**Orange Report**

For this project I have used two datasets:

1. The heart diseases dataset, which consists of 303 instances of heart diseases patient. The dataset provides measurements for age and cholesterol, and BP and Chest pain.
   1. Features:
      1. Age (numerical)
      2. Gender (numerical)
      3. Chest Pain (numerical)
      4. Cholesterol(numerical)
2. The Wine quality dataset, which consists of 4898 instances. The quality of the white variants of the Portuguese "Vinho Verde" wine. Wine samples are described with results of physicochemical tests that include information on acidity, sugar and alcohol content, density and other. This data set is related to its sister data set on the quality of red wines.
   1. Features:
      1. Quality (numerical)
      2. Fixed Acidity (numerical, target)
      3. Size (numerical)
      4. Volatile Acidity (numerical)
      5. Citric Acid (time)
      6. Residual Sugar (categorical)
      7. Density (categorical)

By:

Name: V Shashank

PRN: 20030121120

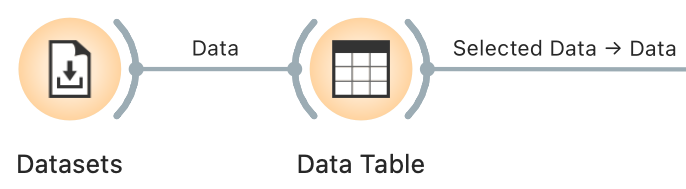
