GC Image[©] Users' Manual

Investigator ML™

Version 2018r1.0 (15 June 2018), GC Image, LLC

Contents

---

1. Introduction

Investigator ML is a tool for supervised pattern recognition and classification. With input of a list of chromatograms, each with name, sample class, and feature values, Investigator ML allows the user to:

1. Configure the chromatograms into training and testing sets and select features to be used for pattern recognition and classification.
2. Build and apply a classification model for pattern recognition.

Investigator ML provides various tools for visualization and reporting on the chromatograms, features, feature patterns, classification models, and classification results.

Although there are many details to understanding and performing pattern recognition and classification, the three basic steps for using Investigator ML are simple, as illustrated in **Figure 1**:

1. Import classification feature data.
2. Perform classification, e.g., using linear discriminant analysis (LDA) with **Action -> LDA**.
3. Save the results, including a Summary Report and associated tables with **File -> Save As…**.

Of course, Investigator ML has many additional options for visualization, refinement, and analysis, but getting started is as easy as one-two-three. The remainder of this *Users' Manual* is organized into four sections: Input/Output, Item and Feature Analysis, Pattern Recognition and Classification, and Configuration.

2. Input/Output

There are three ways to open data in Investigator ML.

- Export classification data from the Investigator™ program, using the Investigator ML plugin.
- Start Investigator ML, then import data from a CSV-format spreadsheet file, using the **File -> Import…** menu option.
- Start Investigator ML, open data that was saved to a file during a previous Investigator ML session, using **File -> Open…** menu option.

These three methods are described in the following subsections.

2.1. Export Data from Investigator

The GC Image Investigator program can be equipped with a plugin to open Investigator ML and export data to it. In the Investigator program, select Investigator ML from the Tools menu, as shown in **Figure 2**. (If the Tools menu does not show Investigator ML, it must be installed.) Then, a popup dialog, also shown in **Figure 2**, allows the user to configure the export of data to Investigator ML:

- The Feature Type selector designates exporting of either a Blob feature or Area feature.
- The Attribute selector designates the feature to be exported, e.g., Volume, Percent Response, etc.
- The Class selectors designate the classes of samples to be used for classification. (Currently, only binary classification, i.e., with two classes, can be performed.)
- The Unused and Used Chromatograms lists specify the chromatograms for which feature vectors are exported. Only vectors for the chromatograms in the Used list are exported. When a class selector is changed, these lists are updated with available chromatograms in those classes.

After the export configuration is set, click the Submit button to export feature vectors into Investigator ML.

2.2. Import Data from a CSV-Format Spreadsheet File

After starting Investigator ML, the **File -> Import…** menu option, illustrated in **Figure 1**, Step 1, invokes a dialog to select a CSV-format data file for input. If the import operation is selected while the current session already has active classification data, then the user first is prompted to confirm that the current analysis will be abandoned.

The CSV-format spreadsheet, such as is produced by GC Image Investigator, must have column headers in the first row. There are three types of columns:

- NAME: One and only one column contains the name of the chromatogram. By default, the column header for the Chromatogram Name column is “NAME”. Investigator ML can be configured to recognize a different header for the Chromatogram Name column. (See **Section 7**.)
- CLASS: One and only one column contains the class of the chromatogram. By default, the column header for the Chromatogram Class column is “CLASS”. Investigator ML can be configured to recognize a different header for the Chromatogram Class column. (See **Section 7**.)
- FEATURE: The other column(s) contain feature values. The column header for each feature is the feature name, typically the compound name or other identifier for the feature, e.g., “octanoic acid”. There must be at least one feature column.

Each subsequent row of the spreadsheet gives the name, class, and feature value(s) of a unique chromatogram. As Investigator ML imports each line, it assigns each chromatogram a unique ID from its ordinal position in the list, for convenient reference.

The convention for naming chromatograms in GC Image allows parsing of the Sample Name (or source Vial Name) for each chromatogram. In this convention, the Chromatogram Name begins with the Sample Name, which is terminated by a delimiter string. For example, given the chromatogram name “CabernetCC2011_Batalha_Run01_Img01.gci” and the run delimiter string “_Run”, the sample name is “CabernetCC2011_Batalha”. As each new sample name is recorded, a Sample ID is assigned for convenient reference. The Sample ID may not be unique because multiple chromatograms from the same sample (e.g., chromatographic replicates) will have the same Sample ID.

Information about replicate chromatograms from the same sample is important for avoiding biased classification results. Investigator ML recognizes when the same sample name recurs and assigns a Replicate ID for convenient reference of the chromatograms from that sample. For example, given chromatogram names “CabernetCC2011_Batalha_Run01_Img01.gci” and CabernetCC2011_Batalha_Run02_Img01.gci”, both chromatograms are recognized to have the same source (“CabernetCC2011_Batalha”) and so will have the same Sample ID, but the first chromatogram from that sample will be assigned Replicate ID of 1 and the second chromatogram will be assigned Replicate ID of 2.

The Class Name for each chromatogram is a character string. For example, if the goal of classification is to distinguish Merlot wines from Chardonnay wines, then the class names could be “Merlot” and “Chardonnay”. Every chromatogram from the same source sample must have the same class.

Currently, the software supports only binary classification, so exactly two different class names must appear in the Class Name column and the class name in each row must be one of those two names.

All other columns for a chromatogram contain the measured feature values. Currently, the software supports only numeric feature values.

Part of a CSV-format spreadsheet, opened in Microsoft Excel, is illustrated in **Figure 3**.

2.3. Save and Open Data

If an Investigator ML session has active classification data, the analysis, i.e., the data and any classifiers built for the data, can be saved to a file and for opening later. The **File -> Save As…** menu option is available whenever the session has active classification data. The **File -> Save…** menu option is available immediately after opening a previously saved analysis and after an analysis with imported data is saved.

The **File -> Open…** menu option is always available. Just as for the import operation, if the open operation is selected while the session already has active classification data, then the user is prompted to confirm that the current analysis will be abandoned.

2.4. Close and Exit

If the session has active classification data, the analysis of that data can be closed by selecting the **File -> Close** menu option. When this option selected, the user first is prompted to confirm that the current analysis will be abandoned.

The **File -> Exit** menu option first closes the current analysis (if there is active classification data and the user confirms that it will be abandoned) and then exits the program.

3. Item and Feature Analysis

The input data for classification is summarized in two tabbed panes:

- Chromatograms: The Chromatograms tab supports visualization, analysis, and selection of the items, i.e., chromatogram, for classification.
- Features: The Features tab supports visualization, analysis, and selection of the features, e.g., the vectors of detector responses for peaks or regions that characterize each chromatogram.

3.1. Item Analysis

The Item Analysis tab, illustrated in **Figure 4**, presents two tools: an **Items Table**, which lists all chromatograms, and an **Item Analysis Graph**, which plots the Used features for a designated chromatogram item.

Items Table

The Items Table lists all of the chromatograms, one per line, with associated metadata and tool interfaces. The metadata are:

- ID: The Chromatogram ID is a unique identification number assigned from the item's ordinal position in the original list of chromatograms. The Chromatogram ID allows easy differentiation of the chromatograms.
- Sample: The Sample ID number is assigned from the ordinal position of the first replicate of the sample (or vial) in the original list of chromatograms. The Sample ID conveniently differentiates

source samples, but replicate chromatograms from the same source sample have the same Sample ID.

- Replicate: The Replicate ID number is assigned from the ordinal position of replicates from the same sample in the original list of chromatograms. The Replicate ID conveniently differentiates between replicates from the same sample.
- Name: The Chromatogram Name is given in the original data.
- Class: The Chromatogram Class is the class of the sample given in the original data. The Class of each chromatogram from the same sample (i.e., replicates) must be identical.
- Use: The Use Status of the chromatogram item indicates its role for classification analysis, one of: TRAINING, TESTING, or UNUSED. The role of the Use Status setting is more fully explained in **Section 4** in relation to Classification. Initially, the Use Status of all imported chromatograms is set to TRAINING.
- Z-Compare: The Z-Compare Value for each chromatogram is computed from its feature values and the feature value statistics of the training-set items. This value is useful for sorting items from large positive values (feature values indicative of Class 0) to large negative values (feature values indicative of Class 1).

Figure 4 shows the Items Table for the sample CSV-format spreadsheet shown in **Figure 1**. For example, ID=10, Sample=5, and Replicate=2, indicates that the chromatogram was represented on the 10th line of the original data, is from the 5th different sample in those first 10 lines, and is the 2nd replicate for that sample.

The Items Table is sortable on any column: a left-click on a column header cycles through forward order, reverse order, and unordered (for multiple sorting levels) for the column values.

The decimal number columns provide control over the number of decimal digits shown for each value: a right-click on the column header opens a precision dialog with up and down arrows to increase and decrease the number of visible digits.

The Use Status of items for analysis can be set by first selecting items and then clicking one of the buttons below the table. Items are selected by left-click on a single row, shift-left-click to designate a range of rows, and control-left-click to toggle selection of a single row. After selection, the buttons below the table allow setting the Use Status of the selected items to TRAINING, TESTING, or UNUSED. Alternatively, the values in individual cells of the Use Status column can be set directly by double-left-click in the Use cell and selecting from the pull-down options.

The chromatogram item that is designated for the Item Analysis Graph is selected by clicking the associated radio button in the Plot column of the Items Table.

The Items Table can be saved to a file in CSV or XLS format using the buttons above the top-right of the table.

Items Graph

The Items Graph, illustrated in **Figure 5**, employs the new Graphical Classification Interface (GCI)™ to plot the features of the chromatogram that is designated in the Items Table. Along the *x*-axis of the GCI plot is the relative difference of the feature value with the mid-point between the means of the classes. Along the *y*-axis is the relative difference of the class means, which is a Signed F Ratio. Then, given a specific chromatogram, the coordinates of each point depends on the value of that feature in the chromatogram:

$$GCI(f,i) := \left(\left(f(i) - \frac{\left(\mu_0(i) + \mu_1(i)\right)}{2}\right)v(i), \left(\frac{\left(\mu_0(i) - \mu_1(i)\right)}{2}\right)v(i)\right)$$

where $f(i)$ is the i^{th} feature value of chromatogram item *f*, $\mu_c(i)$ is the mean for Feature *i* in Class *c*, and *v*(*i*) is the reciprocal of the square-root of the summed variances for Feature *i* in Class c:

$$v(i) = (\sigma_0^2(i) + \sigma_1^2(i))^{-2}$$

where $\sigma_c^2(i)$ is the variance for Feature *i* in Class *c*. With this coordinate system, the mean values for Class 0 fall on the diagonal from bottom-left to top-right and the mean values for Class 1 fall on the diagonal from top-left to bottom-right. The *x*-axis values (abscissas) are the relative differences with the midpoint between means, so features with values greater than the mid-point between means are to the right and values less than the mid-point of the means are to the left. The y-axis values (ordinates) are Signed F Ratios, so features with large positive Signed F Ratios are at the top of the graph and features with large negative Signed F Ratios are at the bottom of the graph.

The abscissa value (plotted on the *x*-axis) for each feature is computed with reference to the midpoint between class means and relative to the feature's summed class variances. That is desirable for using the same axis for all features and both classes. However, it also is interesting to know for which class each item's value has a smaller *z*-score, i.e., the variances are considered separately for each class. Points are shown with different shapes and colors to distinguish which feature values are relatively closer to the mean for Class 0 and which features are relatively closer to Class 1, using *z*-scores as the distance measures.

The GCI plot is useful for both: (a) quickly visualizing the trend of many feature values for a specific chromatogram and (b) visually focusing on individual features that are either strongly indicative of the chromatogram's class or significantly atypical for the chromatogram's class. Those values that are closer to the means for Class 0 than the means for Class 1 are plotted in the first and third quadrants (top-right and bottom-left) and those values that are closer to the means for Class 1 than the means for Class 0 are in the second and fourth quadrants (top-left and bottom-right). So, for a specific chromatogram from Class 0, the trend of features should be along the center diagonal through the first and third quadrants and any features plotted well into the second or fourth quadrants have atypical values for that chromatogram's class. Likewise, for a specific chromatogram from Class 1, the trend of features should be along the center diagonal through the second and fourth quadrants and any features plotted well into the first or third quadrants have atypical values for that chromatogram's class. **Figure 5** illustrates two example GCI plots (for wine analysis): (A) with a trend consistent with Class 0 (Cabernet) and (B) with a trend consistent with Class 1 (Merlot).

The Item Analysis Graph has mouse-overs such that if the cursor hovers above a plotted point, then the feature ID and name for that point are shown. The mouse-overs are especially useful for identifying those features that are most strongly indicative of the class (i.e., far from the origin, along the desirable diagonal) and those that deviate most from the class expectations (i.e., far from the origin, along the undesirable diagonal).

