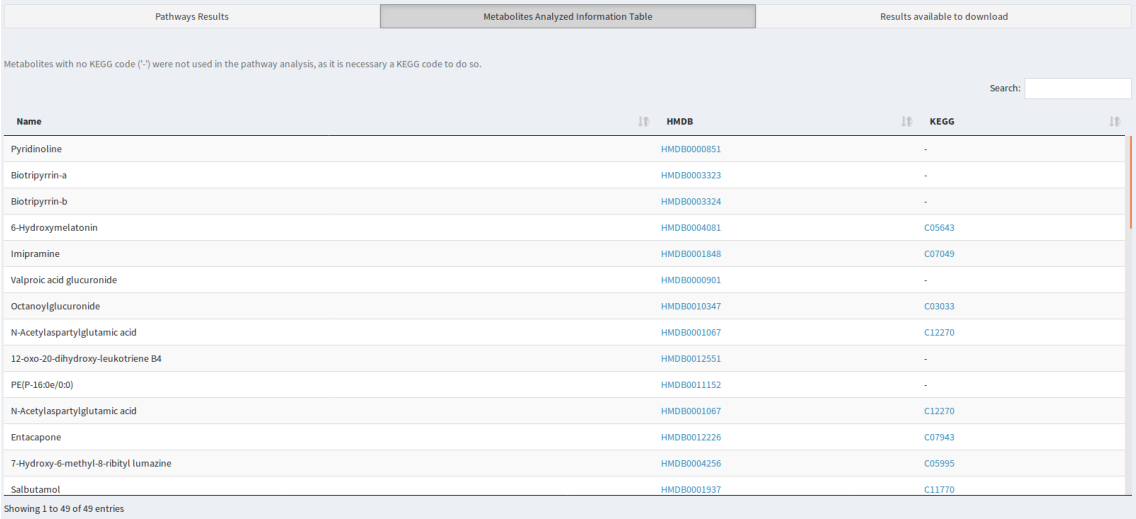

### 2.9.8 Pathway Analysis

There are three sets of information for this type of results. These can be accessed through the buttons positioned at the top of the page:

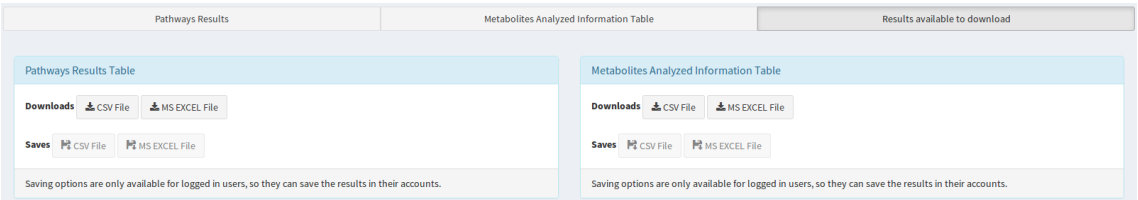
- *Pathways Results*: this page contains a table containing information on the pathways identified, the given compounds that are present in each of those pathways, the respective compounds KEGG codes and the ratio between the number of compounds given in the analysis that are in each pathway and the number of total compounds in the pathway. At the right, you can see the pathway map of the pathway that you click to see in the table.



- *Metabolites Analyzed Information Table*: this page contains a table with the KEGG and HMDB codes for each compound. Each compound is represented by the name. Only the compounds that have a KEGG code associated were taken into consideration for the analysis.




- *Results available to download*: it is in this page where you can download tables, in CSV or MS EXCEL formats, of the tables present in the two pages mentioned above.



# II Use Examples

# 3. NMR Peak Lists: Propolis

## 3.1 Where to find the data

The study here reproduced aimed to get insights of important features associated with the chemical composition, seasons and geographical origin of the propolis produced in the Santa Catarina state, in southern Brazil [1].

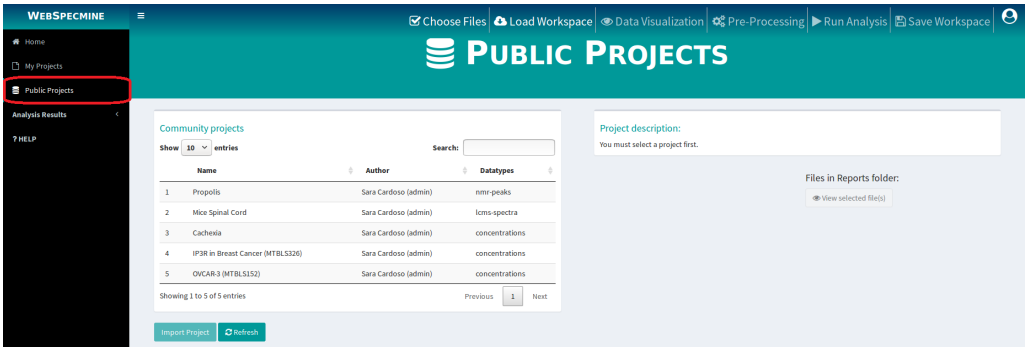
The samples used in this study, acquired using the NMR technique, were stored in the public project called *Propolis*, under the data folder *NMR Peaks Data*. Regarding the metadata, the file *propolis_nmr_metadata.csv* is given.

There are a total of 59 samples, 15 from autumn (AU) and spring (SP), 13 from winter (WI) and 16 from summer (SM). They were collected in 2010 from *Apis mellifera* hives located in southern Brazil (Santa Catarina State). The samples are also separated in three agroecological regions for the different apiaries: 12 samples from the Highlands, 11 from the Plain, and 36 from the Plateau.

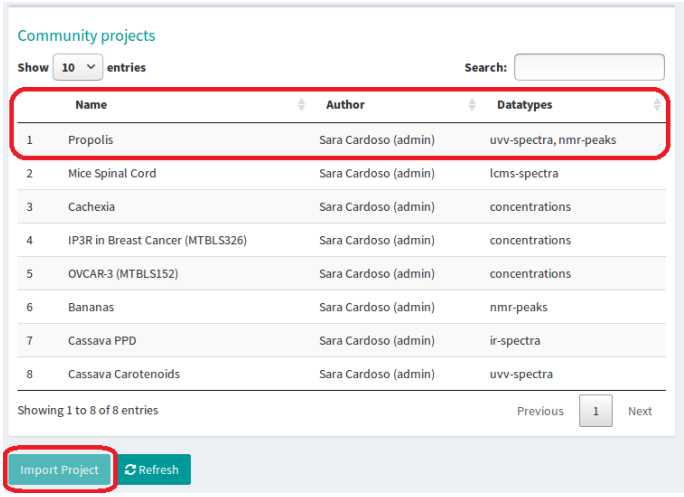
The analysis pipeline here demonstrated followed one available in `http://pubs.acs.org/doi/suppl/10.1021/acs.jnatprod.5b00315/suppl_file/np5b00315_si_001.pdf`.

## 3.2 Choosing the files for analysis

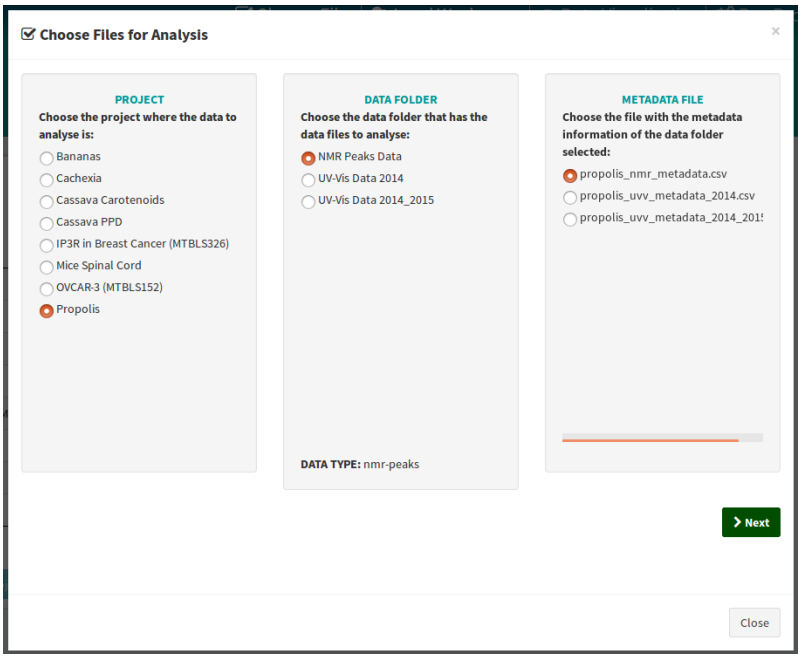
1. Enter your user account;
2. Copy the public project in question, named *Propolis*, into your account:
   Go to the "Public Projects" page, accessible through the sidebar panel;



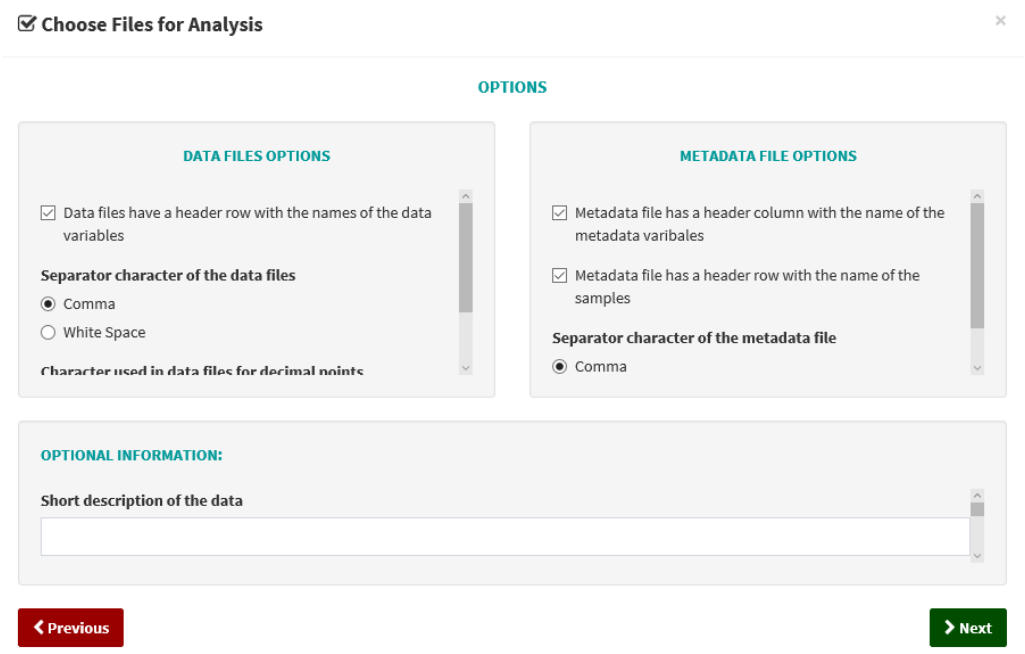

Select the project in the table of the *Community projects* box and click the button "Import Project";



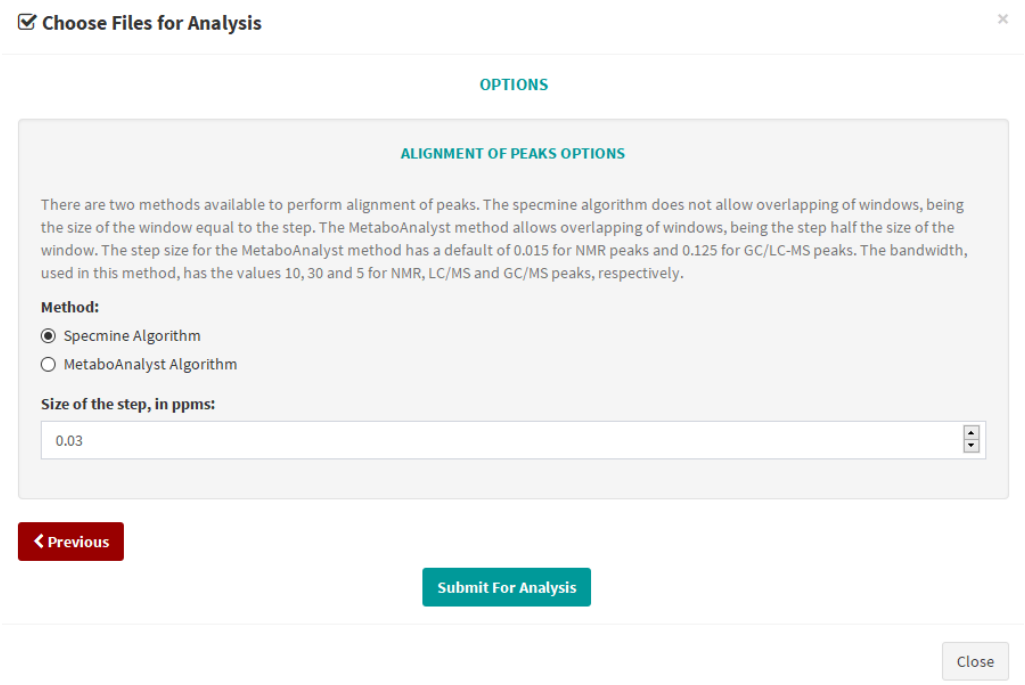

3. Click the "Choose Files" button, present in the header panel;
4. Choose the project, data folder and metadata file in question and click the "> Next" button;



5. This will lead to the window were the options regarding the data and metadata files are set, so that they are read and processed correctly. In this case, the options are the default ones, so no change is needed;