Group - 2

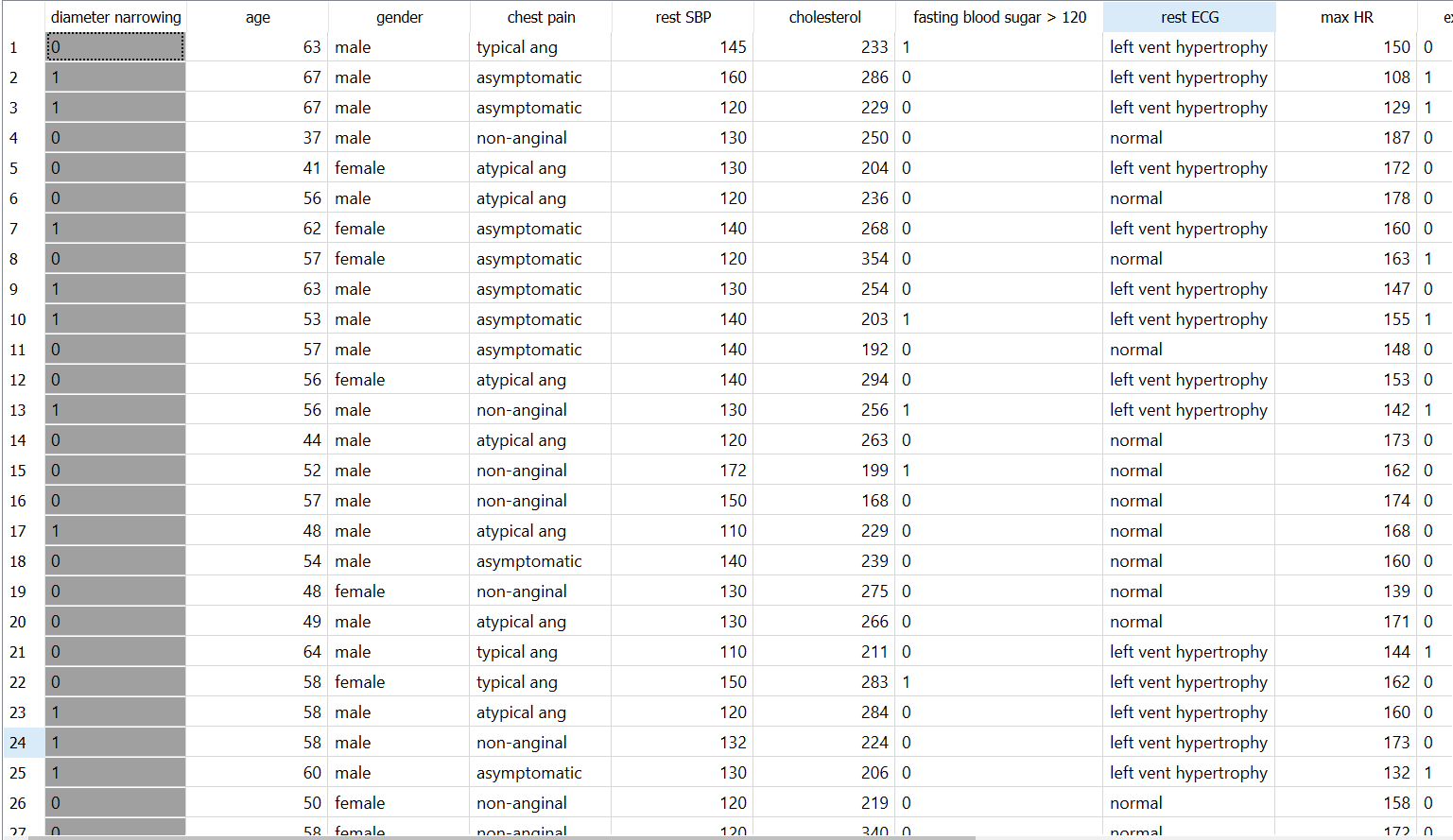
**Data Category**

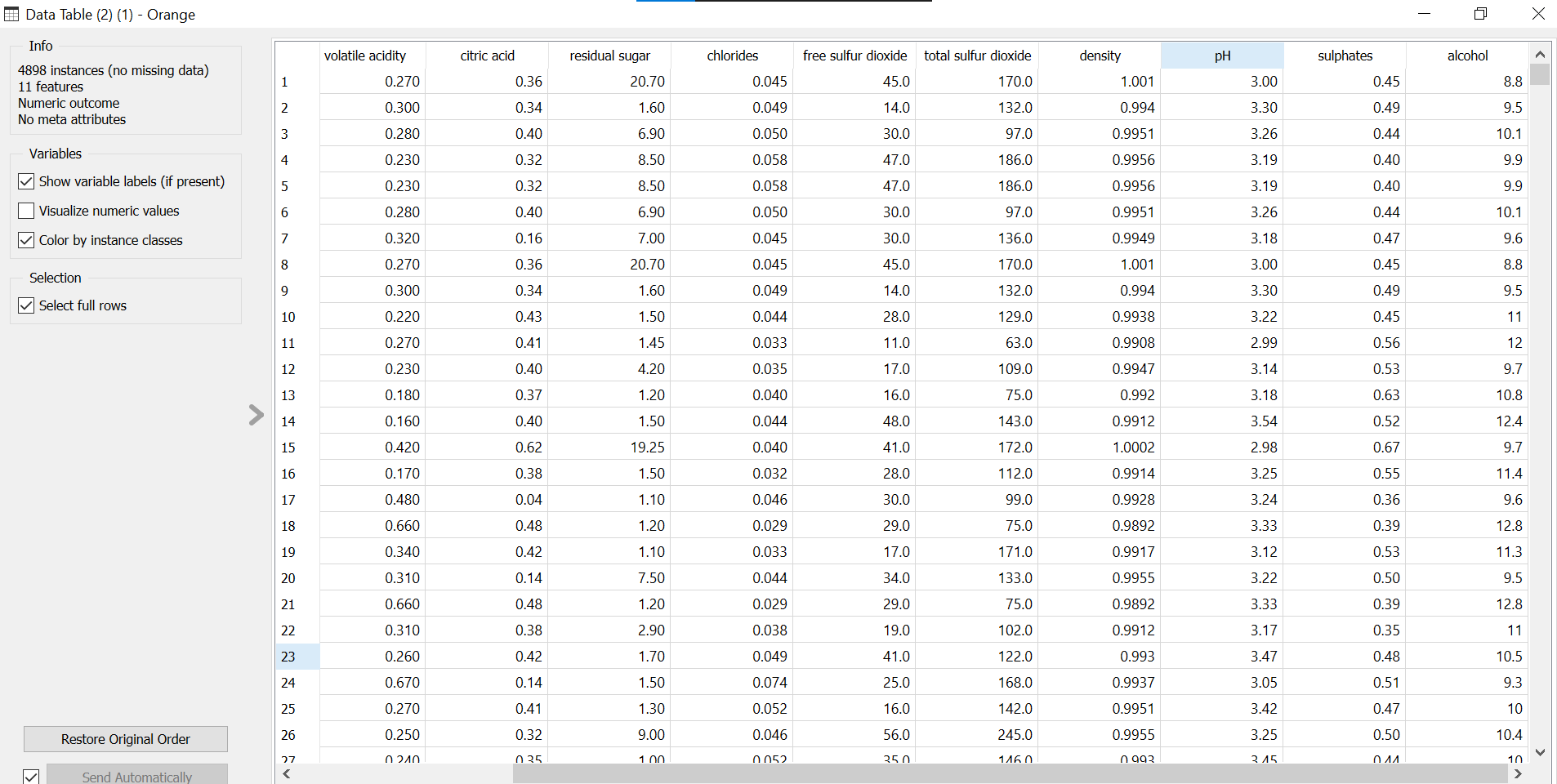
**1. Data Table**

Need of Data Table:

* Data table is needed to represent data instances as a set of 2D tables representing the independent variables and dependent variables, and the corresponding weights and meta attributes.







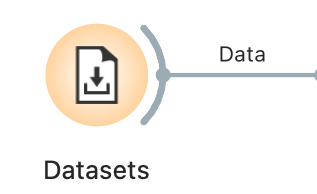
Interpretation:

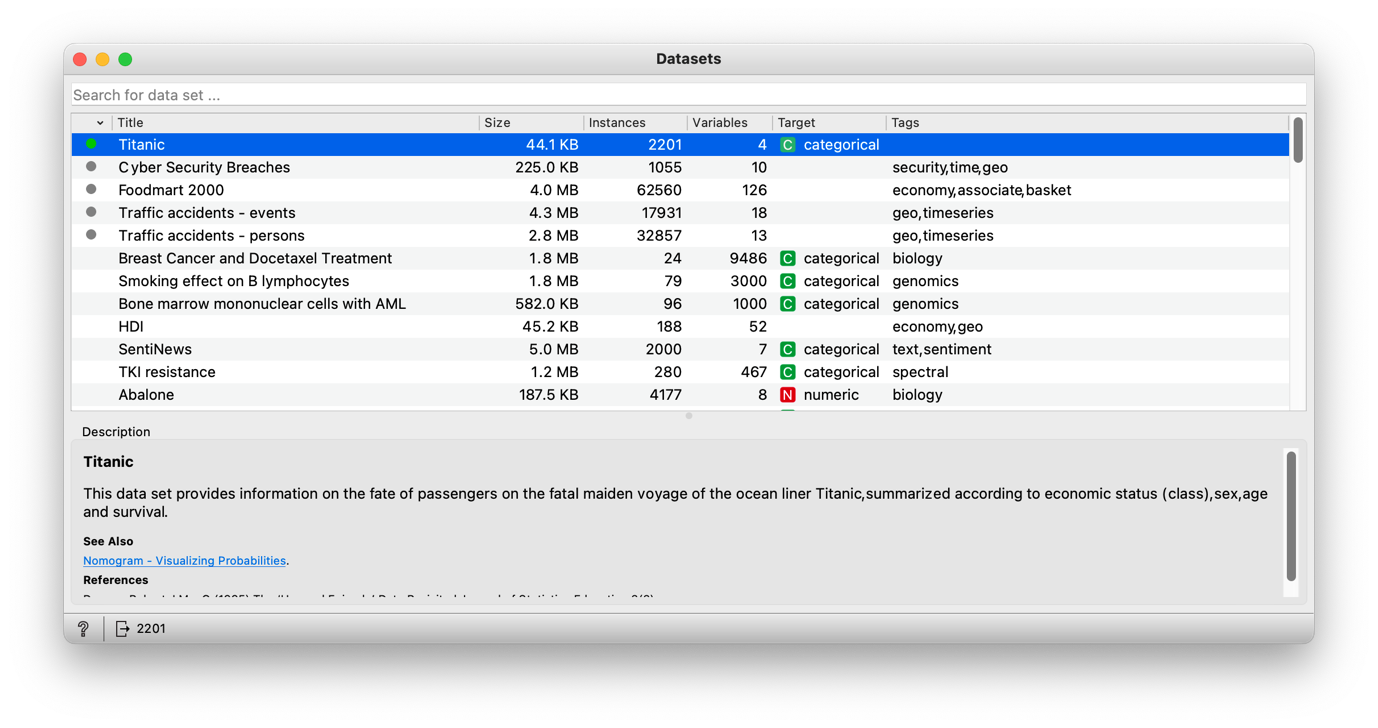
* Using the Data Table widget, we are able to see a tabular representation of our dataset.