The GCI plot can be converted to an image and either copied to the clipboard, saved to a file, or printed. To initiate one of these actions, right-click on the Item Analysis Graph, then select from the popup menu.

3.2. Feature Analysis

The Feature Analysis tab, illustrated in **Figure 6**, presents two tools: a **Features Table**, which lists all features, and a **Feature Analysis Graph**, which plots training-set chromatogram values for a designated feature.

Features Table

The Features Table lists all of the features, one per line, with associated metadata and tool interfaces. The metadata are:

- ID: The Feature ID is a unique identification number assigned from the feature's ordinal position in the original data, which allows easy differentiation of the features.
- Name: The Feature Name is given in the column header of the original data.
- Use: The Use Status of the feature indicates its role for classification analysis, one of: USED or UNUSED. The role of the Use Status setting is more fully explained in **Section 4**, in relation to classification. Initially, the Use Status of an imported feature is set to USED.
- F Ratio: The Signed F Ratio is a measure of the relative difference of class means for a specific feature from training-set items. It is computed as:

$$Signed\_F\_Ratio(i) := \frac{\big(\mu_0(i) - \mu_1(i)\big)/2}{\sqrt{\sigma_0^2(i) + \sigma_1^2(i)}}$$

  where $\mu_c(i)$ and $\sigma_c^2(i)$ are the mean and variance in Class *c* for Feature *i*. This is the ordinate for the Items Analysis Graph. For each feature, if the mean value for Class 0 in the training set is larger than the mean value for Class 1 in the training set, then the Signed F Ratio is positive; but, if the mean value for Class 0 in the training set is smaller than the mean value for Class 1 in the training set, then the Signed F Ratio is negative. If the difference-magnitude between the class means in the numerator is large, relative to the within-class deviations in the denominator, then the magnitude of the Signed F Ratio is large, meaning that the feature is a good differential marker; but, if the difference-magnitude between the class means is small, relative to the within-class deviations, then the magnitude of the Signed F Ratio is small, meaning the feature is not a good differential marker. The Features Table shows Signed F Ratios in numeric text and F Ratio magnitudes in a bar chart.
- P Value: The P Value is a measure of the probability that there is no statistical difference between the class means, i.e., the null hypothesis. In addition to the statistics of class means and variances (used in the signed F Ratio), the P Value depends on the number of training-set items used for those statistical estimates. A relative difference of means observed for a larger number of items has greater statistical significance expressed in the P Value. The table shows P Values in numeric text and a bar chart.

Figure 6 shows the Features Table for the sample CSV-format spreadsheet shown in **Figure 1**.

The Figures Table is sortable on any column except the bar charts: a left-click on a column header cycles through forward order, reverse order, and unordered (for multiple sorting levels).

The decimal number columns provide control over the number of decimal digits shown for each value: a right-click on the column header opens a precision dialog with up and down arrows to increase and decrease the number of visible digits.

The Use Status of features for analysis can be set by first selecting features and then clicking one of the buttons below the table. Features are selected by left-click on a single row, shift-left-click to designate a range of rows, and control-left-click to toggle selection of a single row. After selection, the buttons allow setting the Use Status of the selected features to USED or UNUSED. Alternatively, the values in individual cells of the Use Status column can be set directly by double-left-click in the Use cell and selecting from the pull-down options.

The feature that is designated for the Feature Analysis Graph is selected by clicking the associated radio button in the Plot column of the Features Table.

The Features Table can be saved to a file in CSV or XLS format using the buttons above the top-right of the table.

Features Graph

The Features Graph, illustrated in **Figure 7**, plots the values for the feature that is designated in the Features Table from each of the training chromatograms. Along the *x*-axis is the feature value. Along the *y*-axis is the Gaussian Distribution Model probability for the feature values in each of the classes. Then, given a specific chromatogram, if the feature value is equal to the mean for its sample class, it will be plotted at the peak of the Gaussian Distribution Model. Feature values that deviate from the class mean have smaller probabilities on the *y*-axis as a function of their difference with the mean. The points are plotted with different shapes and colors for each class.

The Feature Analysis Graph is useful for both: (a) quickly visualizing the distributions of the designated feature values for many chromatograms in the different classes and (b) visually focusing on individual chromatograms that have an atypical feature value for the class (i.e., are far out on the tails of the Gaussian Distribution Model). **Figure 7** shows an example Feature Analysis Graph with (A) relatively different class means, indicating a feature that may be useful for distinguishing classes, and (B) relatively similar class means, indicating a feature that may not be useful for distinguishing classes.

The Feature Analysis Graph has mouse-overs such that if the cursor hovers above a plotted point, then the chromatogram ID and name for that point are shown. The mouse-overs are especially useful for identifying those chromatograms that are most consistent with the class (i.e., near the center of the Gaussian Distribution Model) and those that deviate most from the class expectations (i.e., are on the tails of the Gaussian Distribution Model).

The Feature Analysis Graph can be converted to an image and either copied to the clipboard, saved to a file, or printed. To initiate one of these actions, right-click on the Feature Analysis Graph.

4. Pattern Recognition and Classification

4.1. Supervised Training, Classification Performance, and Item and Feature Selection

Investigator ML builds a classification model with supervised training. In supervised training, a set of items with known classes provide instances of feature values that are used to construct a model, e.g., using statistics of the feature values in the training set. Ideally, in order to build a valid model, the training set should have representative items from each of the classes. As described in **Section 3.1**, the chromatogram items whose Use Status is set to TRAINING form the training set. Currently, Investigator ML does not support unsupervised classification.

Accuracy, Confusion Matrix, Recall, and Precision

Accuracy is a fundamental measure of the quality of a classification model. Accuracy measures the model's success in predicting the class of all items that are classified and frequently is expressed as a percentage:

$$\%Accuracy \ := \ 100 * N_{correct}/N_{classified}$$

where $N_{correct}$ is the number of correctly classified items and $N_{classified}$ is the number of items for which classification is performed. The training-set accuracy is computed from the results for classifying all training-set items, testing-set accuracy is computed for classifying all testing-set items, and validation accuracy is computed for classifying all validation items. If all items are classified correctly, then the %Accuracy is 100.

A confusion matrix summarizes the numbers of correct and incorrect classifications in each class. There is a row for each true class and a column for every predicted class. Example confusion matrices are illustrated in **Figure 8**, with one for training and one for validation (discussed in **Section 4**). The accuracy is computed from the sum of numbers along the diagonal (i.e., the correct classifications) divided by the sum of numbers in all cells (i.e., the number of items classified).

Two additional measures of classification success, recall and precision, are computed easily from a confusion matrix. The recall for each class is the number of items in the class that are correctly classified divided by the total number of items in that class, i.e., the accuracy for items in the class. The precision for each class is the number of items in the class that are correctly classified divided by the total number of items that are predicted to be in that class. These measures are especially useful if there are differential costs associated with different types of errors, e.g., failing to detect a disease state may be a more costly error than falsely detecting a disease state.

Over-Fitting, Cross-Validation, and Testing

The classification model is built to fit the training set and so will be biased by the particulars and idiosyncrasies of those items in the training set, including atypical samples, problematic chromatography, and noise. In fitting the model to the training set, there is a danger of over-fitting to the training set's particulars, especially if the feature-space is high dimensional (i.e., there are many features).

Investigator ML uses leave-one-out cross-validation to assess over-fitting and gauge a model's general applicability or robustness. In leave-one-out cross-validation, all replicates (or twins) of one source sample are removed from the training set to create a validation set, next a classification model is built from the reduced training set, and then the class of the replicates in the validation set are predicted from the classifier. In this way, no replicates of the validation sample are used in building (or biasing) the classification model. This process is repeated for each of the distinct samples, each with possible replicates, producing as many independent validation predictions as there items in the original training set. The performance for predictions in cross-validation may be a better indicator of the model's general applicability than the performance for the full training set. Accuracy, recall, and precision also can be computed for the classifications performed in cross-validation. **Figure 8** illustrates results for both the training-set classifier and for validation classifications.

Chromatogram items also can be assigned to a testing set (and excluded from the training set) in order to evaluate the performance of the classification model. Ideally, in order to test classification performance, both the training and testing sets should have a representative samplings of items from each class. As described in **Section 3.1**, the chromatogram items whose Use Status is set to TESTING form the testing set. It is not necessary to have a testing set, in which case there are no testing results (other than cross-validation). If the testing set is a representative sampling, then the testing set performance should be a good indicator of the general performance of the classification model.

By default, the Use Status for all imported items are set to TRAINING, but the status of individual items can be changed so that an item is put in the testing set or is unused in either training or testing (e.g., if an item chromatogram exhibits chromatographic problems).

By default, the Use Status for all imported features are set to USED, but the status of individual features can be changed so that they are unused in classification. Reducing the number of features used, e.g., by not using features that are not especially useful for classification, reduces the dimensionality of the feature space. Reducing the feature space in turn reduces the computation for classification, can mitigate over-fitting, and can simplify interpretation of the classification model. Building an initial classification model with all features may provide insights into the roles of individual features that then can be used to guide feature reduction.

4.2. Result Reporting

Classification can be repeated multiple times, e.g., with different training and testing sets and/or different feature sets and/or different classifier settings and/or different classification methods. Each time classification is performed in Investigator ML, a Result tab is added to the interface titled with an integer ordering the results and a string indicating the classification method. Each Result tab has three or four tabs with classification results: Summary Report, Training Set Report, Testing Set Report (if the testing set is not empty), and Feature Set Report.

The first sub-tab for a classification result is the Summary Report. Just above the Summary Report are three buttons too copy, save, or print an image of the report. The Summary report has four sections: a description of the data and classification, accuracy, confusion matrices with recall and precision, and scores. The description, accuracy results, and confusion matrices are shown in **Figure 8** and as are described in **Section 4.1**. The scores reporting on the Summary Report sub-tab and the other result sub-tabs differ according to classification method and so are described below, accompanying descriptions of those methods.

There are many different classification methods. Each method can require different parameters and can result in different models. Currently, Investigator ML supports Linear Discriminant Analysis (LDA) and *k*-Nearest Neighbors (KNN) classification. More methods will be implemented in future versions.

5. Linear Discriminant Analysis

5.1. Algorithm Description

Linear Discriminant Analysis (LDA) LDA is a widely used classification method that yields a model that facilitates direct insights in to the roles of different features in classification and the particulars of the model for different item chromatograms. LDA is based on assumptions about the feature value covariances and under those conditions is Bayesian optimal. Also, it is efficiently computed and has been shown to perform nearly as well as more complex methods for a variety of classification problems with GCxGC data.

LDA mathematically projects the points for items in the feature space onto the line that maximizes the relative separation of classes. The linear projection is simple, i.e., each feature value is multiplied by the weight for that feature and the products are summed to yield a score. So, the projected score for each item is a linear combination of its features. Then, the scores are used to classify items, e.g., items with larger scores are classified in one class and items with smaller scores are classified in the other class.

Mathematically, LDA assumes Gaussian probability distributions with the same covariance matrices for both classes. It is implemented by computing statistical estimates of the class means and covariances, and then minimizing the between-class variance relative to the within-class variances:

$$S = \frac{\vec{w}(\vec{\mu}_1 - \vec{\mu}_0)^2}{\vec{w}^T \Sigma_0 \vec{w} + \vec{w}^T \Sigma_1 \vec{w}}$$

where $\vec{\mu}_i$ is the vector of feature means for Class *i*, Σ_i is the feature covariance matrix for Class *i*, and $\vec{w}$ is the vector of model feature weights for which the expression is minimized. Then, the score s for an Item *f* with feature vector $\vec{f}$ is:

$$s = \vec{w} \cdot \vec{f}$$

Classification is based on delineation of scores at the mid-point between class means (under the assumption of equal covariance matrices, $\Sigma_0 = \Sigma_1$).

5.2. Result Reporting

For each LDA classification, Investigator ML produces three or four classification result sub-tabs: Summary Report, Training Set, Testing Set (if the testing set is not empty), and Feature Set. The description, accuracy, and precision results are reported in the Summary Report sub-tab, as described in **Section 4.2**. The Scores reporting in the Summary Report and the other sub-tabs are described here.

Scores Reporting

A Scores Graph is reported for the training-set classification, for validation classification, and for the testing-set classification (if the testing set is not empty). Examples of a training scores graph and a validation scores graph for LDA are shown in **Figure 9**. Each of the scores plots has the same structure as the Features plot, described in **Section 3.2** and pictured in **Figure 7**, but with the scores rather than the feature values. The Scores Graph for LDA shows the score for each chromatogram along the *x*-axis. Along the *y*-axis is the Gaussian Distribution Model probability for the scores of each of the classes (consistent with LDA's assumptions of Gaussian distributions). Then, for a specific chromatogram, if the score is equal to the mean for its sample class, it is plotted at the peak of the Gaussian Distribution Model. Scores that deviate from the class mean have smaller probabilities on the *y*-axis. The points are plotted with different shapes and colors for each class.