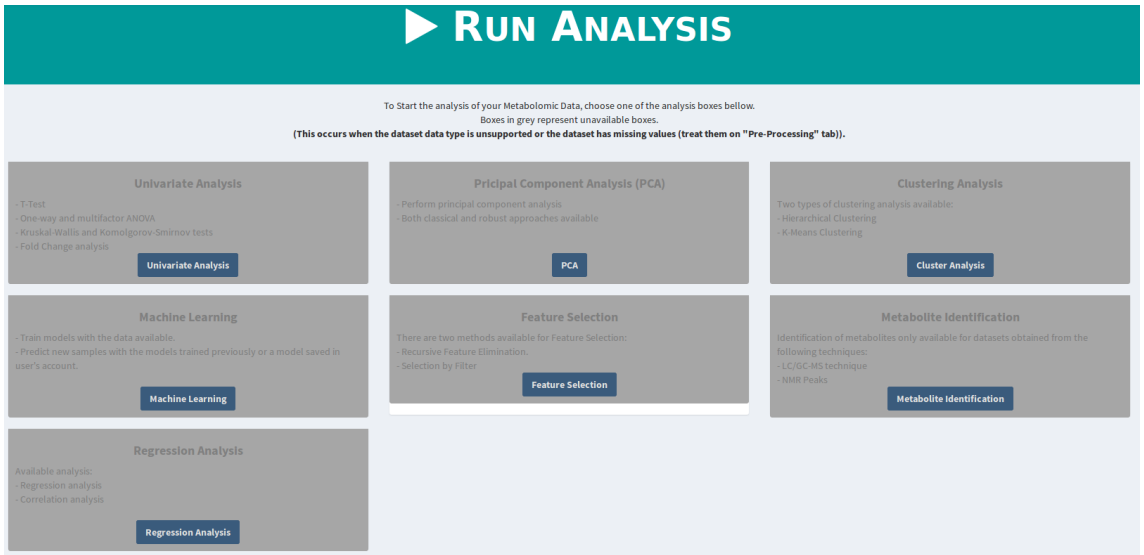



6. After clicking the "> Next" button, you will have to set the options for the alignment of peaks. In this case, the default ones are also used, which consist in the specmine algorithm as the method used and 0.03 ppm as the size of the step. With this, you are able to click the button "Submit For Analysis" to finalize the submission of the data to analyse.



## 3.3 Pre-process the data

After the data files are processed, the user is redirected to the "Run Analysis" page. Here, the user will notice that no box is accessible at this point:

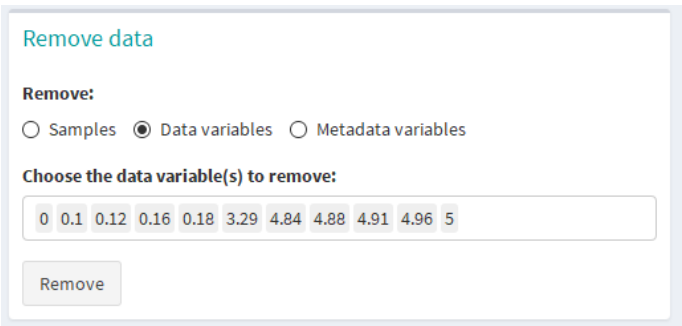



This happens because the dataset created by processing the files (*OriginalData*) has missing values, which makes impossible to proceed with the analysis. Thus, pre-processing of the dataset is needed.

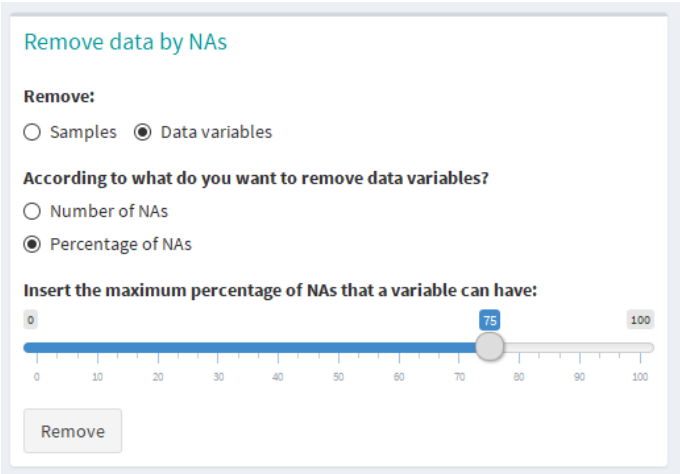
Two different pre-processing pipelines were applied, one to be used in the chemometrics analysis, named *data_chemometrics*, and the other one for metabolite identification, named *data_ID*.

### *Pre-processing the dataset for the chemometrics analysis*

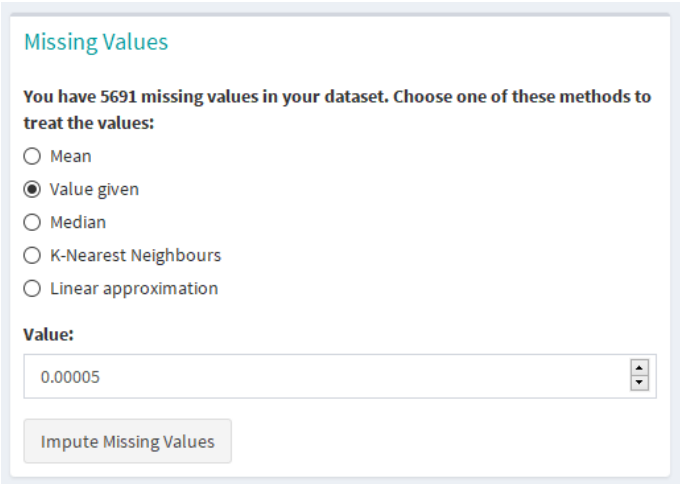
1. Go to the "Pre-Processing" page, accessible through the header panel;
2. Extract the data variables whose ppm values are between 0-0.19, 3.29-3.31 and 4.84-5.00;



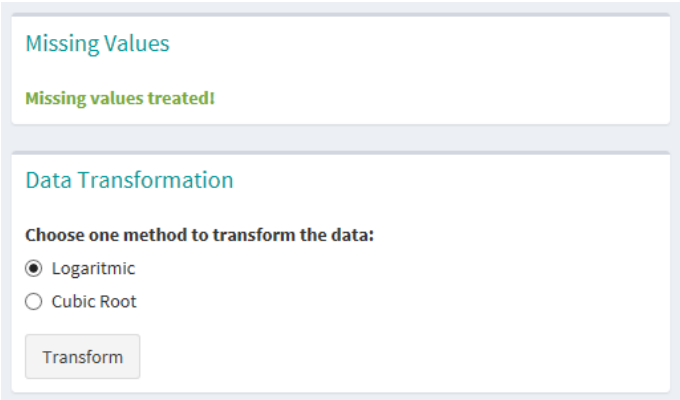

3. Remove any variables with more than 75% of missing values;



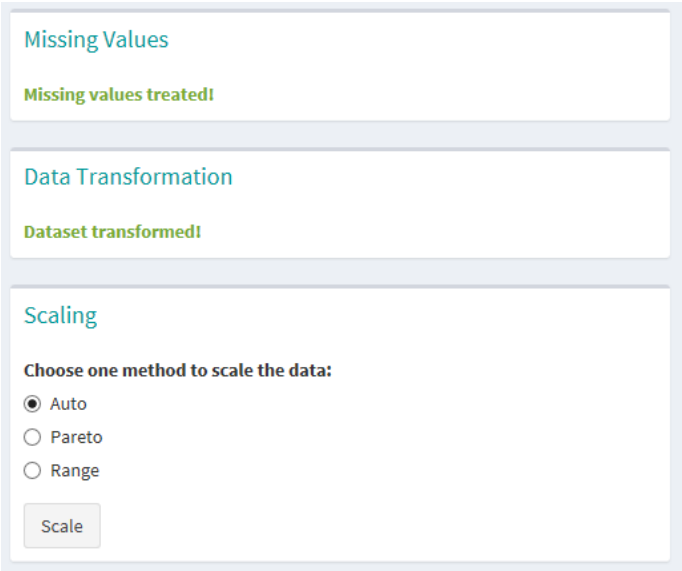

4. Treat missing values, by replacing them with the given value of 0.00005



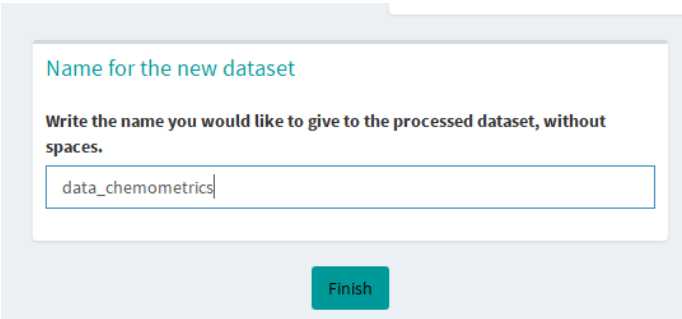

5. Do logaritmic transformation on the dataset;




6. Auto scale the data;



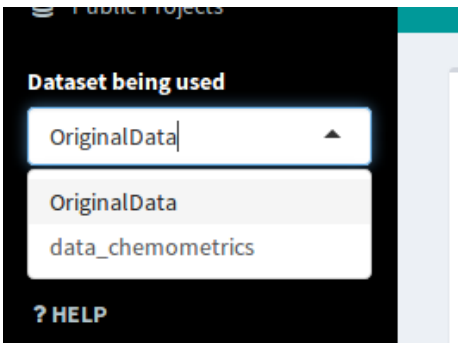

7. Name the dataset (*data_chemometrics*) and click the "Finish" button;



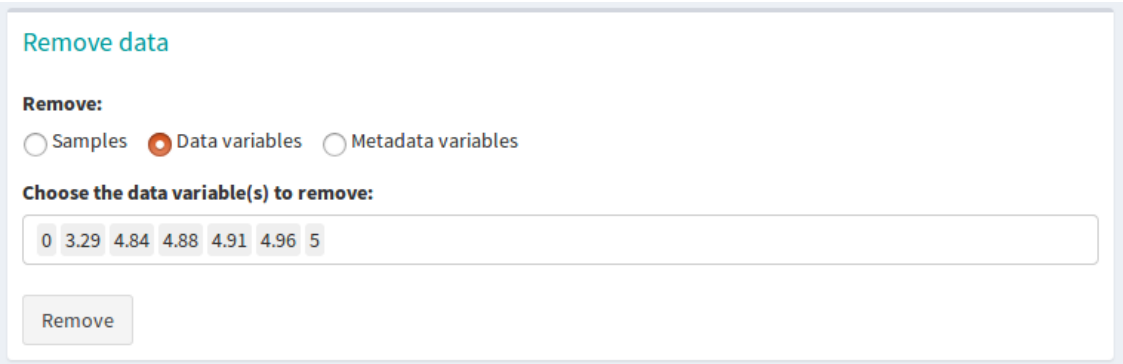

8. With this, the dataset being currently in use will automatically change to the newly created dataset and, by entering the "Run Analysis" page once again, the user will notice that the boxes are now available.

### *Pre-processing the dataset for the metabolite identification*

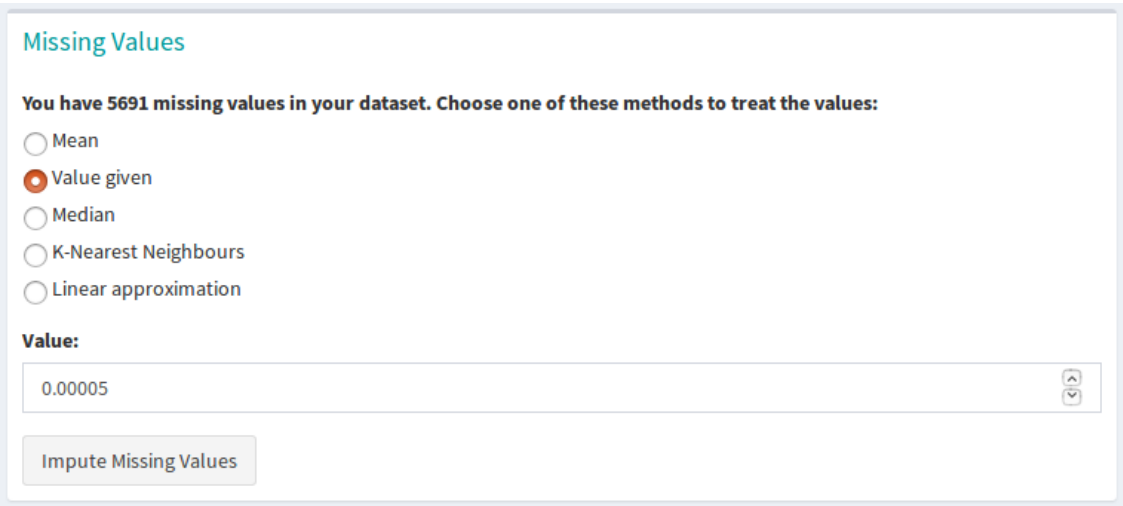
1. Still in the "Pre-Processing" page, you should now set the dataset being used back to *OriginalData* to start the new processing;