In **Figure 9**, note that for training-set classification, LDA over-fits the mapping from the high-dimensional feature space to the score dimension and so achieves large class separations that are misleading. The class separations achieved for the validation classifications are more realistic. The scores incorporate the decision-point between classes, so a positive score predicts Class 0 and a negative score predicts Class 1. In the validation classifications, five chromatograms from Class 0 have negative scores and so are misclassified into Class 1 and 7 chromatograms from Class 1 have positive scores and so are misclassified into Class 0. The summary of these results are shown in **Figure 8**.

The Scores Graph has mouse-overs such that if the cursor hovers above a plotted point, then the chromatogram ID and name for that point are shown. The mouse-overs are especially useful for identifying those chromatograms that are misclassified (i.e., far along the desired radial), those that are most consistent from the class expectations (i.e., at the top of the distribution model), and those that deviate from the class mean (in either direction).

Training Set and Testing Set Reporting

The LDA Training Set and Testing Set sub-tabs have same structure as the Chromatograms tab, described in **Section 3.1**, with an Item Set Table and an Item Score Analysis Graph. An example Training Set tab is shown in **Figure 11**. The Testing Set sub-tab is present only if chromatograms were included in the testing set.

The Item Set Table is similar to the Items Table in the Chromatograms tab, with a row for each chromatogram and columns for Chromatogram ID, Sample ID, Replicate ID, Name, and Class. Columns reporting the classification scores and predicted classes for each chromatogram are given in place of the Z-Compare value. In the Training Set tab, there are scores and predictions for both training and validation classifications, whereas the Testing Set tab reports scores and predictions only for testing classifications. The Item Set Table does not have a Use Status column, because the use of each item was fixed for the classification, but as a convenience, it is possible to select chromatogram row(s) and change the Use Status for a new classification setup in the same manner as in the Items Table. The Items Set Table also can be exported to CSV or XLS format files.

The Item Score Analysis Graph is similar to the Item Analysis Graph in the Chromatograms tab. However, the factors of both the abscissa and ordinate are multiplied by the feature weight instead of by the reciprocal of the square-root of the summed variances, i.e., $v(i)$ is replaced by $w(i)$. Thus, the Item Score Analysis Graph shows the weighted values rather than the relative values. Otherwise, the graph is the same, with mouse-overs and operations to copy, save, or print an image of the graph.

Feature Set Reporting

The structure of LDA Feature Set tab, illustrated in **Figure 11**, is similar to the Features tab, described in **Section 3.2**, with a Feature Set Table and a Weights Graph.

The Feature Set Table is similar to the Features Table in the Features tab, with a row for each feature and columns for Feature ID, Name, Signed F Ratio, and P Value. In addition, there are columns reporting training weights and average validation weights for each feature. The Feature Set Table does not have a Use Status column, because only the features used for the classification are shown, but as a convenience, it is possible to select feature row(s) and change the Use Status for a new classification setup in the same manner as in the Features Table. The Feature Set Table also can be exported to CSV or XLS format files.

The Weights Graph shows the Relative Difference of Averages (i.e., the Signed Fisher Ratios) and the Weighted Difference of Averages for each feature. These are the ordinate values, respectively, for the Items Graph in the Chromatograms tab and the Items Set Graph in the Training Set tab. Many, but not all, Weighted Difference of Averages are in line with the Signed Fisher Ratios. The differences are due to taking the features individually to compute the Signed Fisher Ratios and taking the features, with cross-correlation, as a linear combination to compute the weights. The Weights Graph has mouse-overs and operations to copy, save, or print an image.

6. K-Nearest Neighbors Classification

6.1. Algorithm Description

In contrast to LDA, *k*-Nearest Neighbors (KNN) classification makes no assumptions about the feature-value distributions except that the values of items in the same class are more likely to be similar (i.e., mathematically nearer) to one another than to items in another class(es). Therefore, KNN can be a good alternative to LDA if the feature-value distributions are especially non-Gaussian, e.g., multi-modal. KNN also differs from LDA in that it performs its calculations during classification rather than in building a model.

To classify an item, KNN searches the feature space (or the transformed feature space) for the *k* items that are nearest (i.e., most similar). Then, the number of items (or the weighted numbers) of those nearest neighbors in each class is computed. The item is classified in the class with the largest number (or weighted number) of items.

Investigator ML provides options for four KNN parameters:

- The number of neighbors, *k*, which must be an integer greater than or equal to 1. The default is 1, but that may not be the best setting.
- Normalization, i.e., the transformation of the feature space before distances are computed. Investigator ML has three options to normalize the feature values:
    - NONE, no scaling of the feature values;
    - STDEV, each feature value is divided by the standard deviation of that feature's values in the training set; and
    - RANGE, each feature value is divided by the range of that feature's values in the training set.

  The default normalization is NONE.

- Distance metric, i.e., how the distances between items are computed. Investigator ML has three options to measure distance:
    - MANHATTAN, Manhattan Distance, the distance is the sum of the feature difference magnitudes;
    - EUCLIDEAN, Euclidean Distance, the distance is the square-root of the sum of the squared feature differences;
    - EUCLIDEANSQUARED, Euclidean Squared Distance, the distance is the sum of the squared feature differences; and
    - CHEBYSHEV, Chebyshev Distance, the distance is the maximum of the feature difference magnitudes.

  The default is EUCLIDEAN.

- Weighting, i.e., how items in the set of nearest neighbors are weighted to compute the sum of items in each class. Investigator ML has three options for weighting:

  - ONE, each item is counted as 1 in the sum;
  - INVERSE, each item is counted as the multiplicative inverse (reciprocal) of its distance from the item being classified.; and
  - INVERSESQUARE, each item is counted as the square of the multiplicative inverse of its distance from the item being classified.

The default is ONE.

These settings are set in the Configuration dialog, discussed in **Section 7**.

6.2. Result Reporting

For each KNN classification, Investigator ML produces three or four classification result sub-tabs: Summary Report, Training Set, Testing Set (if the testing set is not empty), and Feature Set. The description, accuracy, and precision results are reported in the Summary Report sub-tab, as described in **Section 4.2**, and shown for KNN in **Figure 11**. The Scores reporting in the Summary Report and the other sub-tabs are described here.

Scores Reporting

A Scores Graph is reported for the training-set classification, for validation classification, and for the testing-set classification (if the testing set is not empty). Examples of a training scores graph and a validation scores graph for KNN are shown in **Figure 12**. The Chromatogram (Item) ID is given on the *x*-axis and the score for each chromatogram (item) is given on the *y*-axis. The scores are plotted as bar graphs with scores for Class 0 as positive scores and scores for Class 1 as negative scores. For each chromatogram (item), both the score for each class is given, with true scores given in a solid bar and false scores given in a striped bar.

For example, in **Figure 12.B** for the validation set, Chromatogram 6, which is from a Cabernet wine sample (Class 0), has a score of 4 for its neighbors in Class 0 (shown with a positive, solid bar) and a score of 1 for its neighbors in Class 1 (shown with a negative, striped bar). Because 4 > 1, Chromatogram 6 was classified correctly. Chromatogram 76, which is from a Merlot wine sample (Class 1), has a validation score of 4 for its neighbors in Class 0 (shown with a positive, striped bar) and a score of 1 for its neighbors in Class 1 (shown as a negative solid bar). Because 1 < 4, Chromatogram 76 was classified incorrectly. The summary of these results are shown in **Figure 8**.

The Scores Graph has mouse-overs such that if the cursor hovers above a plotted bar, then the chromatogram ID and name for that bar are shown. The mouse-overs are useful for identifying those chromatograms that are misclassified.

Training Set and Testing Set Reporting

The Training Set and Testing Set sub-tabs have an Item Set table and an Item Result Analysis graph. An example Training Set tab is shown in **Figure 14**. The Testing Set sub-tab is present only if chromatograms were included in the testing set.

The Item Set Table is similar to the Items Table in the Chromatograms tab, with a row for each chromatogram and columns for Chromatogram ID, Sample ID, Replicate ID, Name, and Class. Columns reporting the classification scores and predicted classes for each chromatogram are given in place of the Z-Compare value. In the Training Set tab, there are scores and predictions for both training and validation classifications, whereas the Testing Set tab reports scores and predictions only for testing classifications. The Item Set Table does not have a Use Status column, because the use of each item was fixed for the classification, but as a convenience, it is possible to select chromatogram row(s) and change the Use Status for a new classification setup in the same manner as in the Items Table. The Items Set Table also can be exported to CSV or XLS format files.

The Item Score Analysis Graph plots each feature for the selected chromatogram. The value on the *y*-axis for each feature is a radius that is computed as the largest feature-value difference with the selected chromatogram's feature value among the selected chromatogram's nearest neighbors. The value on the *x*-axis for each feature is the ratio of the number of same-set items to the number of all items within the radius around the selected chromatogram's feature value. A ratio of 1 means that all chromatograms with features within the radius are in the same class as the selected chromatogram, which indicates that the feature is a good indicator for that chromatogram's class. Features with ratios from 0.5 to 1.0 are somewhat to excellent predictors of the correct class and features with ratios less than 0.5 are counter indicative. The graph has mouse-overs that show the feature ID and name, which allows examinations of good and bad predictive features. Operations to copy, save, or print an image of the graph are accessible with a mouse right-click.

Feature Set Reporting

The structure of KNN Feature Set tab, illustrated in **Figure 15**, is similar to the Features tab, described in **Section 3.2**, with a Feature Set Table and a Weights Graph.

The Feature Set Table is similar to the Features Table in the Features tab, with a row for each feature and columns for Feature ID, Name, Signed F Ratio, and P Value. In addition, there are columns reporting training and validation average radii and average ratios for each feature. The Feature Set Table does not have a Use Status column, because only the features used for the classification are shown, but as a convenience, it is possible to select feature row(s) and change the Use Status for a new classification setup in the same manner as in the Features Table. The Feature Set Table also can be exported to CSV or XLS format files.

The Feature Set Graph plots the mean radius and ratio for each feature averaged over the training set. The Feature Set Graph has mouse-overs displaying the feature ID and name and operations to copy, save, or print an image.

7. Configuration

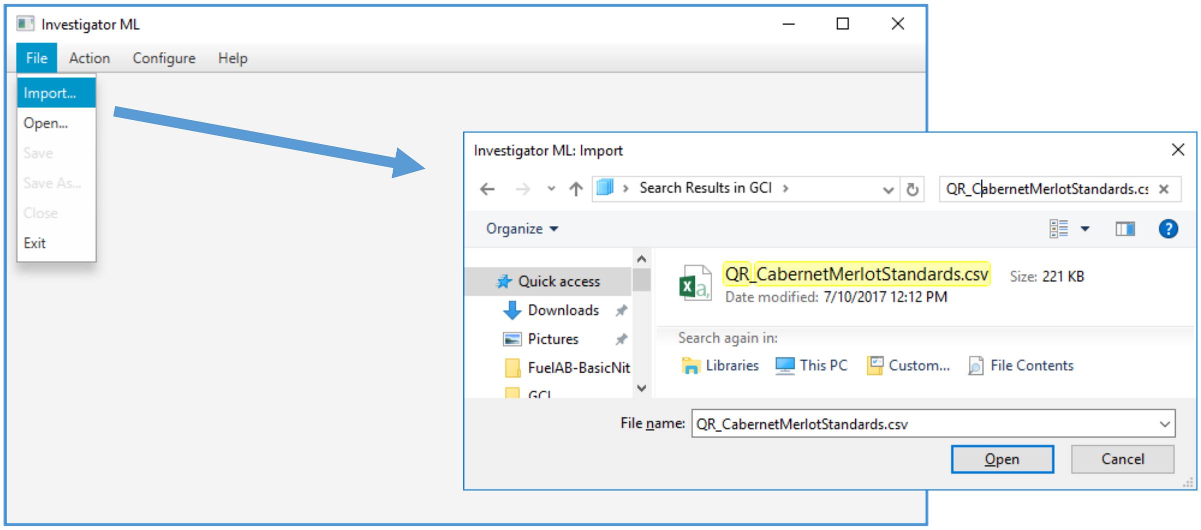
Investigator ML has two configuration tabs: Import and KNN. Both are shown in **Figure 16**.

The Import tab, in **Figure 16.A**, configures the strings for the name and class column headers and the run delimiter.

The KNN tab, in **Figure 16.B**, configures the KNN settings for normalization, distance, number of neighbors and weighting function.

Figure 1: Three simple steps for Investigator ML.

Step 1: Import the data from a CSV file (or open from Investigator).

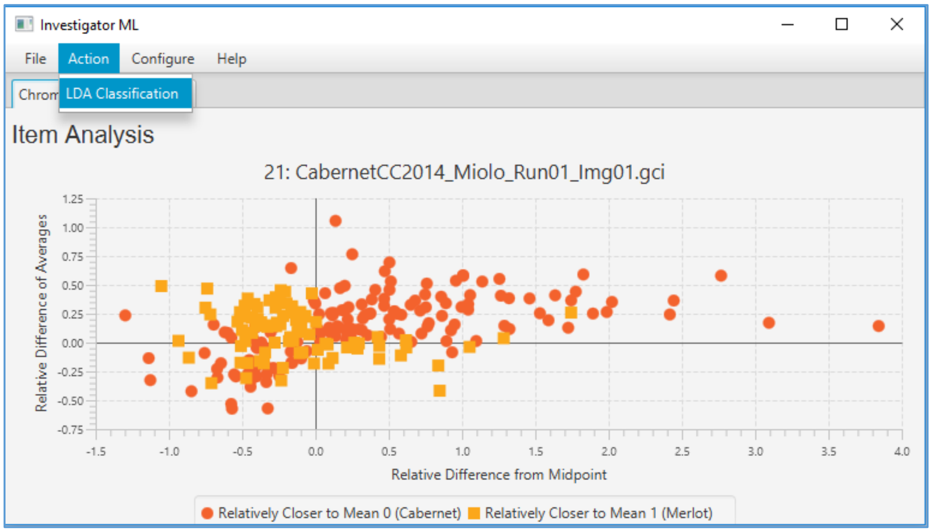


Search: QR_CabernetMerlotStandards.cs

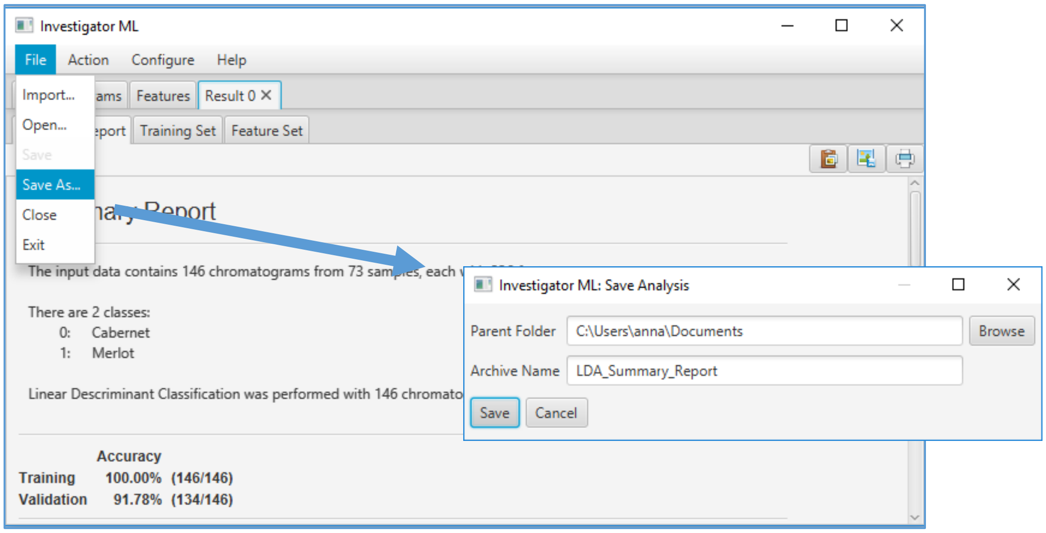
QR_CabernetMerlotStandards.csv Size: 221 KB
Date modified: 7/10/2017 12:12 PM

File name: QR_CabernetMerlotStandards.csv

Step 2: Perform classification.


Investigator ML
File
Action
Configure
Help
Chrom
LDA Classification
Item Analysis
21: CabernetCC2014_Miolo_Run01_Img01.gci
Relative Difference of Averages
1.25
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Relative Difference from Midpoint
● Relatively Closer to Mean 0 (Cabernet)
■ Relatively Closer to Mean 1 (Merlot)

Step 3: Save the results.



The input data contains 146 chromatograms from 73 samples, each

There are 2 classes:
0: Cabernet
1: Merlot

Linear Descriminant Classification was performed with 146 chromato

	Accuracy
Training	100.00% (146/146)
Validation	91.78% (134/146)

Parent Folder: C:\Users\anna\Documents

Archive Name: LDA_Summary_Report

Figure 2: Export classification feature data from Investigator to Investigator ML.

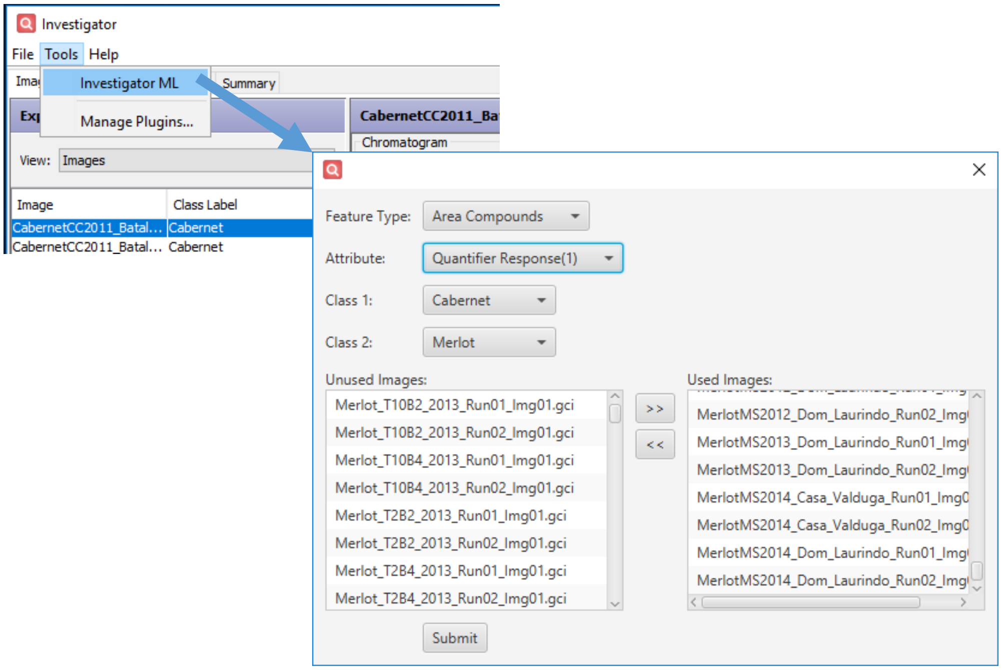


Image	Class Label
CabernetCC2011_Batal…	Cabernet
CabernetCC2011_Batal…	Cabernet

Feature Type: Area Compounds

Attribute: Quantifier Response(1)

Class 1: Cabernet

Class 2: Merlot

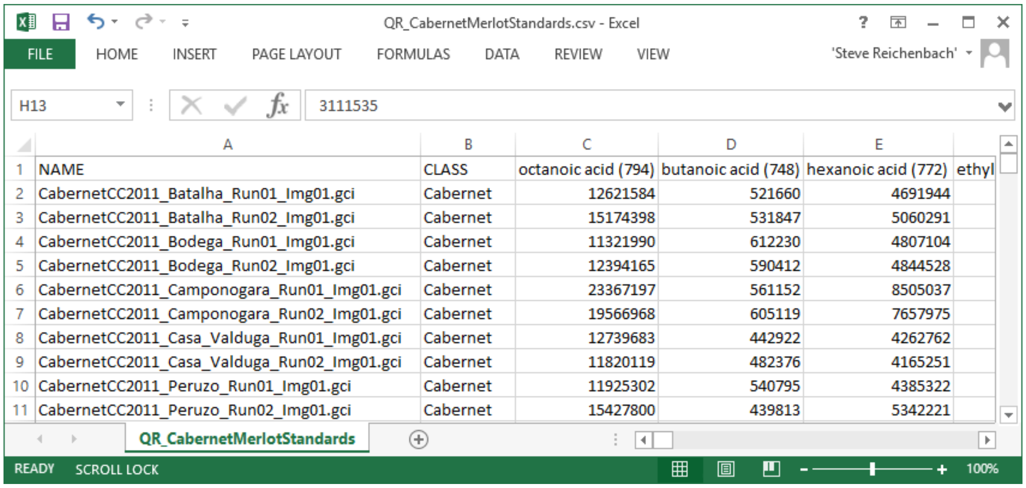
Unused Images:

Merlot_T10B2_2013_Run01_Img01.gci
Merlot_T10B2_2013_Run02_Img01.gci
Merlot_T10B4_2013_Run01_Img01.gci
Merlot_T10B4_2013_Run02_Img01.gci
Merlot_T2B2_2013_Run01_Img01.gci
Merlot_T2B2_2013_Run02_Img01.gci
Merlot_T2B4_2013_Run01_Img01.gci
Merlot_T2B4_2013_Run02_Img01.gci

Used Images:

MerlotMS2012_Dom_Laurindo_Run02_Img
MerlotMS2013_Dom_Laurindo_Run01_Img
MerlotMS2013_Dom_Laurindo_Run02_Img
MerlotMS2014_Casa_Valduga_Run01_Img0
MerlotMS2014_Casa_Valduga_Run02_Img0
MerlotMS2014_Dom_Laurindo_Run01_Img
MerlotMS2014_Dom_Laurindo_Run02_Img

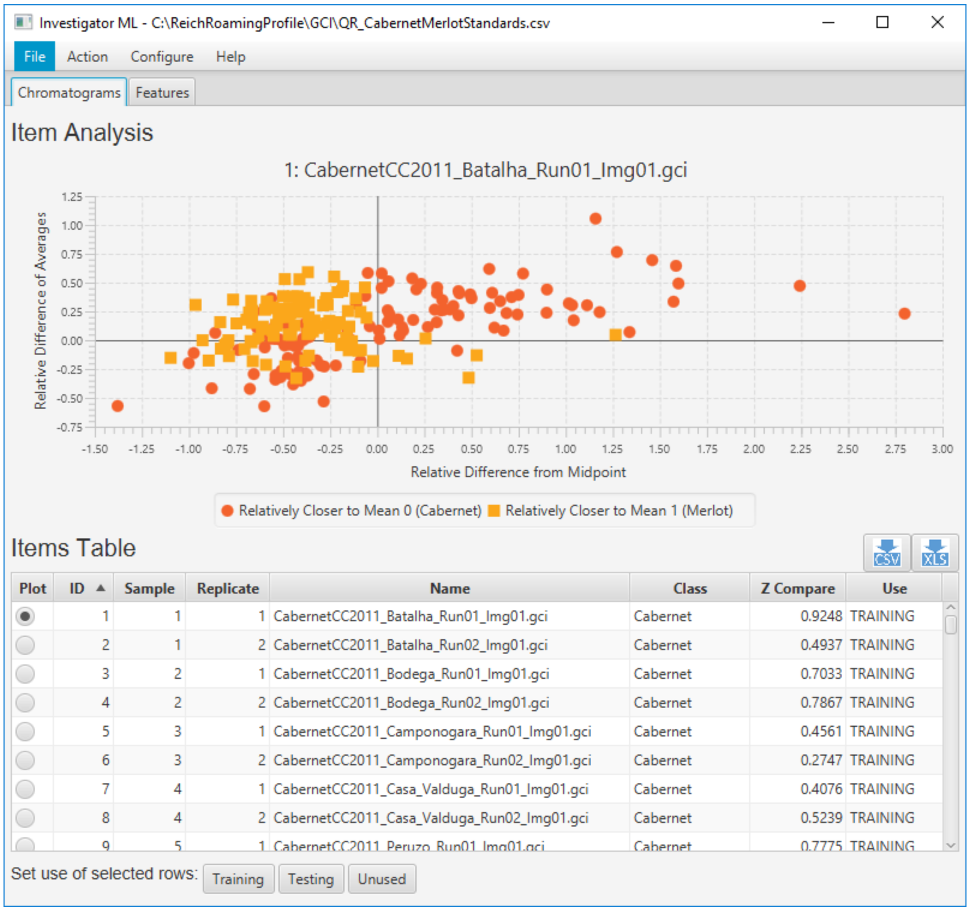
Figure 3: CSV-format data shown in Microsoft Excel®.