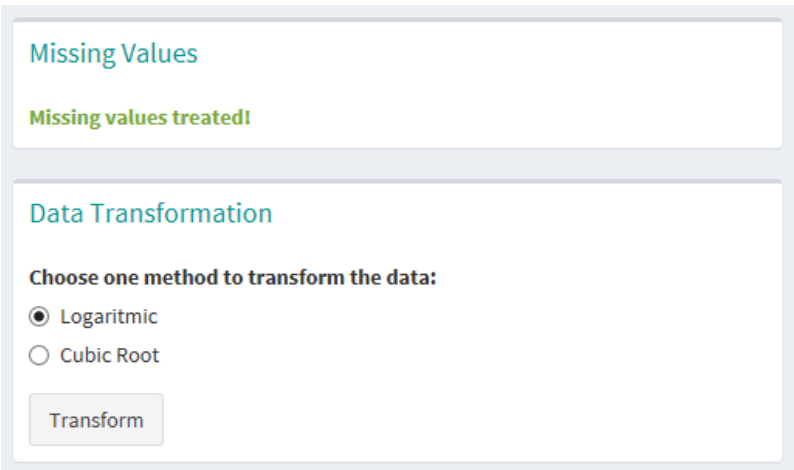

2. Remove the data variables between the ppm values 0, 3.29-3.31 and 4.84-5.00;



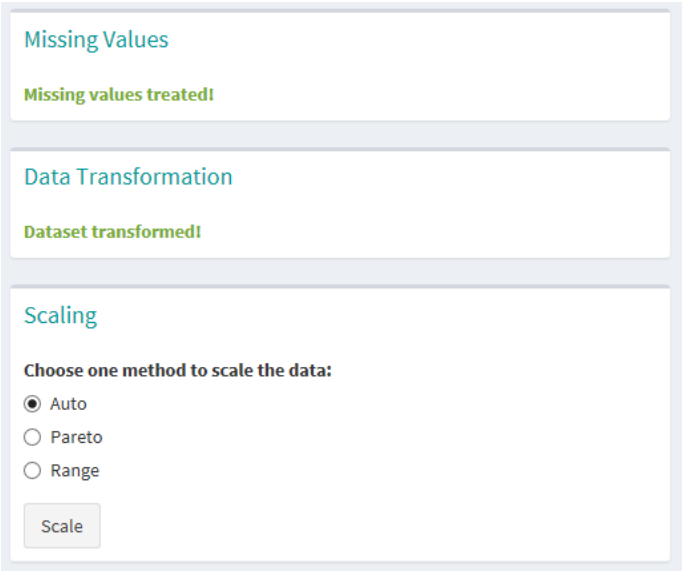

3. Treat the missing values by replacing them with the given value of 0.00005;



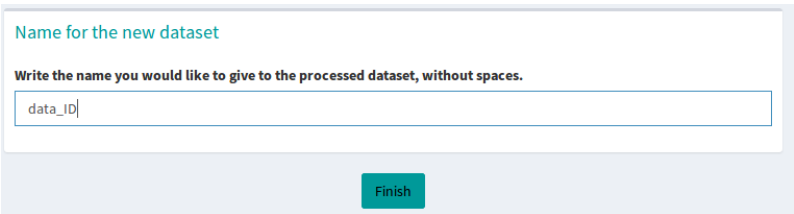

4. Perform logaritmic trasnformation;




5. Auto scale the data;




6. Name the new dataset (*data_ID*) and click the "Finish" button.;

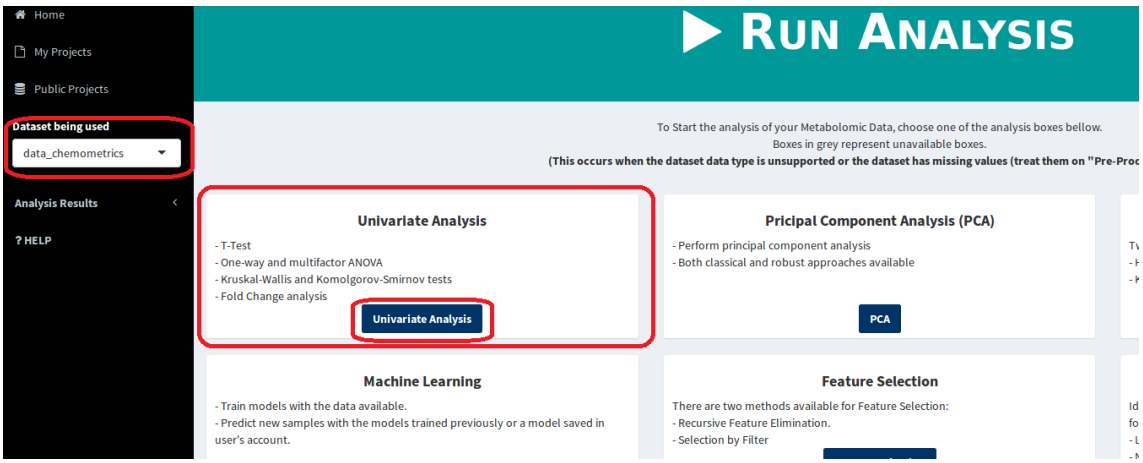



7. With this, the dataset being currently in use will automatically change to the newly created dataset.

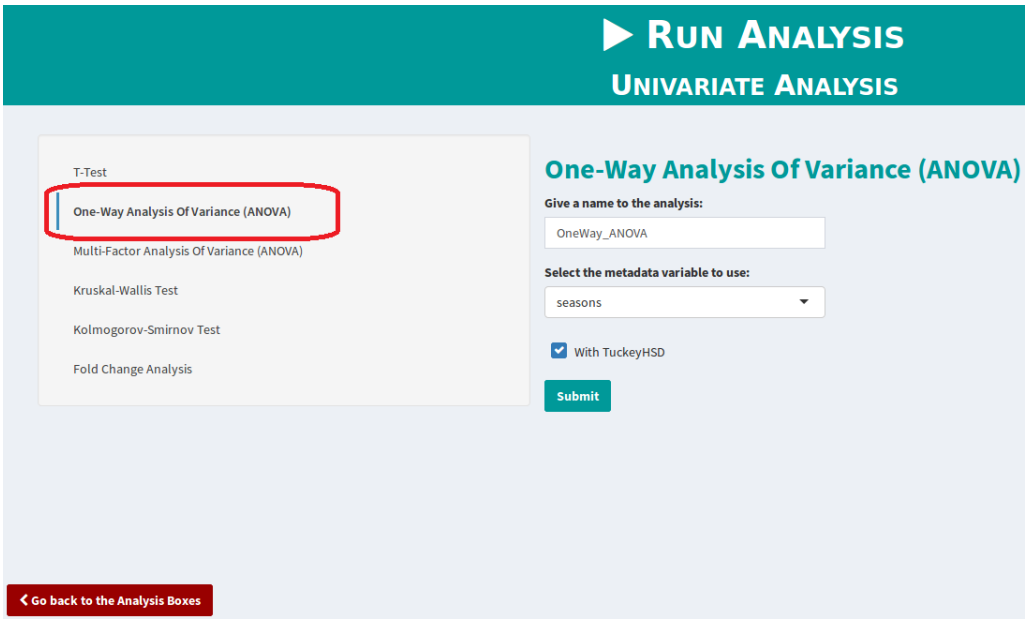
## 3.4 One-way ANOVA Analysis

Here, it is demonstrated how to perform a one-way ANOVA analysis, along with TuckeyHSD test, by using the metadata variable *seasons*.

1. Enter the "Univariate Analysis" box in the "Results Analysis" page while the dataset being used is *data_chemometrics*;



2. Access the "One-Way Analysis of Variance (ANOVA)" tab, in the tab box located at the left of the page. The options regarding this type of analysis will appear at the right;
3. Set the options regarding the analysis and click "Submit" button;

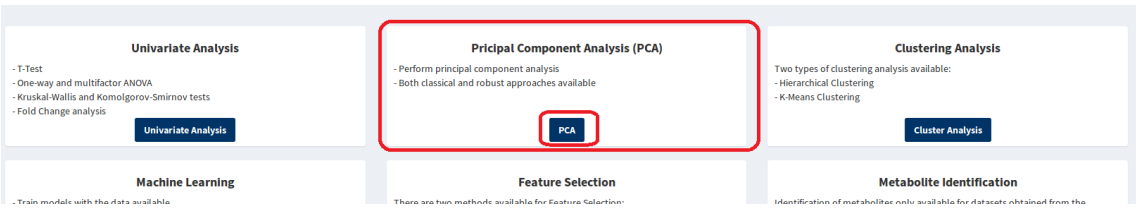



4. Once this analysis is finished, the website redirects the user to the corresponding results page. *For better understanding what information the results contain, go to subsection One-Way ANOVA in section 2.9.1 .*

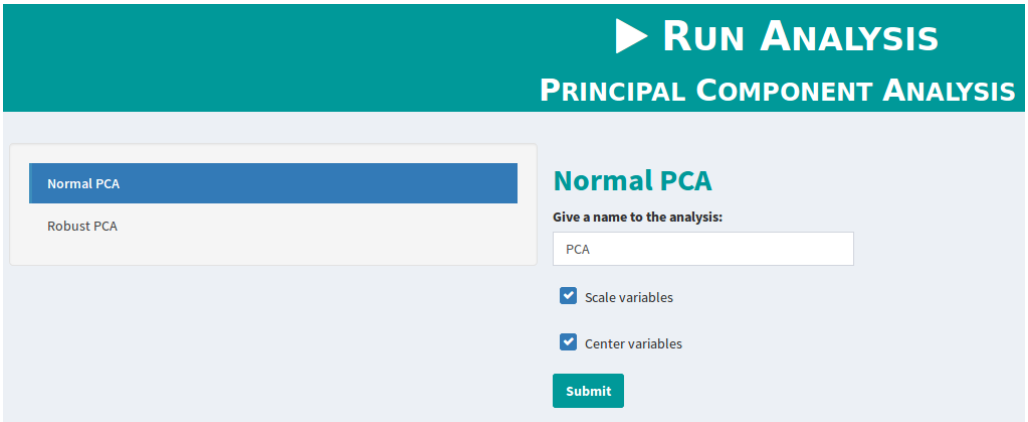
## 3.5 Principal Components Analysis

To perform PCA analysis, you have to:

1. Go back to the "Run Analysis" page, through the header panel, and select the "Principal Component Analysis (PCA)" box;




2. Select the "Normal PCA" tab, in the tab box at the left of the page. The options regarding this type of analysis will appear at the right;
3. Set the options to perform the analysis and click "Submit" button;

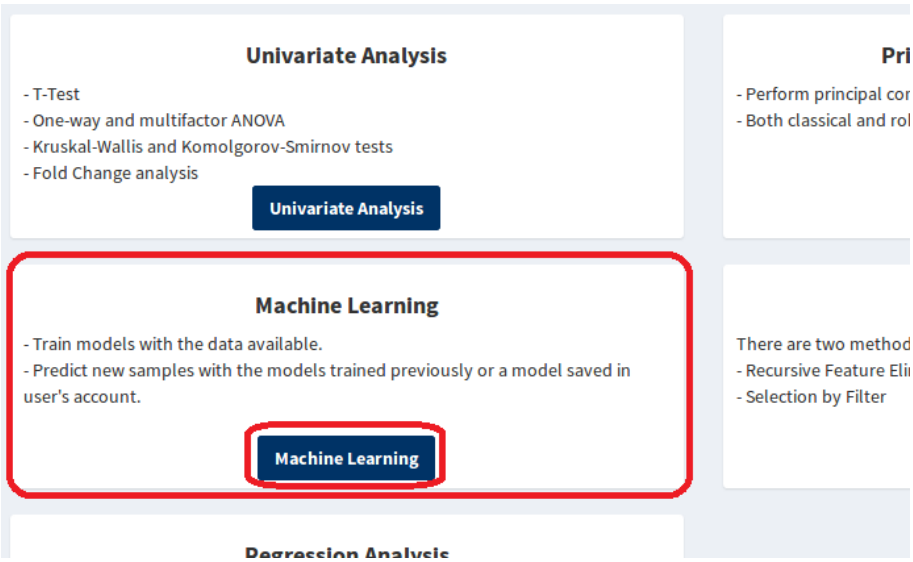



4. Once this analysis is finished, the website redirects the user to the corresponding results page. *For better understanding what information the results contain, go to subsection PCA in section 2.9.1 .*

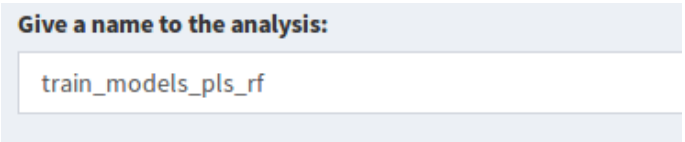
## 3.6 Machine Learning

Finally, to build models to discriminate samples by seasons, do the following:

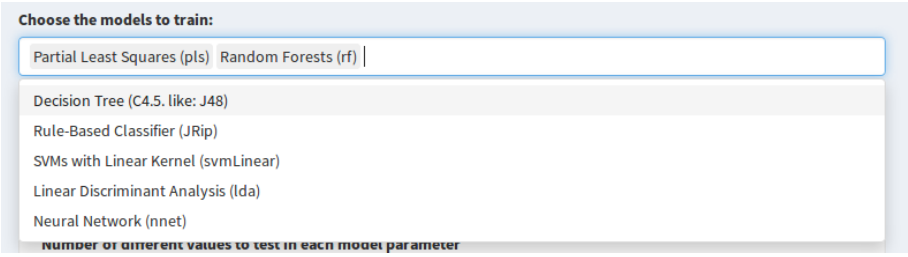
1. Go back to the "Run Analysis" page, through the header panel, and select the "Machine Learning" box;