3111535

	A	B	C	D	E	
1	NAME	CLASS	octanoic acid (794)	butanoic acid (748)	hexanoic acid (772)	ethyl
2	CabernetCC2011_Batalha_Run01_Img01.gci	Cabernet	12621584	521660	4691944	
3	CabernetCC2011_Batalha_Run02_Img01.gci	Cabernet	15174398	531847	5060291	
4	CabernetCC2011_Bodega_Run01_Img01.gci	Cabernet	11321990	612230	4807104	
5	CabernetCC2011_Bodega_Run02_Img01.gci	Cabernet	12394165	590412	4844528	
6	CabernetCC2011_Camponogara_Run01_Img01.gci	Cabernet	23367197	561152	8505037	
7	CabernetCC2011_Camponogara_Run02_Img01.gci	Cabernet	19566968	605119	7657975	
8	CabernetCC2011_Casa_Valduga_Run01_Img01.gci	Cabernet	12739683	442922	4262762	
9	CabernetCC2011_Casa_Valduga_Run02_Img01.gci	Cabernet	11820119	482376	4165251	
10	CabernetCC2011_Peruzo_Run01_Img01.gci	Cabernet	11925302	540795	4385322	
11	CabernetCC2011_Peruzo_Run02_Img01.gci	Cabernet	15427800	439813	5342221	

Figure 4: Chromatograms Tab with Items Table and Item Analysis Graph.

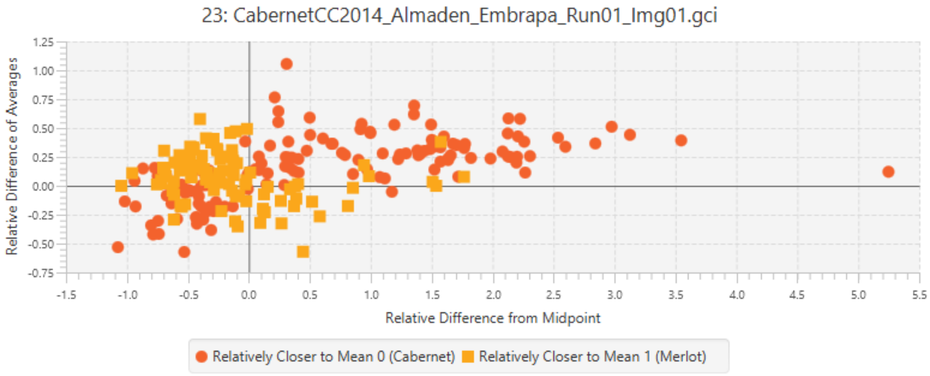


Investigator ML - C:\ReichRoamingProfile\GCI\QR_CabernetMerlotStandards.csv
File
Action
Configure
Help
Chromatograms
Features
Item Analysis
1: CabernetCC2011_Batalha_Run01_Img01.gci
Relative Difference of Averages
1.25
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.50
-1.25
-1.00
-0.75
-0.50
-0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
Relative Difference from Midpoint
● Relatively Closer to Mean 0 (Cabernet) ■ Relatively Closer to Mean 1 (Merlot)
Items Table
CSV
XLS
Set use of selected rows:
Training
Testing
Unused

Plot	ID ▲	Sample	Replicate	Name	Class	Z Compare	Use
●	1	1	1	CabernetCC2011_Batalha_Run01_Img01.gci	Cabernet	0.9248	TRAINING
○	2	1	2	CabernetCC2011_Batalha_Run02_Img01.gci	Cabernet	0.4937	TRAINING
○	3	2	1	CabernetCC2011_Bodega_Run01_Img01.gci	Cabernet	0.7033	TRAINING
○	4	2	2	CabernetCC2011_Bodega_Run02_Img01.gci	Cabernet	0.7867	TRAINING
○	5	3	1	CabernetCC2011_Camponogara_Run01_Img01.gci	Cabernet	0.4561	TRAINING
○	6	3	2	CabernetCC2011_Camponogara_Run02_Img01.gci	Cabernet	0.2747	TRAINING
○	7	4	1	CabernetCC2011_Casa_Valduga_Run01_Img01.gci	Cabernet	0.4076	TRAINING
○	8	4	2	CabernetCC2011_Casa_Valduga_Run02_Img01.gci	Cabernet	0.5239	TRAINING
○	9	5	1	CabernetCC2011_Peruzo_Run01_Img01.gci	Cabernet	0.7775	TRAINING

Figure 5: Item Analysis Graphs.

Example A: Feature values in the 1st and 3rd quadrants are indicative of Class 0.



Example B: Feature values in the 2nd and 4th quadrants are indicative of Class 1.

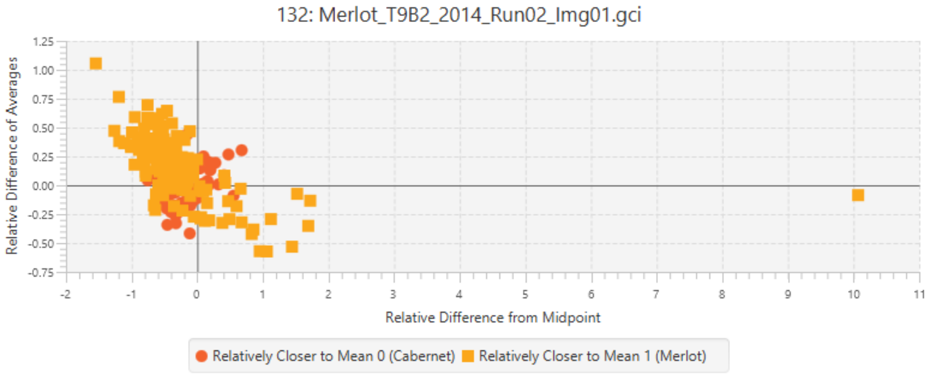
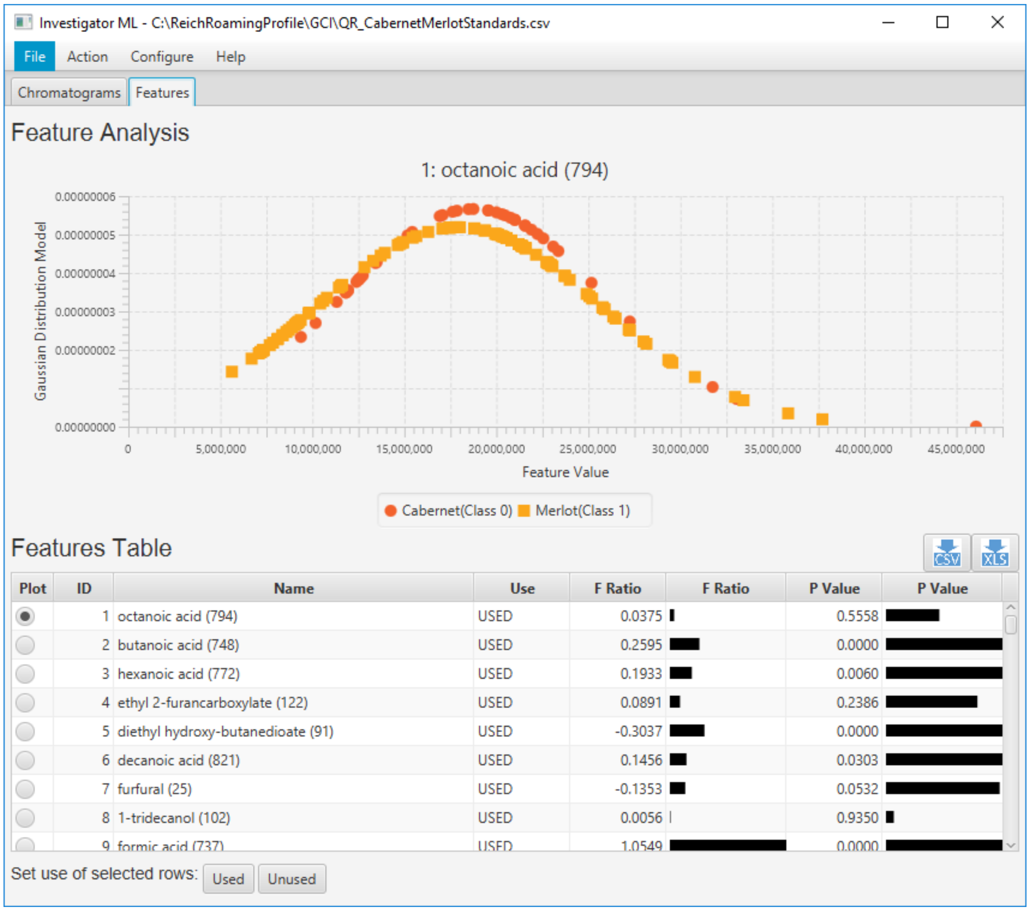


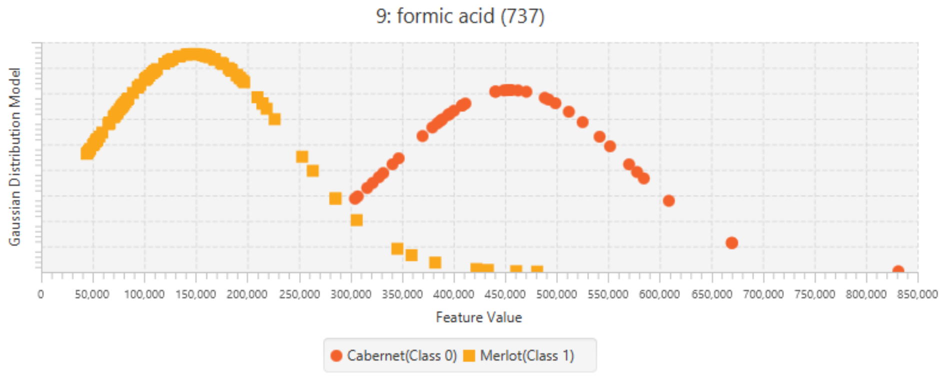
Figure 6: Features Tab with Features Table and Feature Analysis Graph.



Plot	ID	Name	Use	F Ratio	F Ratio	P Value	P Value
●	1	octanoic acid (794)	USED	0.0375	▬	0.5558	▬
○	2	butanoic acid (748)	USED	0.2595	▬	0.0000	▬
○	3	hexanoic acid (772)	USED	0.1933	▬	0.0060	▬
○	4	ethyl 2-furancarboxylate (122)	USED	0.0891	▬	0.2386	▬
○	5	diethyl hydroxy-butanedioate (91)	USED	-0.3037	▬	0.0000	▬
○	6	decanoic acid (821)	USED	0.1456	▬	0.0303	▬
○	7	furfural (25)	USED	-0.1353	▬	0.0532	▬
○	8	1-tridecanol (102)	USED	0.0056	▬	0.9350	▬
○	9	formic acid (737)	USED	1.0549	▬	0.0000	▬

Figure 7: Feature Analysis Graphs.

Example A: Classes have different feature value statistics.



Example B: Classes have similar feature value statistics.

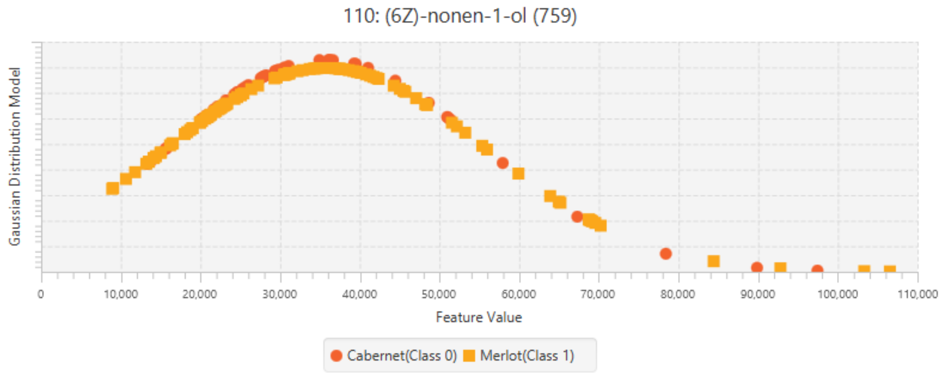
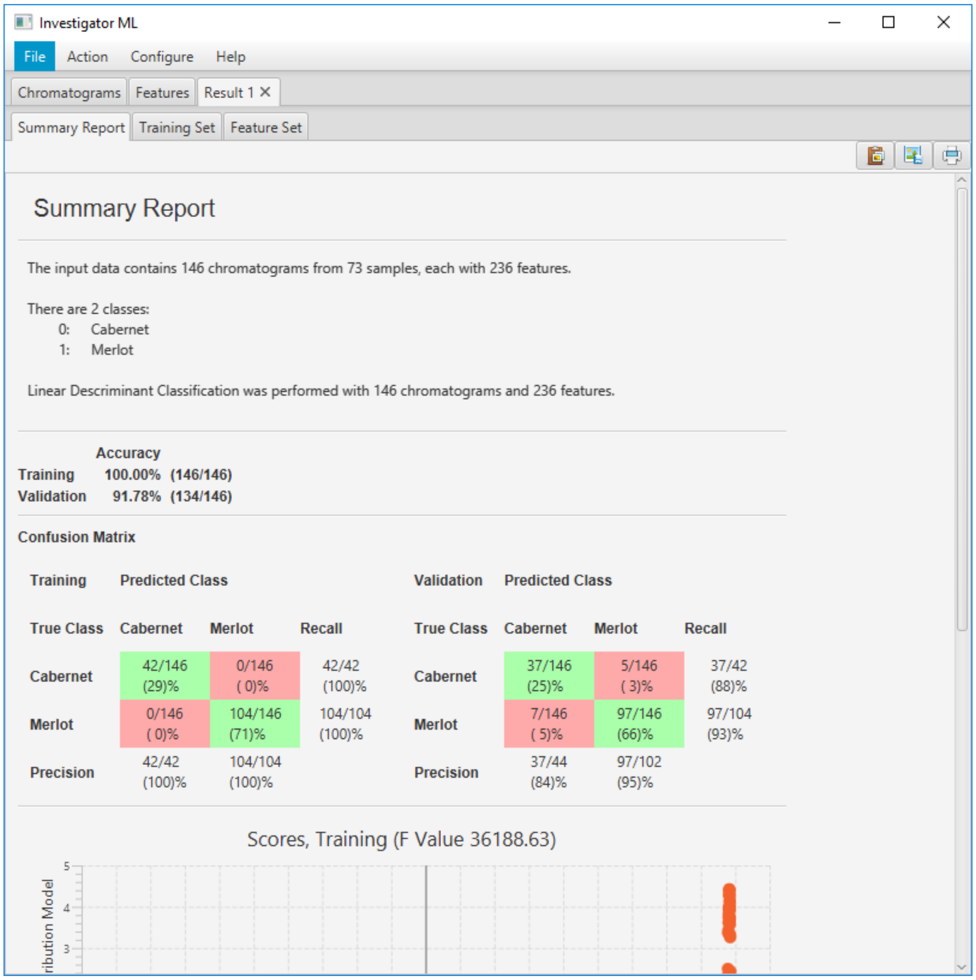