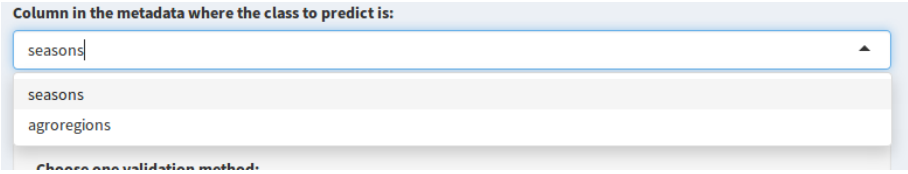

2. Start by giving the name for the analysis;



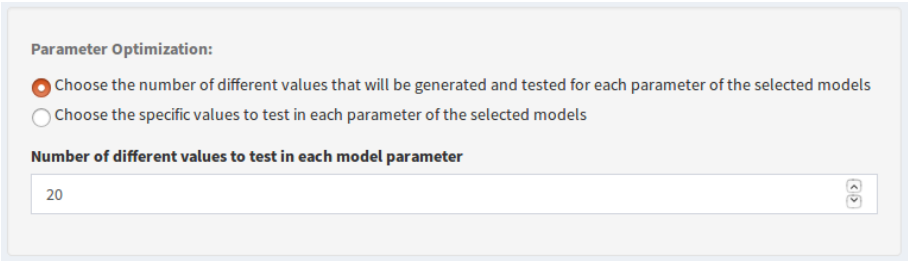

3. To train the two models used in the study, select the PLS and random forests models in the following input option:



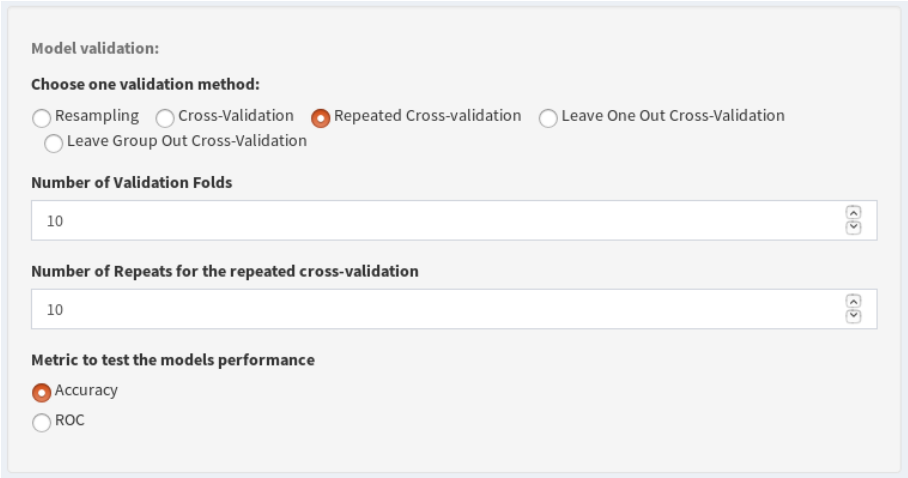

4. Select the metadata variable *seasons* as the one to predict;




5. For parameter optimization, choose 20 different values to test each parameter of the selected models;




6. For model validation, set the following options:

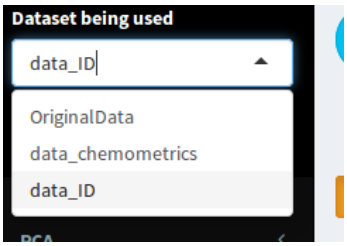



7. Click "Train models" button;
8. Once this analysis is finished, the website redirects you to the corresponding results page. *For better understanding what information the results contain, go to subsection Model Training in section 2.9.4.*

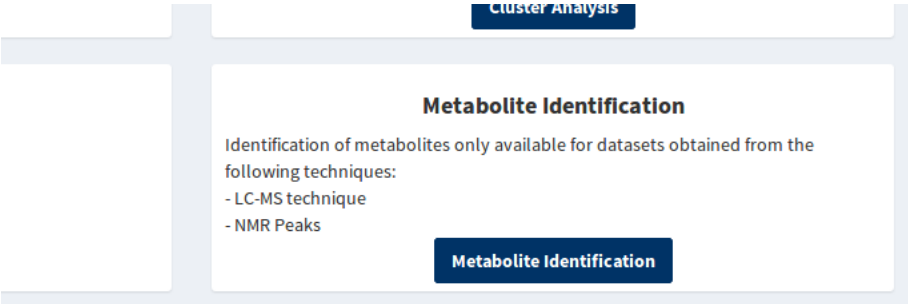
## 3.7 Metabolite Identification

To perform the identification of metabolites present in the samples, do the following:

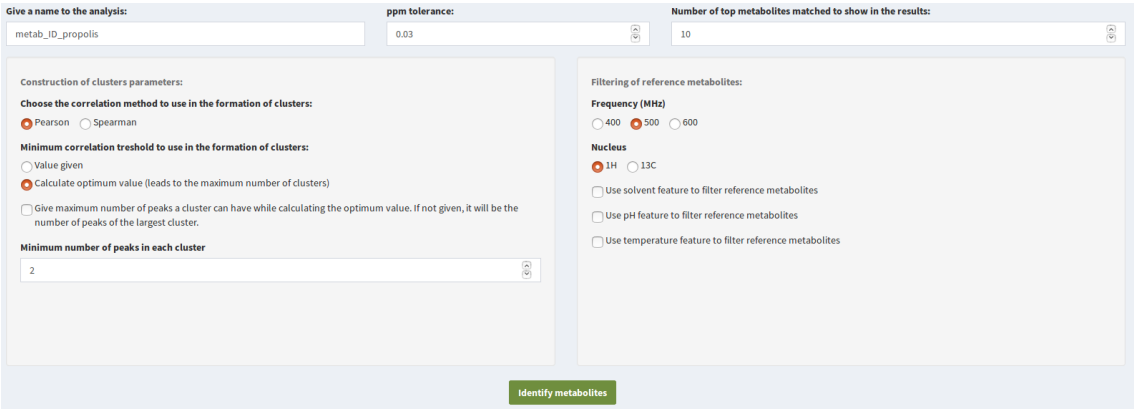
1. Change the selected "Dataset being used", in the sidebar panel, to the dataset *data_ID*;




2. Now go to the "Run Analysis" page, through the header panel, and enter the box "Metabolite Identification";




3. Set the following options to perform this analysis and click the button "Identify metabolites" to perform the identification:




4. Once this analysis is finished, the website redirects you to the corresponding results page. *For better understanding what information the results contain, go to subsection NMR Peaks in section 2.9.6.*

# 4. MS Spectra: Mice Spinal Cord

## 4.1 Where to find the data

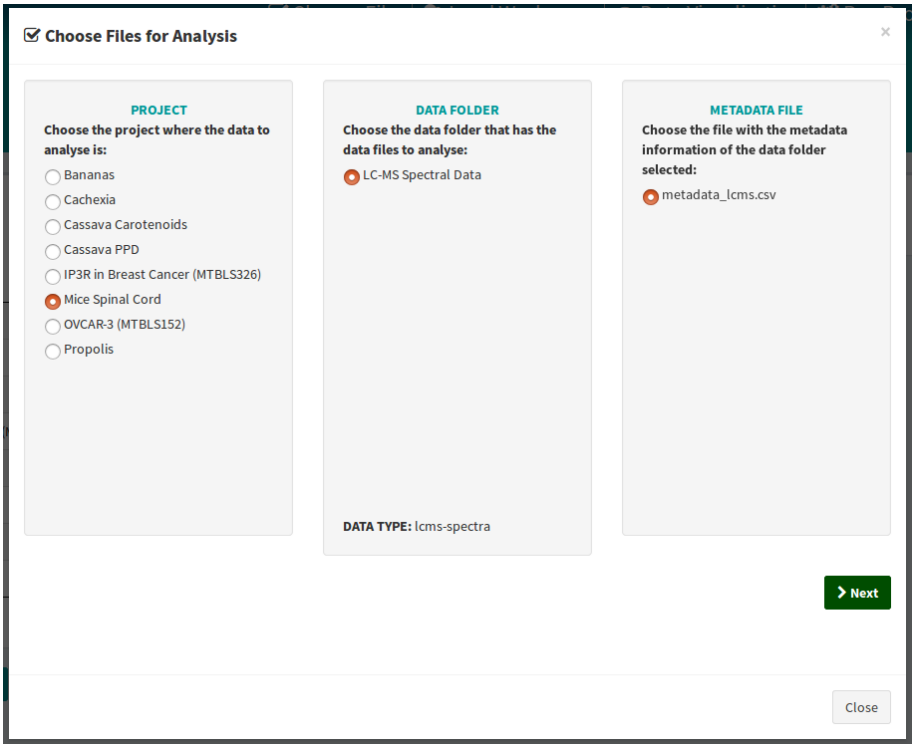
The study here reproduced aimed to identify the endogenous substrates of the FAAH enzyme [2].

The samples used in this study, acquired using the LC-MS technique, were stored in the public project called *Mice Spinal Cord*, under the data folder *LC-MS Spectral Data*. Regarding the metadata, the file *metadata_lcms.csv* is given.

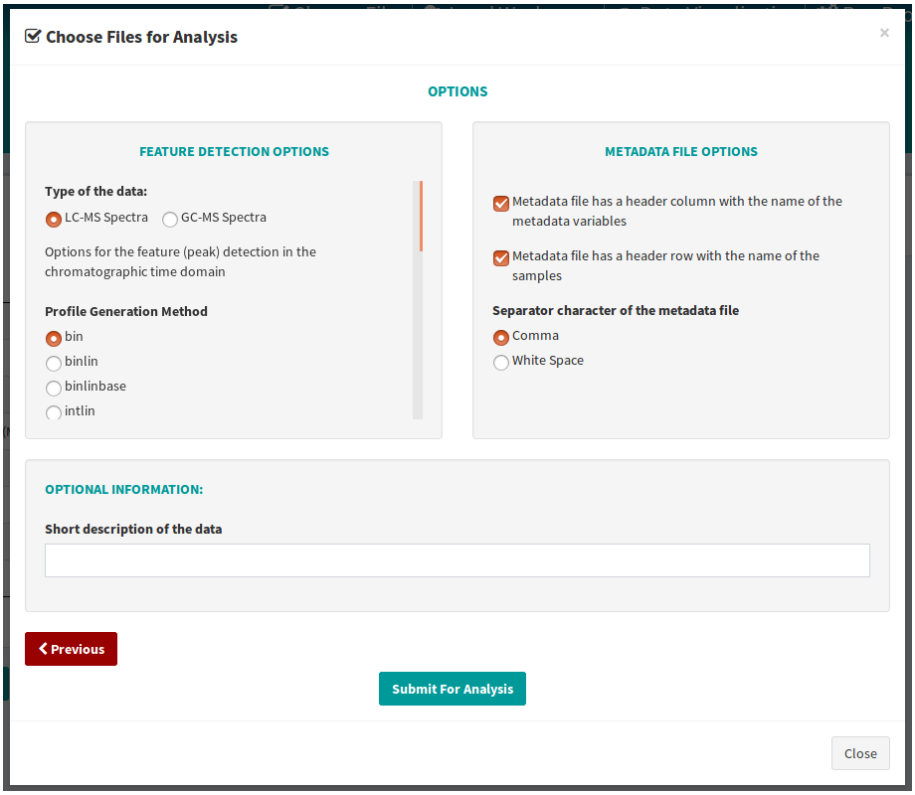
There are a total of 12 samples, in CDF format, 6 from wild-type strains (wt) and 6 not (ko).

## 4.2 Choosing the files for analysis

1. Enter your user account;
2. Copy the public project in question, named *Mice Spinal Cord*, into your account:
   Go to the "Public Projects" page, accessible through the sidebar panel;
   Select the project in the table of the *Community projects* box and click the button "Import Project";
3. Click the "Choose Files" button, present in the header panel;
4. Choose the project, data folder and metadata file in question and click the "> Next" button;




5. This will lead to the window were the options regarding the data and metadata files are set, so that they are read and processed correctly. In this case, the options are the default ones, so no change is needed;




6. With this, you are able to click the button "Submit For Analysis" to finalize the submission

of the data to analyse.

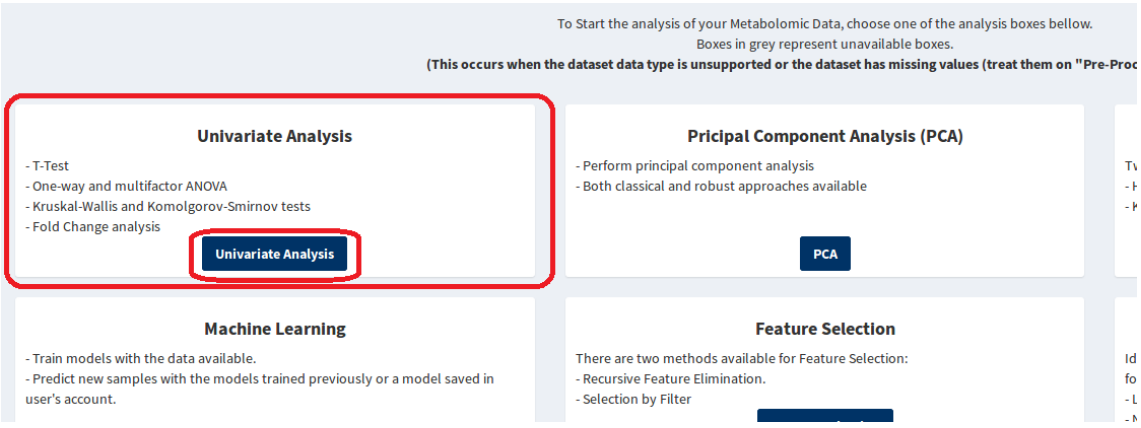
## 4.3 Pre-Process the data

After the data files are processed, the user is redirected to the "Run Analysis" page. Here, the user will notice that, in this case, all boxes are accessible. Following this, no pre-processing is applied, as no processing was conducted by the present study and no missing values were encountered.

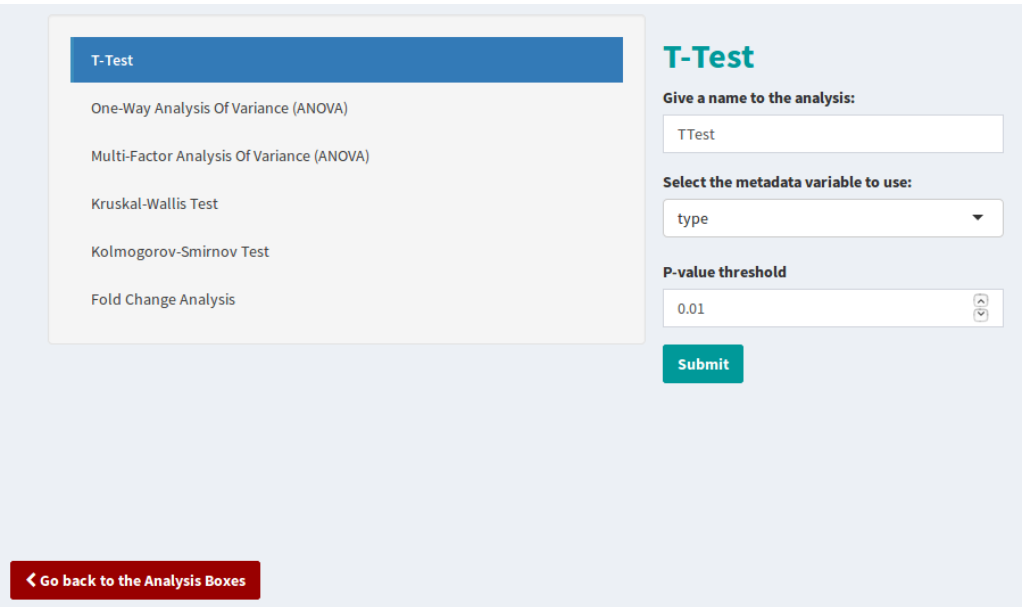
## 4.4 T-Test

To perform T-Test analysis, you have to:

1. Enter the "Univariate Analysis" box in the "Results Analysis" page;




2. Select the "T-Test" tab, in the tab box at the left of the page. The options regarding this type of analysis will appear at the right;
3. Set the options to perform the analysis and click "Submit" button;



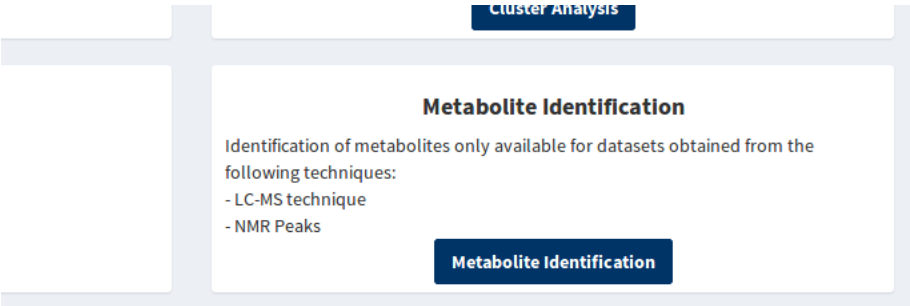

4. Once this analysis is finished, the website redirects the user to the corresponding results page. *For better understanding what information the results contain, go to subsection T-Test in*

*section 2.9.1* .

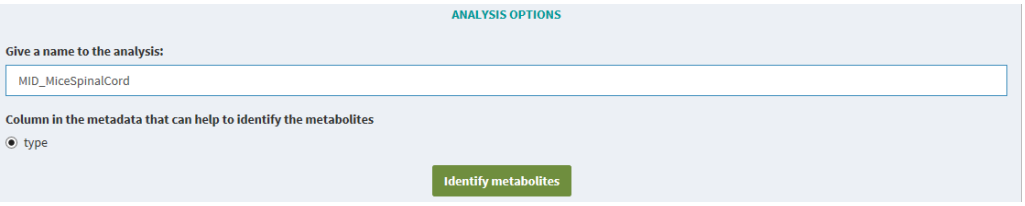
## 4.5 Metabolite Identification

To perform metabolite identification, start by:

1. Going back to the "Run Analysis" page and enter the "Metabolite Identification" box, through the header panel;




2. Set the options to perform the analysis and click "Identify metabolites" button;




3. After the identification is concluded, the website redirects the user to the corresponding results page. *For better understanding what information the results contain, go to subsection LC-MS Data in section 2.9.6.*

# 5. UV-Vis Spectra: Propolis

## 5.1 Where to find the data

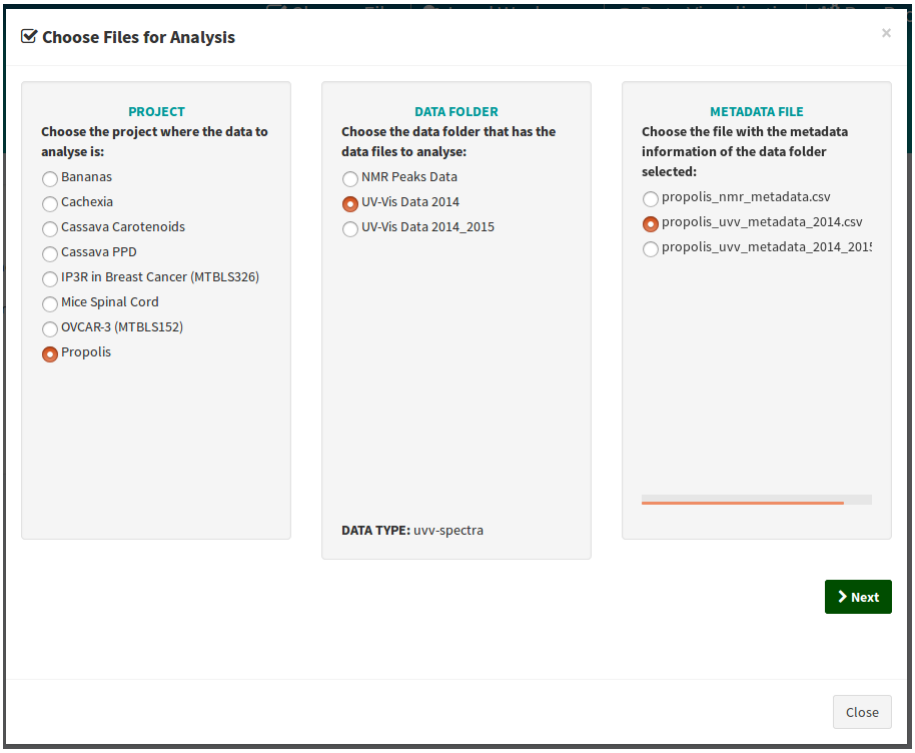
The main scope of the study here reproduced was to determine the harvest season effect on the chemical profile of the propolis in the Santa Catarina state, southern Brazil, throughout the year 2014 [3].

The samples used in this study, acquired using the UV-Vis spectroscopy with a spectral window from 280 to 800 $\eta$m, were stored in the public project *Propolis*, under the data folder *UV-Vis data 2014*. Regarding the metadata, the file *propolis_uvv_metadata_2014.csv* is given.

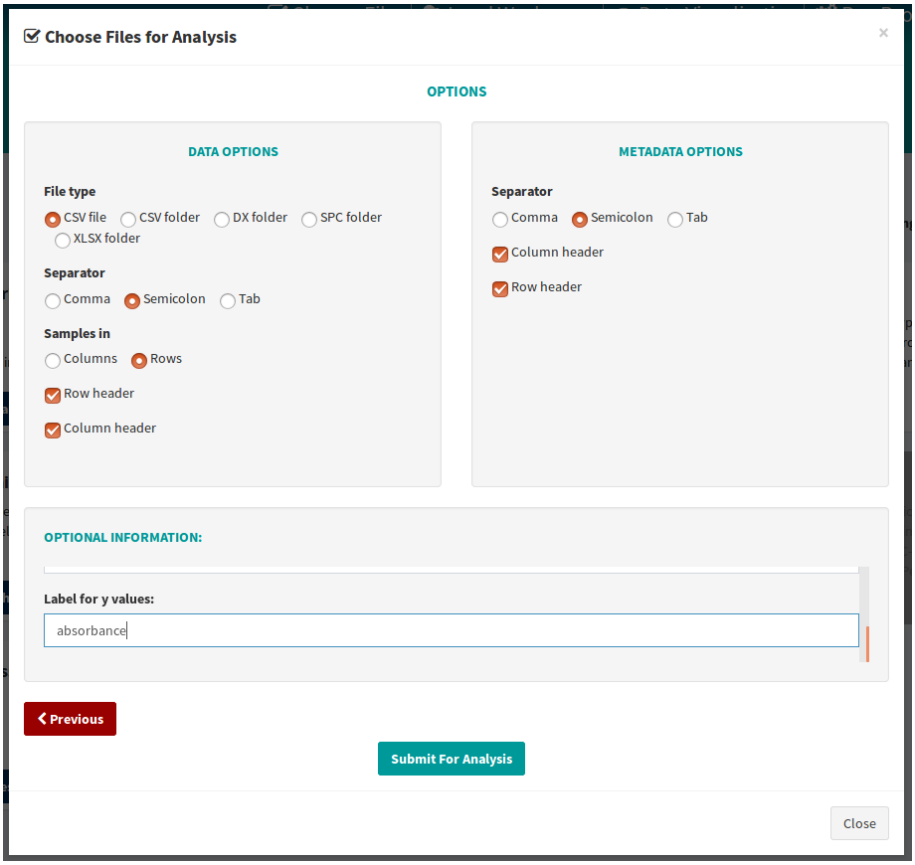
There are a total of 165 samples, whose collected data is all present in one CSV file. Three spectra were collected for each "original" sample, hence having 55 "original" samples, with each one having 3 replicates. The "original" sample to which each sample corresponds to is specified in the metadata variable "names" and the replicates numbers in "replicates". The samples can be further distinguished according to the seasons from when they were collected and color of the sample.

## 5.2 Choosing the files for analysis

1. Enter your user account;
2. Copy the public project in question, named *Propolis*, into your account, if not already, as it is the same project for the *NMR Peak Lists* example in chapter 3:
   Go to the "Public Projects" page, accessible through the sidebar panel;
   Select the project in the table of the *Community projects* box and click the button "Import Project";
3. Click the "Choose Files" button, present in the header panel;
4. Choose the project, data folder and metadata file in question and click the "> Next" button;




5. This will lead to the window were the options regarding the data and metadata files are set, so that they are read and processed correctly. The options to set are the following:

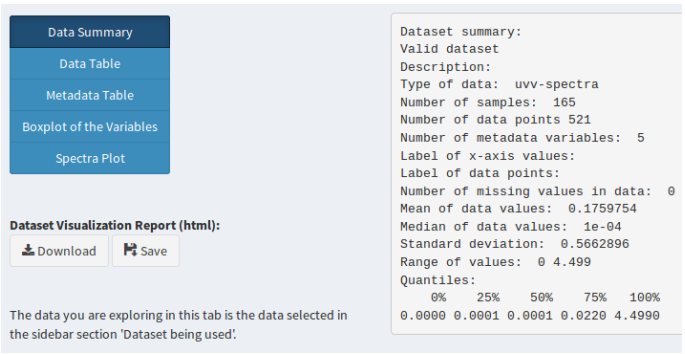


6. With this, you are able to click the button "Submit For Analysis" to finalize the submission of the data to analyse.

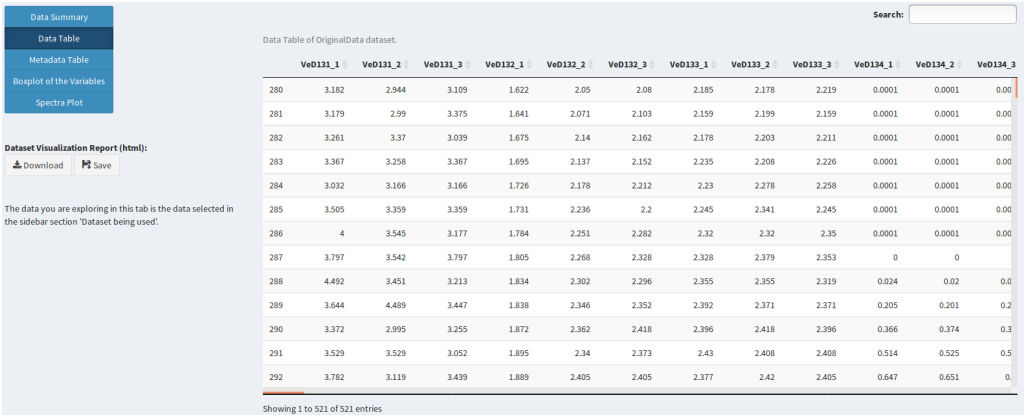
## 5.3 Data Visualization

To have an idea of what is the data being worked on, you can perform the following:

1. Go to "Data Visualization" page, through the header panel;
2. To see a summary of the data, click in the tab "Data Summary" of the tabset panel at the left of the page;