Figure 8: Summary Report for Linear Discriminant Analysis.



Summary Report

The input data contains 146 chromatograms from 73 samples, each with 236 features.

There are 2 classes:
0: Cabernet
1: Merlot

Linear Descriminant Classification was performed with 146 chromatograms and 236 features.

	Accuracy
Training	**100.00% (146/146)**
Validation	**91.78% (134/146)**

Confusion Matrix

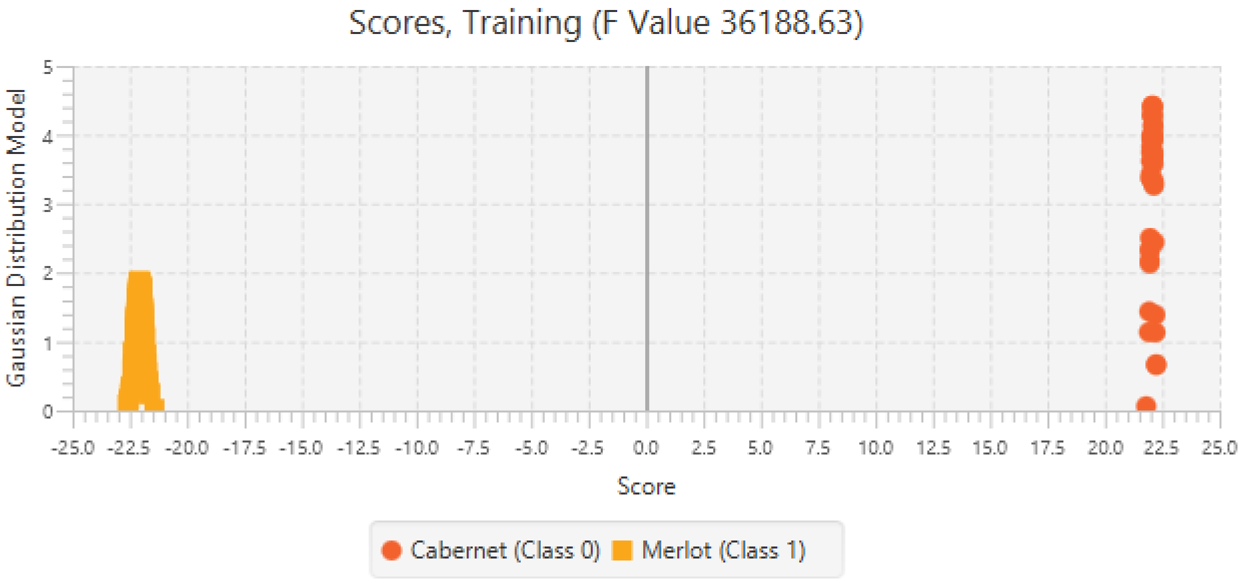
Training	**Predicted Class**		
True Class	**Cabernet**	**Merlot**	**Recall**
Cabernet	42/146 (29)%	0/146 (0)%	42/42 (100)%
Merlot	0/146 (0)%	104/146 (71)%	104/104 (100)%
Precision	42/42 (100)%	104/104 (100)%	

Validation	**Predicted Class**		
True Class	**Cabernet**	**Merlot**	**Recall**
Cabernet	37/146 (25)%	5/146 (3)%	37/42 (88)%
Merlot	7/146 (5)%	97/146 (66)%	97/104 (93)%
Precision	37/44 (84)%	97/102 (95)%	

Scores, Training (F Value 36188.63)

Figure 9: LDA Scores Graphs.

A) Training Scores Graph.



B) Validation Scores Graph.

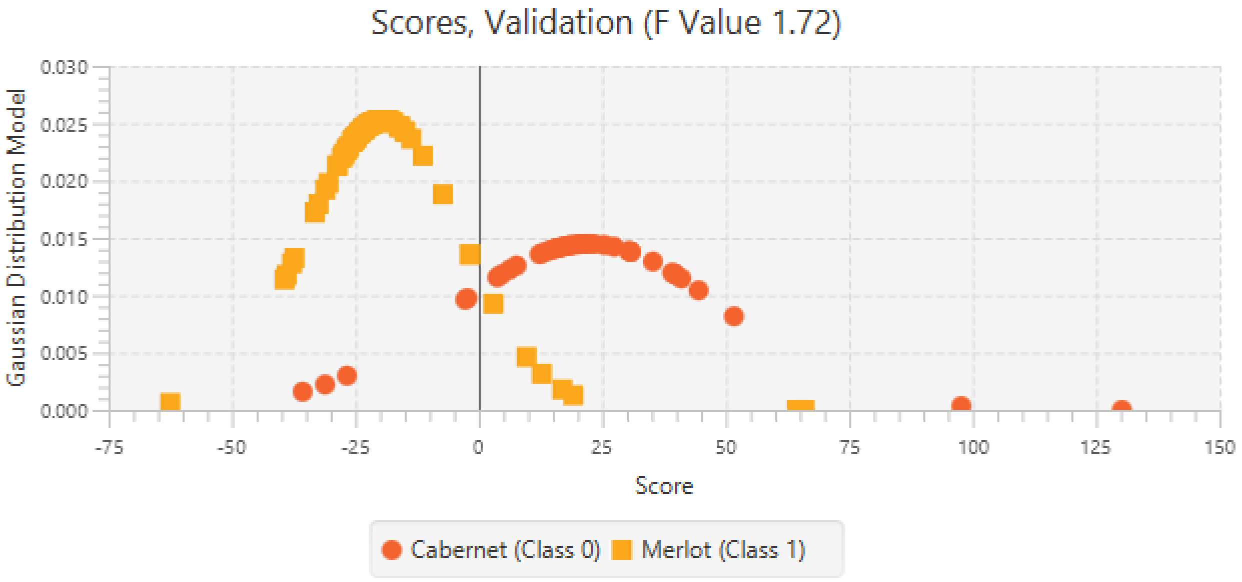
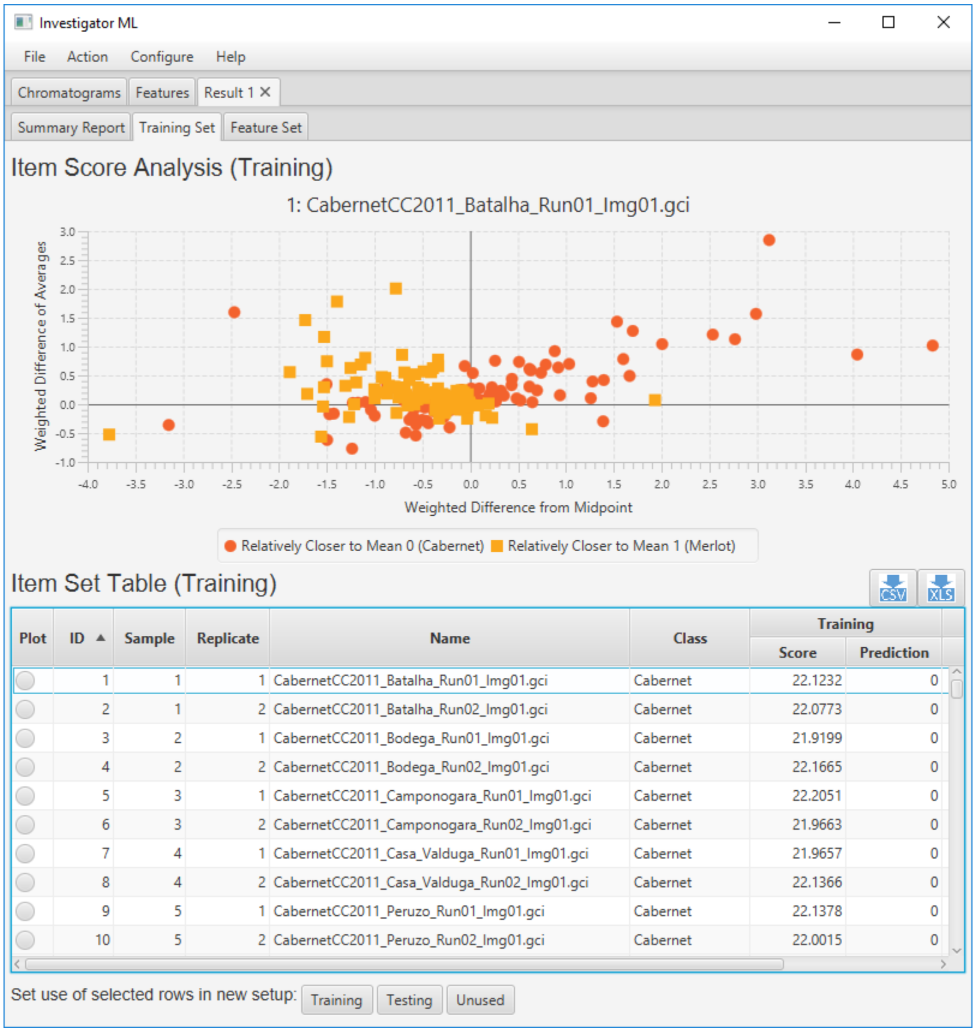


Figure 10: LDA Item Set Tab with Item Set Table and Item Score Graph.



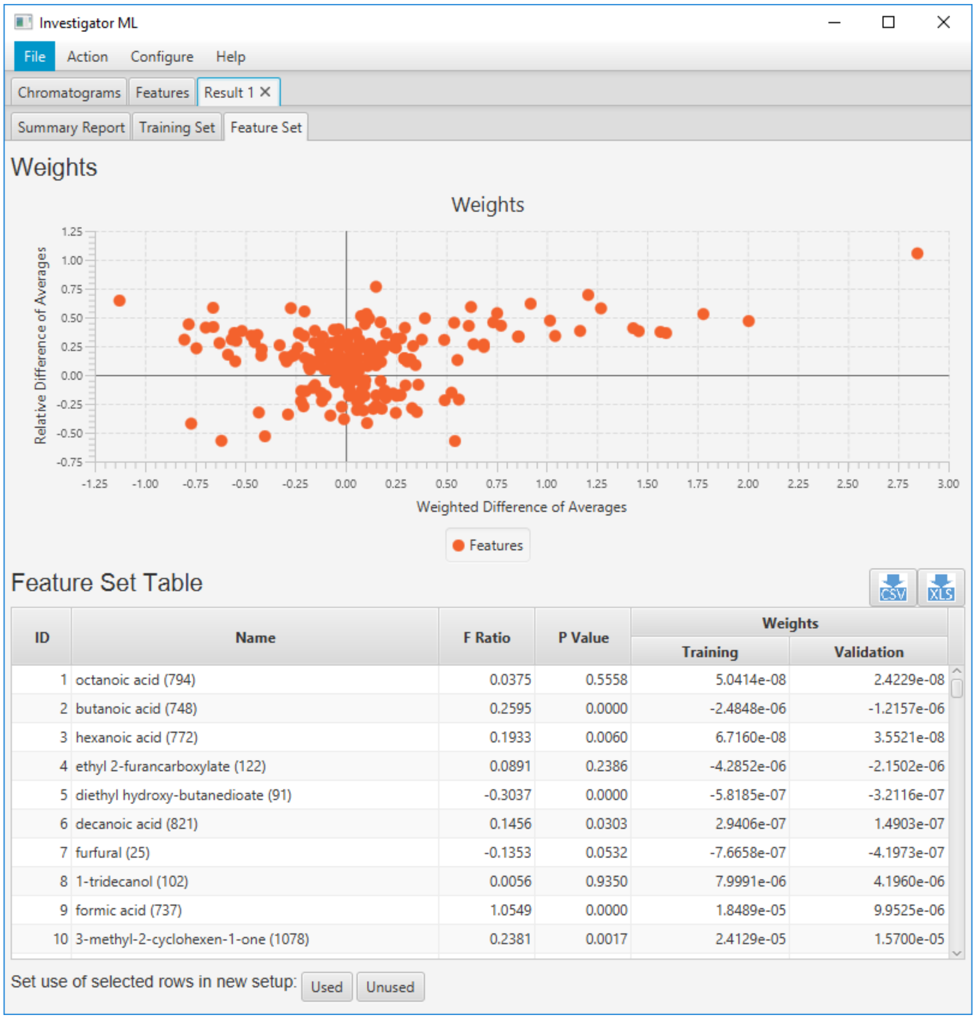
Item Score Analysis (Training)

Item Set Table (Training)

Plot	ID ▲	Sample	Replicate	Name	Class	Training	
						Score	Prediction
○	1	1	1	CabernetCC2011_Batalha_Run01_Img01.gci	Cabernet	22.1232	0
○	2	1	2	CabernetCC2011_Batalha_Run02_Img01.gci	Cabernet	22.0773	0
○	3	2	1	CabernetCC2011_Bodega_Run01_Img01.gci	Cabernet	21.9199	0
○	4	2	2	CabernetCC2011_Bodega_Run02_Img01.gci	Cabernet	22.1665	0
○	5	3	1	CabernetCC2011_Camponogara_Run01_Img01.gci	Cabernet	22.2051	0
○	6	3	2	CabernetCC2011_Camponogara_Run02_Img01.gci	Cabernet	21.9663	0
○	7	4	1	CabernetCC2011_Casa_Valduga_Run01_Img01.gci	Cabernet	21.9657	0
○	8	4	2	CabernetCC2011_Casa_Valduga_Run02_Img01.gci	Cabernet	22.1366	0
○	9	5	1	CabernetCC2011_Peruzo_Run01_Img01.gci	Cabernet	22.1378	0
○	10	5	2	CabernetCC2011_Peruzo_Run02_Img01.gci	Cabernet	22.0015	0

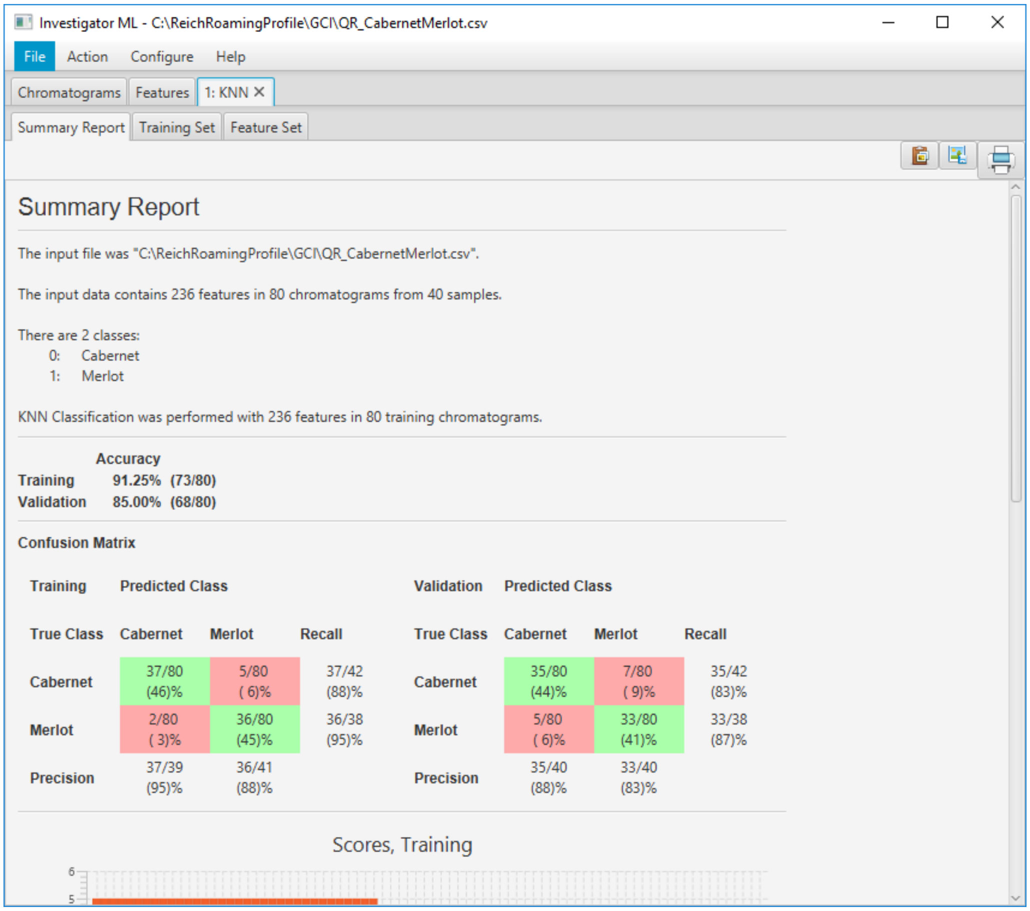
Set use of selected rows in new setup:

Figure 11: LDA Feature Set Tab with Feature Set Table and Weights Graph.



ID	Name	F Ratio	P Value	Weights	
				Training	Validation
1	octanoic acid (794)	0.0375	0.5558	5.0414e-08	2.4229e-08
2	butanoic acid (748)	0.2595	0.0000	-2.4848e-06	-1.2157e-06
3	hexanoic acid (772)	0.1933	0.0060	6.7160e-08	3.5521e-08
4	ethyl 2-furancarboxylate (122)	0.0891	0.2386	-4.2852e-06	-2.1502e-06
5	diethyl hydroxy-butanedioate (91)	-0.3037	0.0000	-5.8185e-07	-3.2116e-07
6	decanoic acid (821)	0.1456	0.0303	2.9406e-07	1.4903e-07
7	furfural (25)	-0.1353	0.0532	-7.6658e-07	-4.1973e-07
8	1-tridecanol (102)	0.0056	0.9350	7.9991e-06	4.1960e-06
9	formic acid (737)	1.0549	0.0000	1.8489e-05	9.9525e-06
10	3-methyl-2-cyclohexen-1-one (1078)	0.2381	0.0017	2.4129e-05	1.5700e-05

Figure 12: Summary Report for *k*-Nearest Neighbor Classification.



Summary Report

The input file was "C:\ReichRoamingProfile\GCI\QR_CabernetMerlot.csv".

The input data contains 236 features in 80 chromatograms from 40 samples.

There are 2 classes:
0: Cabernet
1: Merlot

KNN Classification was performed with 236 features in 80 training chromatograms.

	Accuracy
Training	**91.25% (73/80)**
Validation	**85.00% (68/80)**

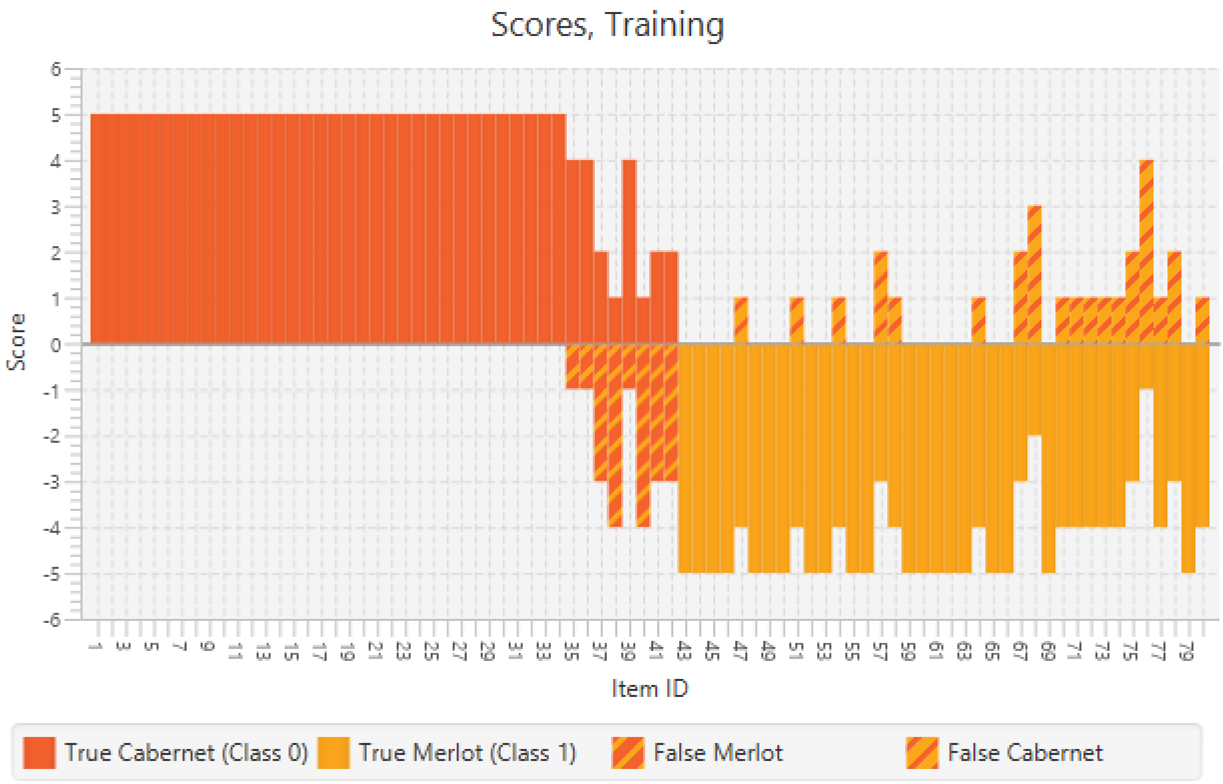
Confusion Matrix

Training	**Predicted Class**		
True Class	**Cabernet**	**Merlot**	**Recall**
Cabernet	37/80 (46)%	5/80 (6)%	37/42 (88)%
Merlot	2/80 (3)%	36/80 (45)%	36/38 (95)%
Precision	37/39 (95)%	36/41 (88)%	

Validation	**Predicted Class**		
True Class	**Cabernet**	**Merlot**	**Recall**
Cabernet	35/80 (44)%	7/80 (9)%	35/42 (83)%
Merlot	5/80 (6)%	33/80 (41)%	33/38 (87)%
Precision	35/40 (88)%	33/40 (83)%	

Figure 13: KNN Scores Graphs.

A) Training Scores Graph.



B) Validation Scores Graph.