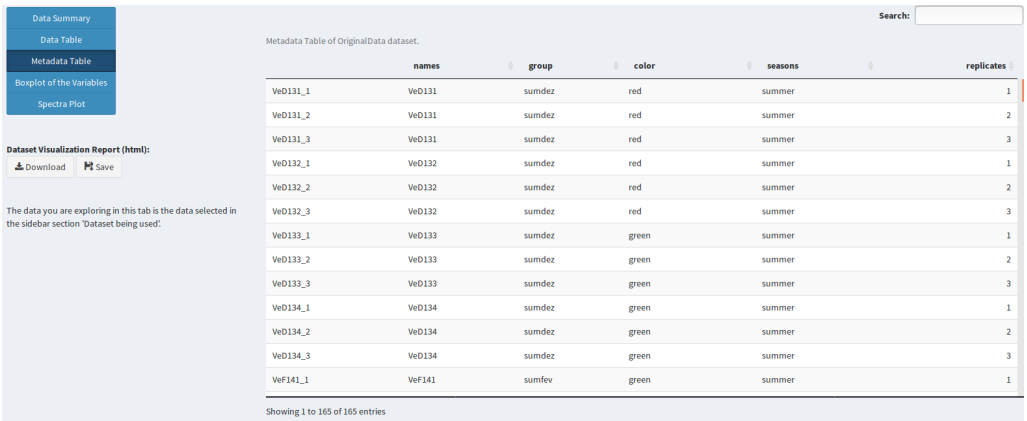



3. You can also see a table with the data values (tab "Data Table");




| | VeD131_1 | VeD131_2 | VeD131_3 | VeD132_1 | VeD132_2 | VeD132_3 | VeD133_1 | VeD133_2 | VeD133_3 | VeD134_1 | VeD134_2 | VeD134_3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 280 | 3.182 | 2.944 | 3.109 | 1.622 | 2.05 | 2.08 | 2.185 | 2.178 | 2.219 | 0.0001 | 0.0001 | 0.00 |
| 281 | 3.179 | 2.99 | 3.375 | 1.641 | 2.071 | 2.103 | 2.159 | 2.199 | 2.159 | 0.0001 | 0.0001 | 0.00 |
| 282 | 3.261 | 3.37 | 3.039 | 1.675 | 2.14 | 2.162 | 2.178 | 2.203 | 2.211 | 0.0001 | 0.0001 | 0.00 |
| 283 | 3.367 | 3.258 | 3.367 | 1.695 | 2.137 | 2.152 | 2.235 | 2.208 | 2.226 | 0.0001 | 0.0001 | 0.00 |
| 284 | 3.032 | 3.166 | 3.166 | 1.726 | 2.178 | 2.212 | 2.23 | 2.278 | 2.258 | 0.0001 | 0.0001 | 0.00 |
| 285 | 3.505 | 3.359 | 3.359 | 1.731 | 2.236 | 2.2 | 2.245 | 2.341 | 2.245 | 0.0001 | 0.0001 | 0.00 |
| 286 | 4 | 3.545 | 3.177 | 1.784 | 2.251 | 2.282 | 2.32 | 2.32 | 2.35 | 0.0001 | 0.0001 | 0.00 |
| 287 | 3.797 | 3.542 | 3.797 | 1.805 | 2.268 | 2.328 | 2.328 | 2.379 | 2.353 | 0 | 0 | |
| 288 | 4.492 | 3.451 | 3.213 | 1.834 | 2.302 | 2.296 | 2.355 | 2.355 | 2.319 | 0.024 | 0.02 | 0.0 |
| 289 | 3.644 | 4.489 | 3.447 | 1.838 | 2.346 | 2.352 | 2.392 | 2.371 | 2.371 | 0.205 | 0.201 | 0.2 |
| 290 | 3.372 | 2.995 | 3.255 | 1.872 | 2.362 | 2.418 | 2.396 | 2.418 | 2.396 | 0.366 | 0.374 | 0.3 |
| 291 | 3.529 | 3.529 | 3.052 | 1.895 | 2.34 | 2.373 | 2.43 | 2.408 | 2.408 | 0.514 | 0.525 | 0.5 |
| 292 | 3.782 | 3.119 | 3.439 | 1.889 | 2.405 | 2.405 | 2.377 | 2.42 | 2.405 | 0.647 | 0.651 | 0. |

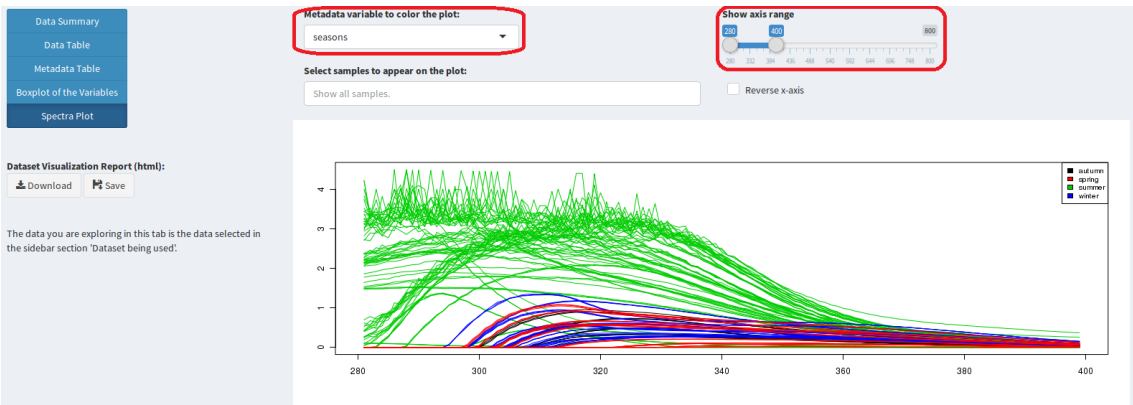
4. And a table with the metadata values (tab "Metadata Table");




| | names | group | color | seasons | replicates |
|---|---|---|---|---|---|
| VeD131_1 | VeD131 | sumdez | red | summer | 1 |
| VeD131_2 | VeD131 | sumdez | red | summer | 2 |
| VeD131_3 | VeD131 | sumdez | red | summer | 3 |
| VeD132_1 | VeD132 | sumdez | red | summer | 1 |
| VeD132_2 | VeD132 | sumdez | red | summer | 2 |
| VeD132_3 | VeD132 | sumdez | red | summer | 3 |
| VeD133_1 | VeD133 | sumdez | green | summer | 1 |
| VeD133_2 | VeD133 | sumdez | green | summer | 2 |
| VeD133_3 | VeD133 | sumdez | green | summer | 3 |
| VeD134_1 | VeD134 | sumdez | green | summer | 1 |
| VeD134_2 | VeD134 | sumdez | green | summer | 2 |
| VeD134_3 | VeD134 | sumdez | green | summer | 3 |
| VeF141_1 | VeF141 | sumfev | green | summer | 1 |

5. You can also view boxplots on one or more variables, by choosing the ones wanted at the top of the page (tab "Boxplot of the Variables");

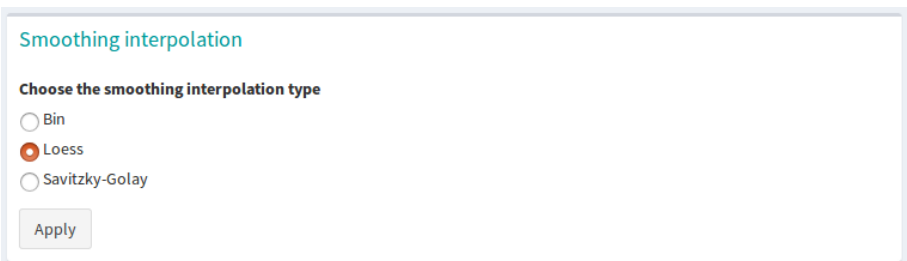

6. Finally, as it is spectral data, you can also see a spectra plot (tab "Spectra Plot"). Here, the plot was personalized so that the spectra are colored by the seasons and it is only shown the values between 280 and 400 $\eta$m.



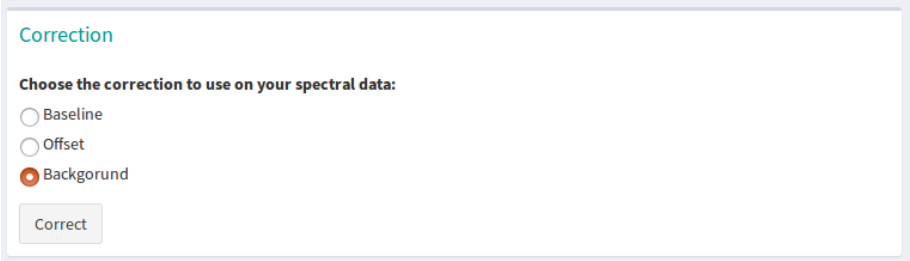

## 5.4 Pre-Process the data

A pre-processing pipeline with four steps is here applied.

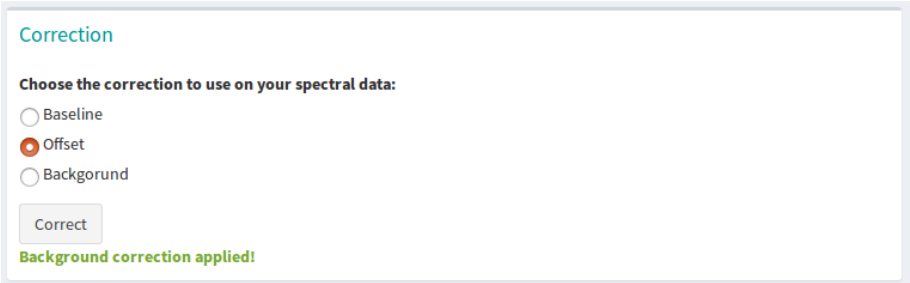
1. Go to the "Pre-Processing" page, accessible through the header panel;
2. Perform smooth interpolation, by selecting the method "Loess";



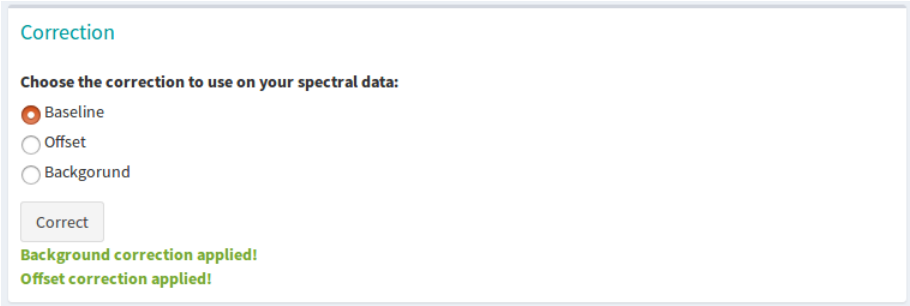
3. Perform background correction;



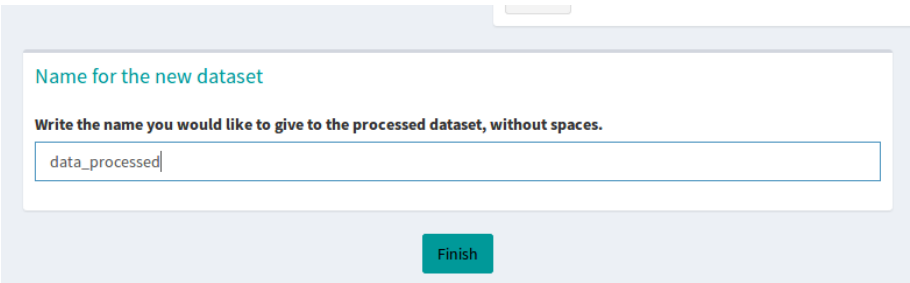

4. Perform offset correction;




5. Perform baseline correction.




6. Name the dataset (*data_processed*) and click the "Finish" button;

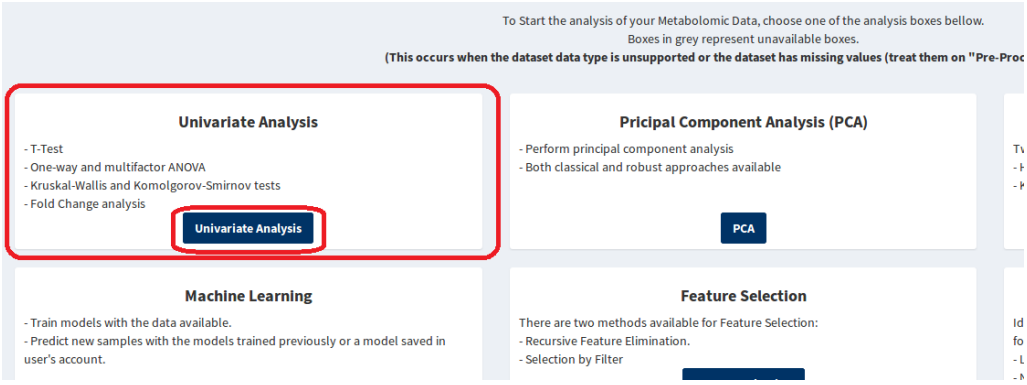



7. With this, the dataset being currently in use will automatically change to the newly created dataset.

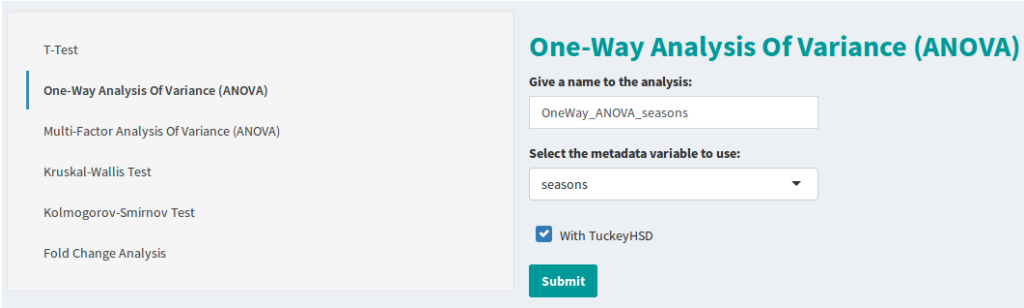
## 5.5 one-way ANOVA Analysis

Here, it is demonstrated how to perform a one-way ANOVA analysis, along with TuckeyHSD test, by using the metadata variable *seasons*, as it has more than two possible values.