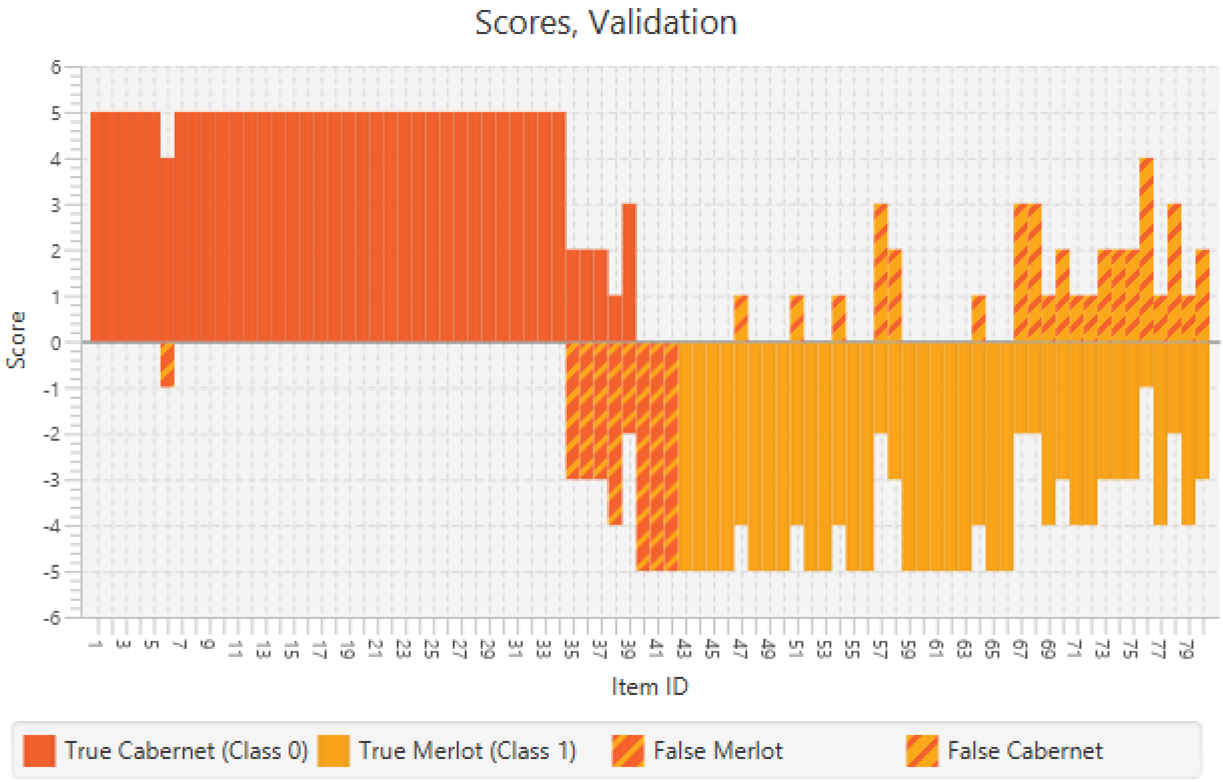
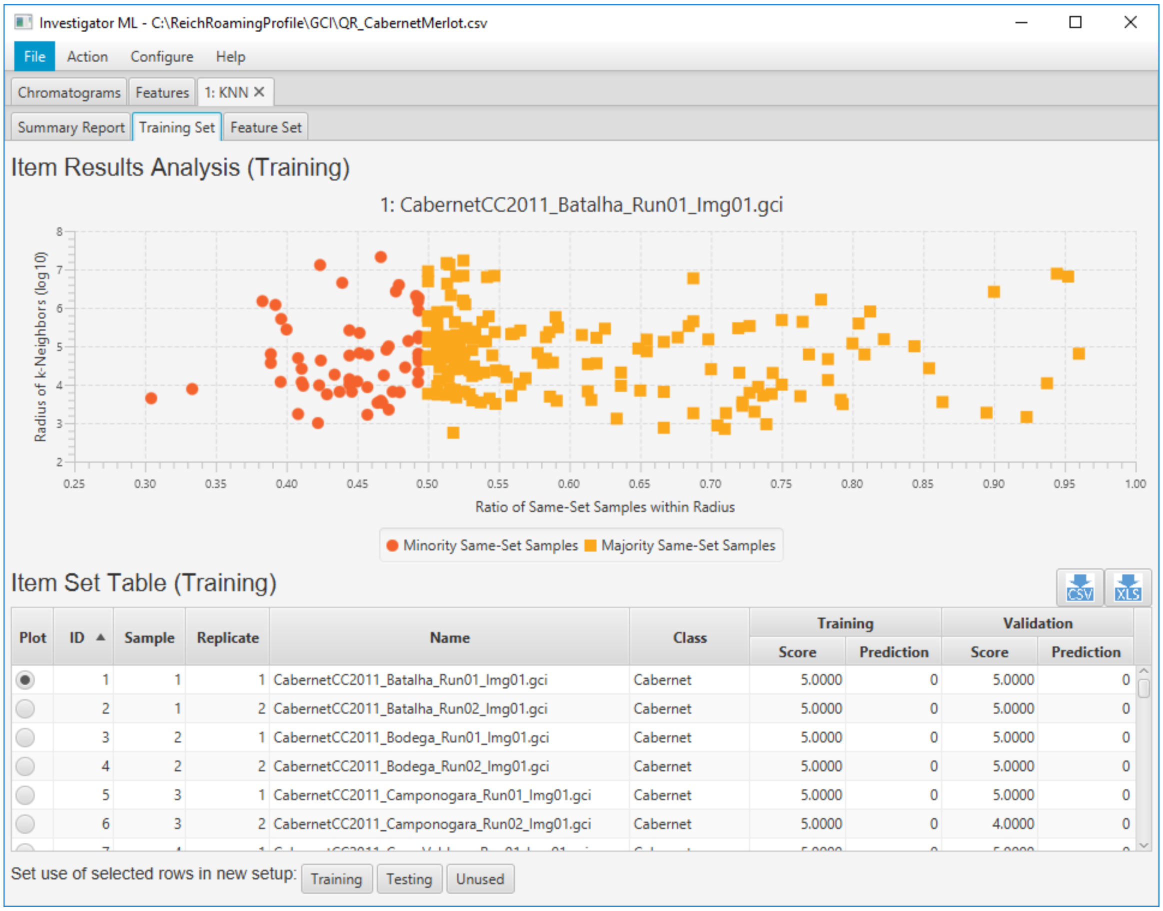
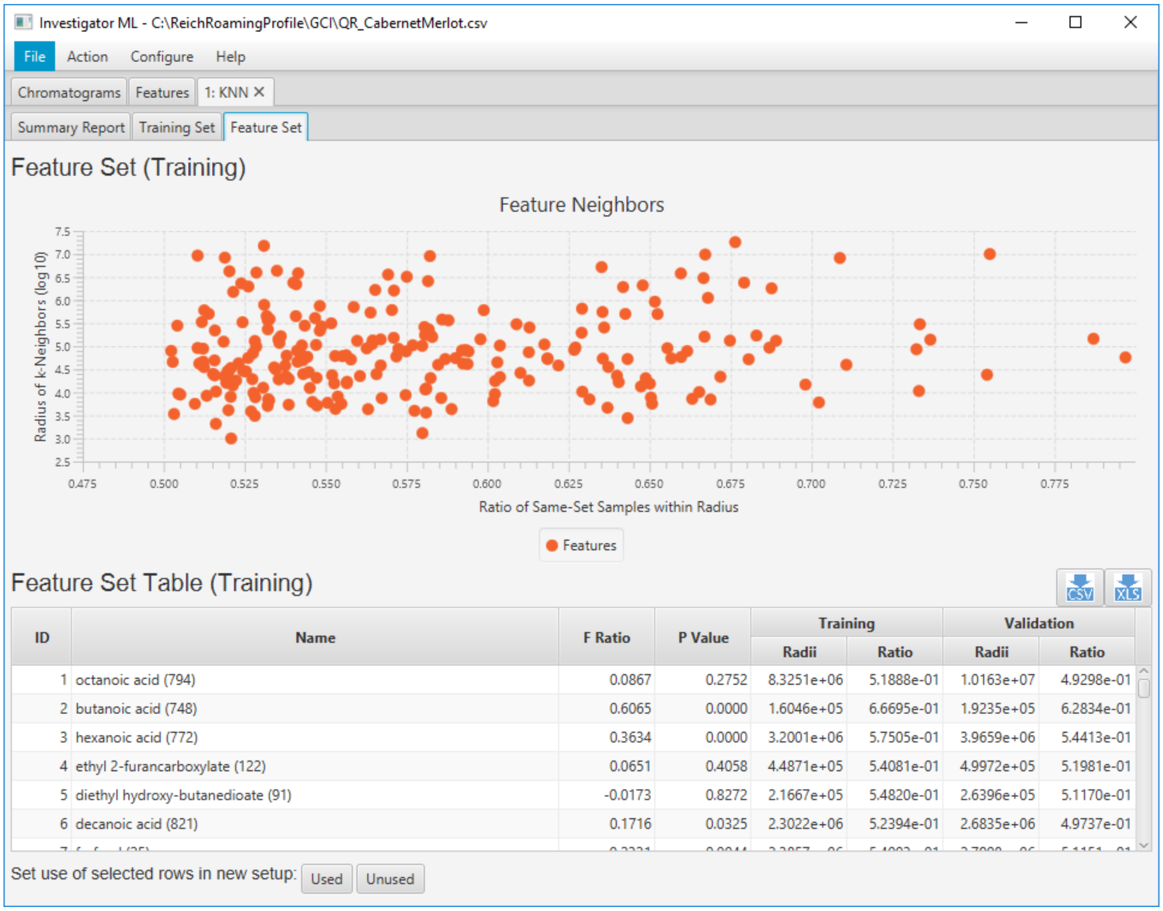


Figure 14: KNN Item Set Tab with Item Set Table and Item Score Graph.


Investigator ML - C:\ReichRoamingProfile\GCI\QR_CabernetMerlot.csv
File
Action
Configure
Help
Chromatograms
Features
1: KNN ✕
Summary Report
Training Set
Feature Set
Item Results Analysis (Training)
1: CabernetCC2011_Batalha_Run01_Img01.gci
Radius of k-Neighbors (log10)
2
3
4
5
6
7
8
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Ratio of Same-Set Samples within Radius
● Minority Same-Set Samples
■ Majority Same-Set Samples
Item Set Table (Training)
CSV
XLS

Plot	ID ▲	Sample	Replicate	Name	Class	Training		Validation	
						Score	Prediction	Score	Prediction
●	1	1	1	CabernetCC2011_Batalha_Run01_Img01.gci	Cabernet	5.0000	0	5.0000	0
○	2	1	2	CabernetCC2011_Batalha_Run02_Img01.gci	Cabernet	5.0000	0	5.0000	0
○	3	2	1	CabernetCC2011_Bodega_Run01_Img01.gci	Cabernet	5.0000	0	5.0000	0
○	4	2	2	CabernetCC2011_Bodega_Run02_Img01.gci	Cabernet	5.0000	0	5.0000	0
○	5	3	1	CabernetCC2011_Camponogara_Run01_Img01.gci	Cabernet	5.0000	0	5.0000	0
○	6	3	2	CabernetCC2011_Camponogara_Run02_Img01.gci	Cabernet	5.0000	0	4.0000	0

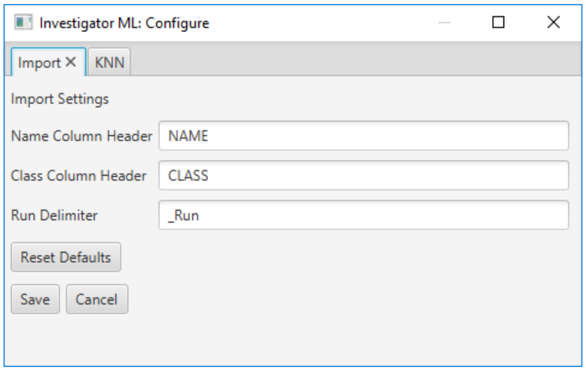
Figure 15: KNN Feature Set Tab with Feature Set Table and Weights Graph.



ID	Name	F Ratio	P Value	Training		Validation	
				Radii	Ratio	Radii	Ratio
1	octanoic acid (794)	0.0867	0.2752	8.3251e+06	5.1888e-01	1.0163e+07	4.9298e-01
2	butanoic acid (748)	0.6065	0.0000	1.6046e+05	6.6695e-01	1.9235e+05	6.2834e-01
3	hexanoic acid (772)	0.3634	0.0000	3.2001e+06	5.7505e-01	3.9659e+06	5.4413e-01
4	ethyl 2-furancarboxylate (122)	0.0651	0.4058	4.4871e+05	5.4081e-01	4.9972e+05	5.1981e-01
5	diethyl hydroxy-butanedioate (91)	-0.0173	0.8272	2.1667e+05	5.4820e-01	2.6396e+05	5.1170e-01
6	decanoic acid (821)	0.1716	0.0325	2.3022e+06	5.2394e-01	2.6835e+06	4.9737e-01

Figure 16: Configuration.

A) Import Configuration Tab.



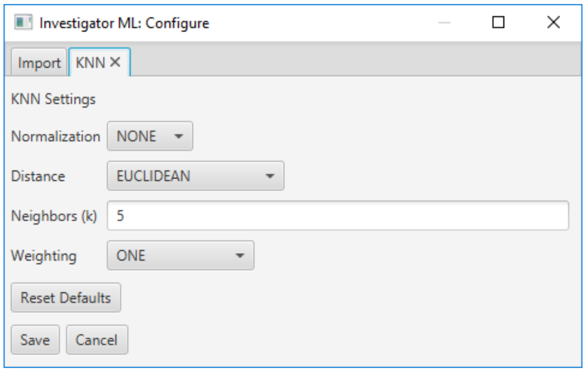
Import Settings

Name Column Header: NAME

Class Column Header: CLASS

Run Delimiter: _Run

B) KNN Configuration Tab.



KNN Settings

Normalization: NONE

Distance: EUCLIDEAN

Neighbors (k): 5

Weighting: ONE