1. Enter the "Univariate Analysis" box in the "Results Analysis" page while the dataset being used is *data_processed*;




2. Access the "One-Way Analysis of Variance (ANOVA)" tab, in the tab box located at the left of the page. The options regarding this type of analysis will appear at the right;
3. Set the options regarding the analysis and click "Submit" button;

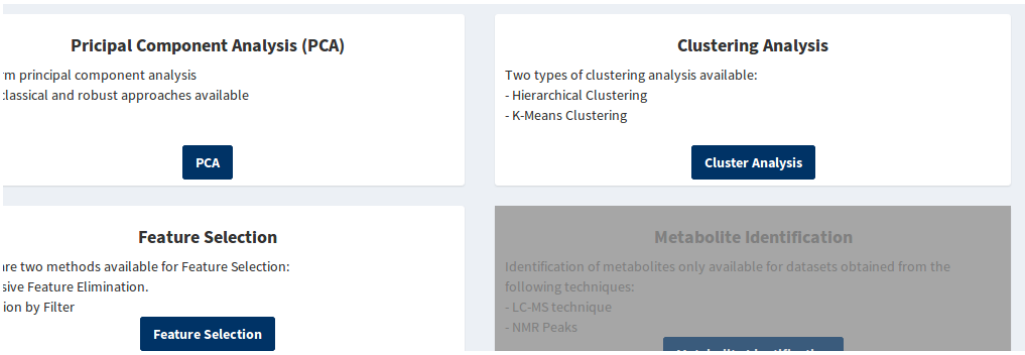



4. Once this analysis is finished, the website redirects the user to the corresponding results page. *For better understanding what information the results contain, go to subsection One-Way ANOVA in section 2.9.1.*

## 5.6 Hierarchical Clustering Analysis

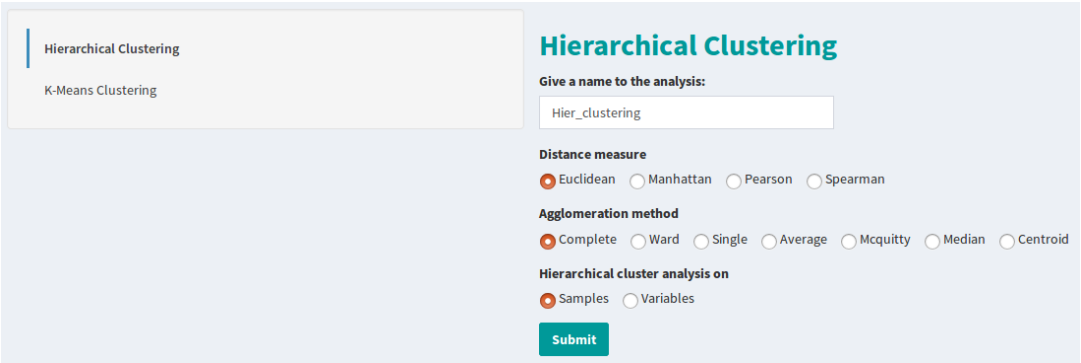
To perform hierarchical clustering on this data, the following could be done:

1. Enter the "Clustering Analysis" box in the "Results Analysis" page while the dataset being used is *data_processed*;




2. Access the "Hierarchical Clustering" tab, in the tab box located at the left of the page. The options regarding this type of analysis will appear at the right;

3. Set the options regarding the analysis and click "Submit" button;

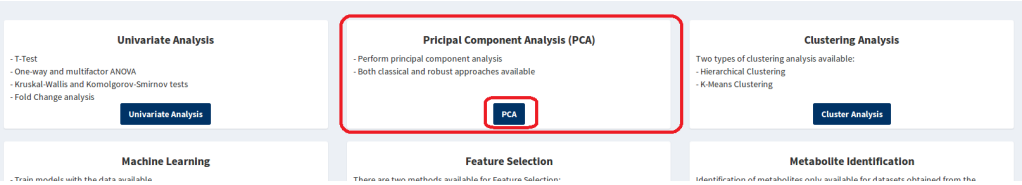



4. Once this analysis is finished, the website redirects the user to the corresponding results page. *For better understanding what information the results contain, go to subsection Hierarchical Clustering in section 2.9.3.*

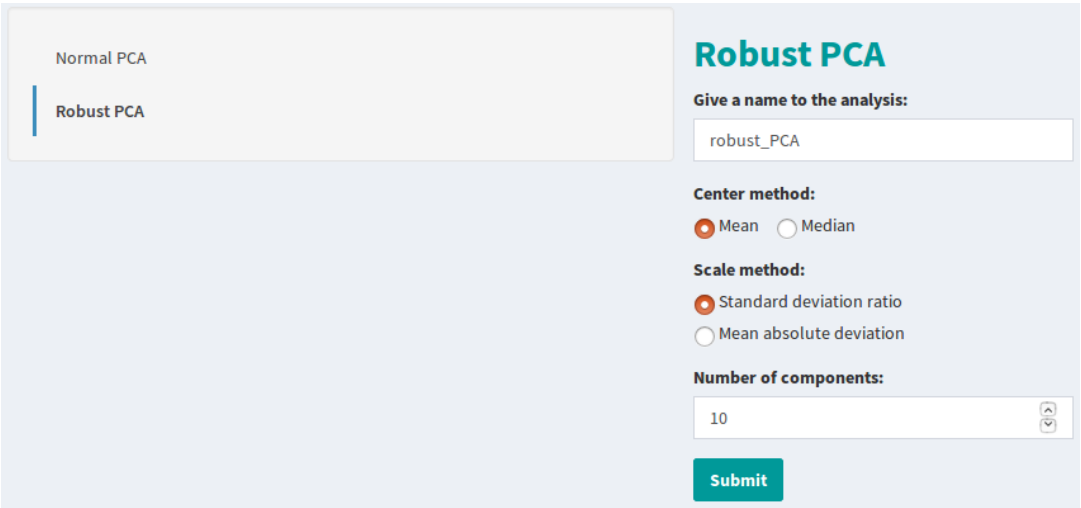
## 5.7 Principal Components Analysis

A PCA can also be performed on this dataset:

1. Enter the "Principal Component Analysis (PCA)" box in the "Results Analysis" page while the dataset being used is *data_processed*;




2. Access the "Robust PCA" tab, in the tab box located at the left of the page. The options regarding this type of analysis will appear at the right;
3. Set the options regarding the analysis and click "Submit" button;




4. Once this analysis is finished, the website redirects the user to the corresponding results page. *For better understanding what information the results contain, go to section 2.9.2.*

# 6. IR Spectra: Cassava PPD

## 6.1 Where to find the data

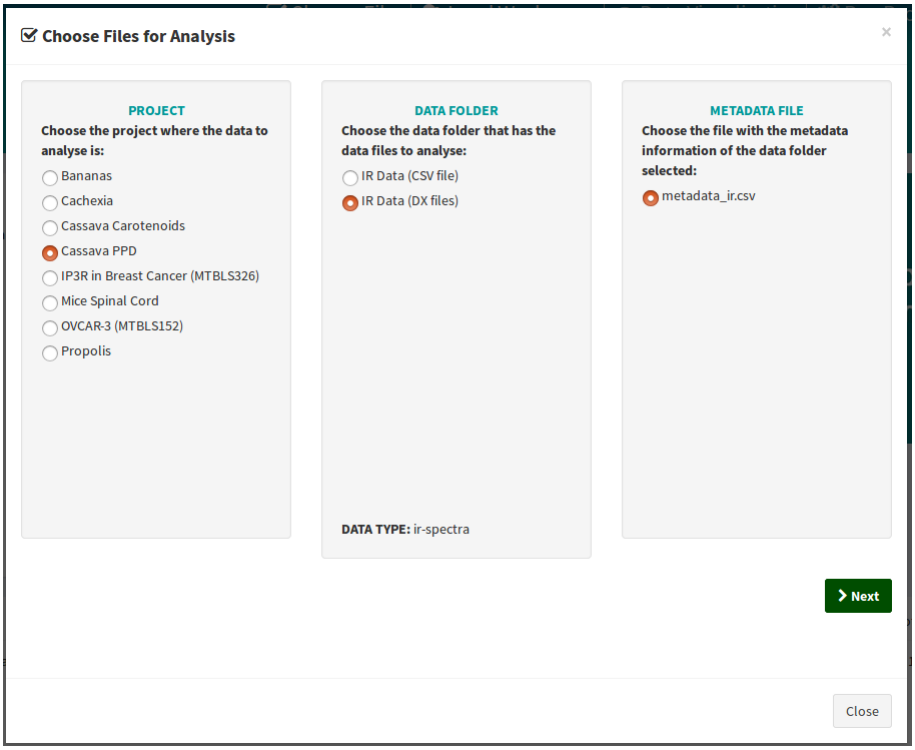
The aim of the present study [4] was to identify and discriminate changes in the chemical and enzymatic composition of cassava genotypes samples during post-harvest deterioration.

The samples used in this study, acquired using the IR spectroscopy with a spectral window of 4000 to 400 $cm^{-1}$, were stored in the public project *Cassava PPD*, under the data folder *IR Data (DX files)*. Regarding the metadata, the file *metadata_ir.csv* is given.

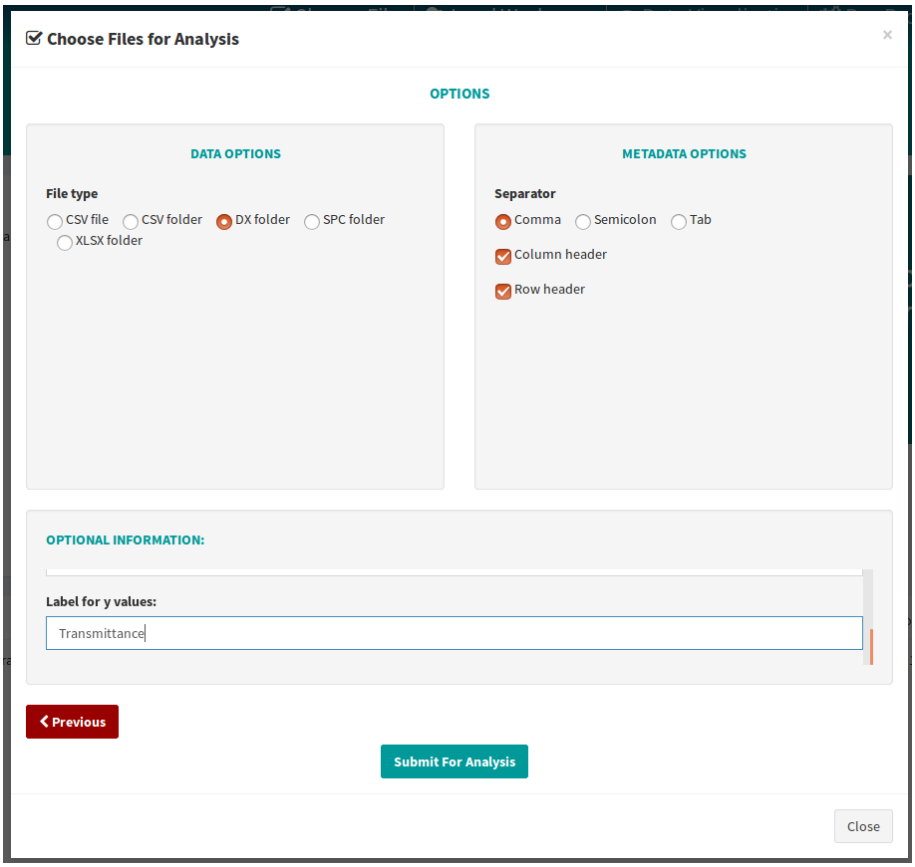
There are a total of 80 samples were collected, 16 samples with 5 replicates each. Samples were collected fresh (0 days of deterioration), and with 3, 5, 8 and 11 days. Samples were from four different varieties SCS 253 Sangão (SAN); Branco (BRA); IAC576-70-Instituto Agronômico de Campinas (IAC); and Oriental (ORI).

## 6.2 Choosing the files for analysis

1. Enter your user account;
2. Copy the public project in question, named *Cassava PPD*, into your account:
   Go to the "Public Projects" page, accessible through the sidebar panel;
   Select the project in the table of the *Community projects* box and click the button "Import Project";
3. Click the "Choose Files" button, present in the header panel;
4. Choose the project, data folder and metadata file in question and click the "> Next" button;




5. This will lead to the window were the options regarding the data and metadata files are set, so that they are read and processed correctly. The options to set are the following:



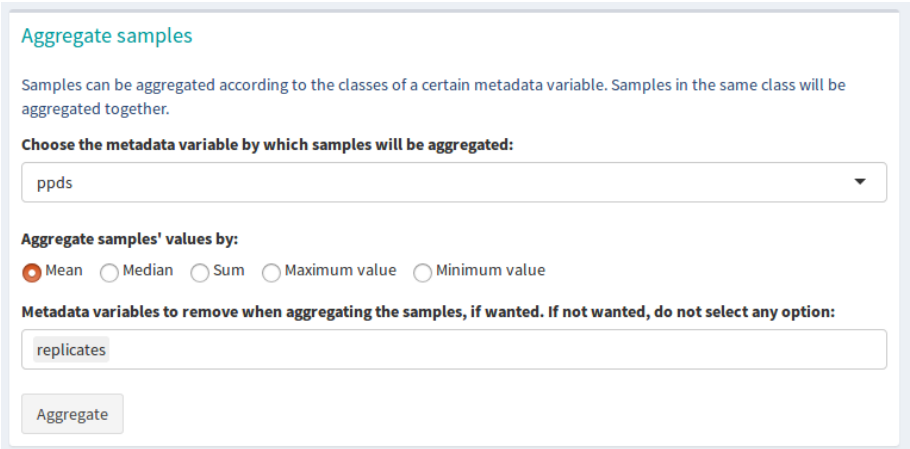
6. With this, you are able to click the button "Submit For Analysis" to finalize the submission of the data to analyse.

## 6.3 Pre-Process the data

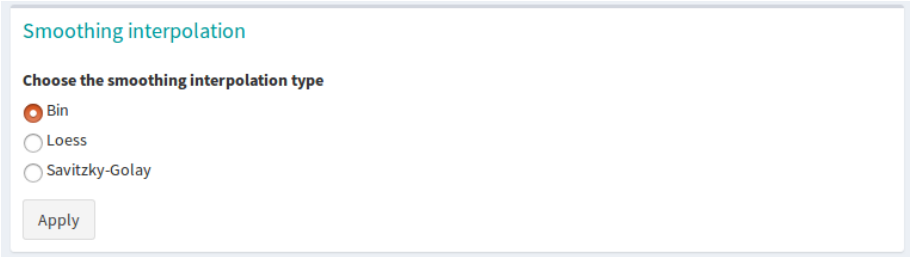
The following pre-processing pipeline should be applied to perform the analysis mentioned below:

1. Go to the "Pre-Processing" page, accessible through the header panel;
2. Convert the metadata variable representing the days of post-harvest physiological deterioration (ppds) from numeric to factor (so it can be used for the classification models in machine learning);

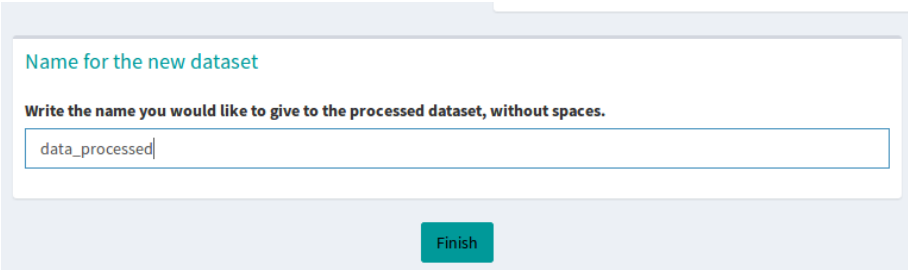

3. Aggregate the different replicates of each sample in one single sample. Because there different replicates at each days of sample collection, the sample aggregation is done according to the ppds metadata variable;




4. Perform smooth interpolation, by selecting the method "Bin";




5. Name the dataset (*data_processed*) and click the "Finish" button;

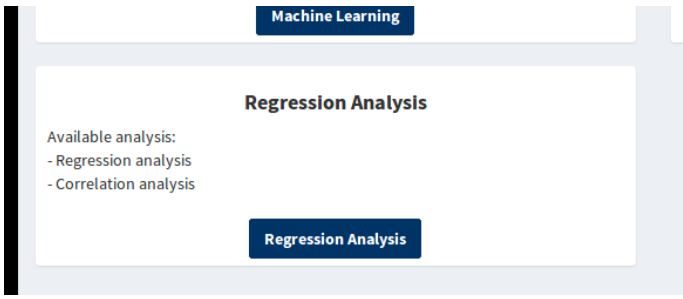



6. With this, the dataset being currently in use will automatically change to the newly created dataset.

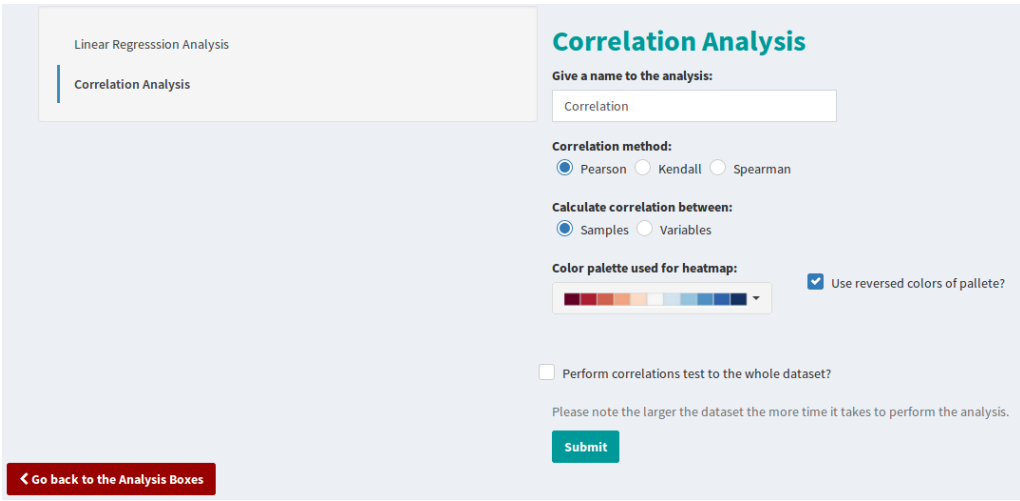
## 6.4 Correlation Analysis

To perform a correlation analysis, you could perform the following:

1. Enter the "Regression Analysis" box in the "Results Analysis" page while the dataset being used is *data_processed*;




2. Access the "Correlation Analysis" tab, in the tab box located at the left of the page. The options regarding this type of analysis will appear at the right;
3. Set the options regarding the analysis and click "Submit" button;

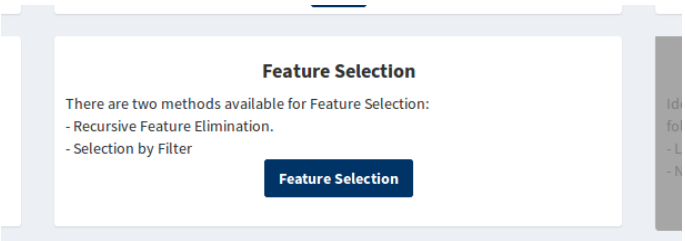



4. Once this analysis is finished, the website redirects the user to the corresponding results page. *For better understanding what information the results contain, go to subsection Linear Regression Analysis in section 2.9.7.*

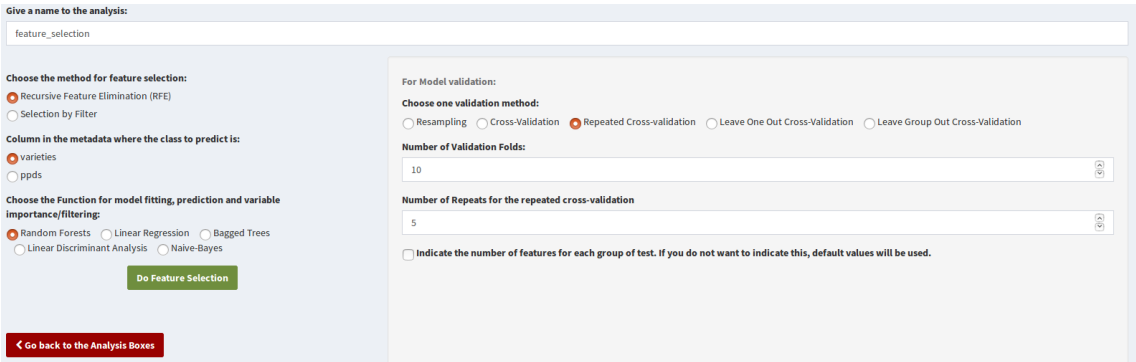
## 6.5 Feature Selection

To perform feature selection, you could do the following:

1. Enter the "Feature Selection" box in the "Results Analysis" page while the dataset being used is *data_processed*;




2. Set the options regarding the analysis and click "Do Feature Selection" button;

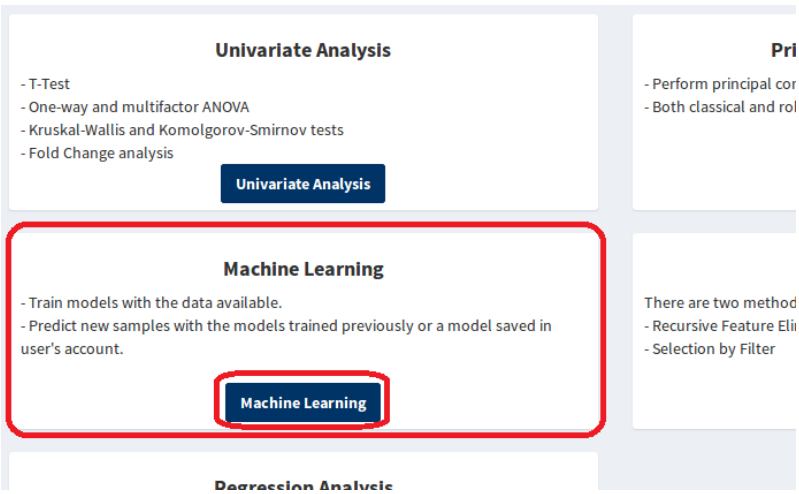



3. Once this analysis is finished, the website redirects the user to the corresponding results page. *For better understanding what information the results contain, go to section 2.9.5.*

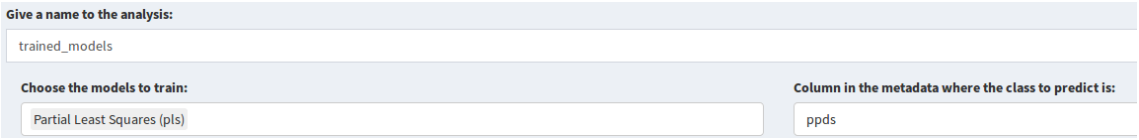
## 6.6 Machine Learning

Finally, to perform machine learning, you could perform as follows:

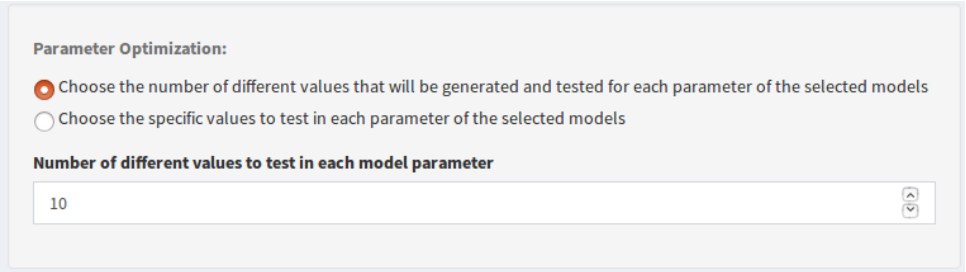
1. Enter the "Machine Learning" box in the "Results Analysis" page while the dataset being used is *data_processed*;



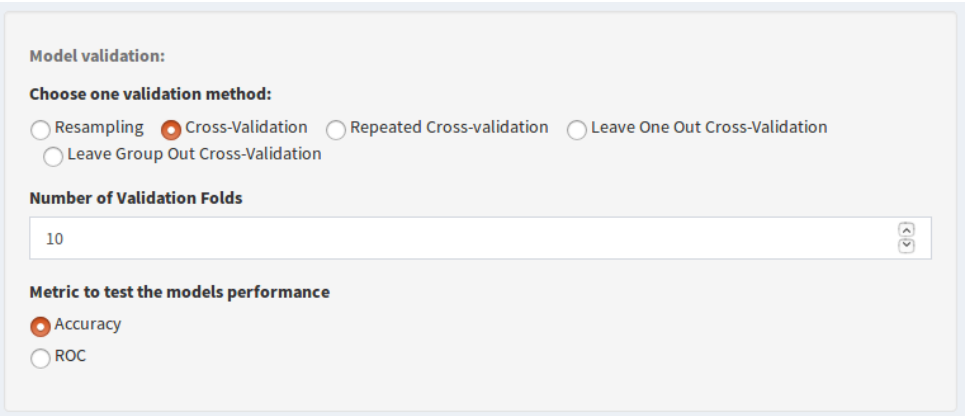
2. Access the "Train Models" page through the button with the same name located at the top of the page;
3. Set the options regarding the analysis:
   Give a name to the analysis, the type of models to train and the metadata variable that will be used to predict:




   Set the parameter optimization options:




   And set the model validation options:




4. Click the "Train models";
5. Once this analysis is finished, the website redirects the user to the corresponding results page. *For better understanding what information the results contain, go to section 2.9.4.*

# Bibliography

## Articles


[Mar+16] Marcelo Maraschin et al. "Metabolic Profiling and Classification of Propolis Samples from Southern Brazil: An NMR-Based Platform Coupled with Machine Learning". In: *Journal of Natural Products* (2016) (cited on page 105).

[Sag+04] Alan Saghatelian et al. "Assignment of endogenous substrates to enzymes by global metabolite profiling". In: *Biochemistry* (2004) (cited on page 117).

[Tom+15] MM Tomazzoli et al. "Discrimination of Brazilian propolis according to the seasoning using chemometrics and machine learning based on UV-Vis scanning data". In: *Journal of Integrative Bioinformatics* (2015) (cited on page 121).

[Uar+14] VG Uarrota et al. "Metabolomics combined with chemometric tools (PCA, HCA, PLS-DA and SVM) for screening cassava (Manihot esculenta Crantz) roots during postharvest physiological deterioration." In: *Food Chemistry* (2014) (cited on page 129).