**2. Datasets**

Need of Datasets:

* This widget is needed to retrieve a selected dataset from a server and send it to the output channel. The file is downloaded to the local memory and thus instantly available even without the internet connection.





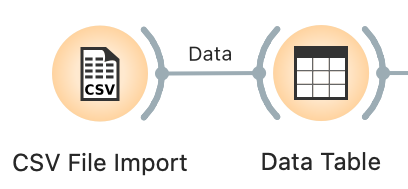
Interpretation:

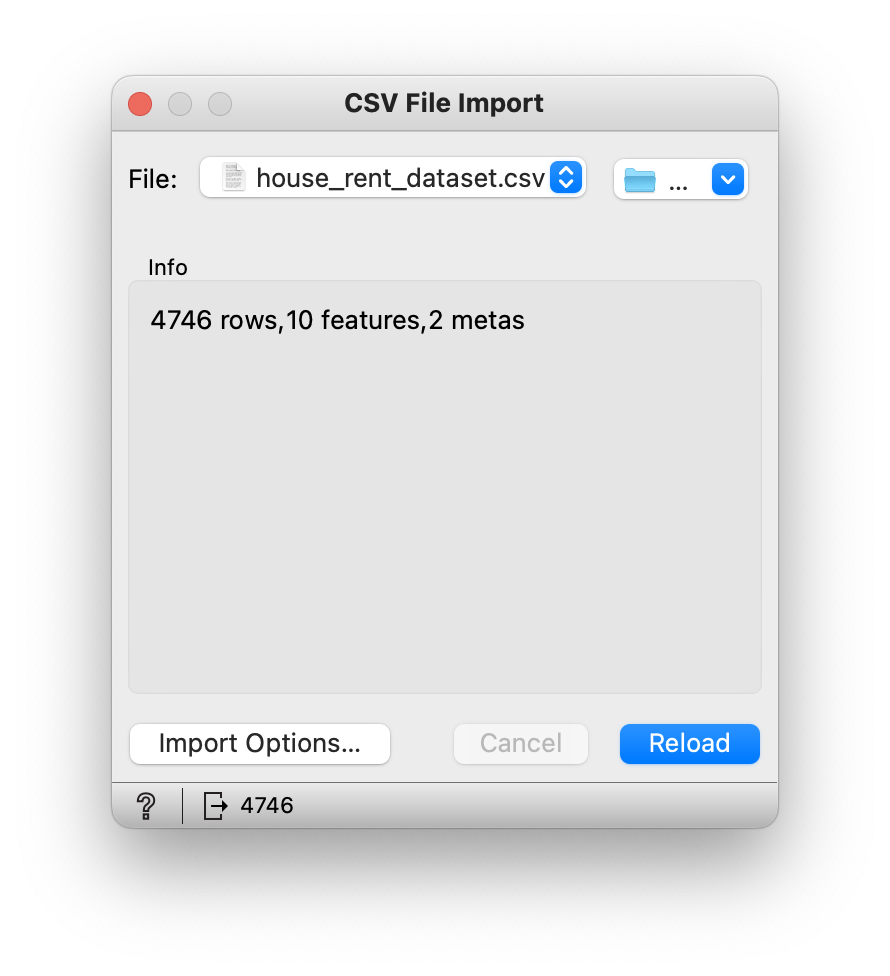
* Using the Datasets widget, we can import datasets from an online repository. We can also see a description of the selected dataset.

**3. CSV File Import**

Need of CSV File Import:

* We need this widget to import a dataset from a csv formatted file. It also gives an overview of the dataset like the number of instances (rows), variables (features or columns) and meta variables (special columns).





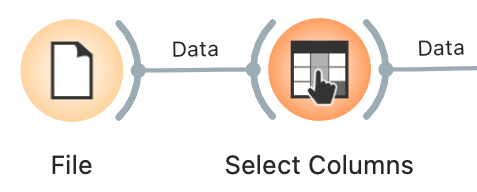
Interpretation:

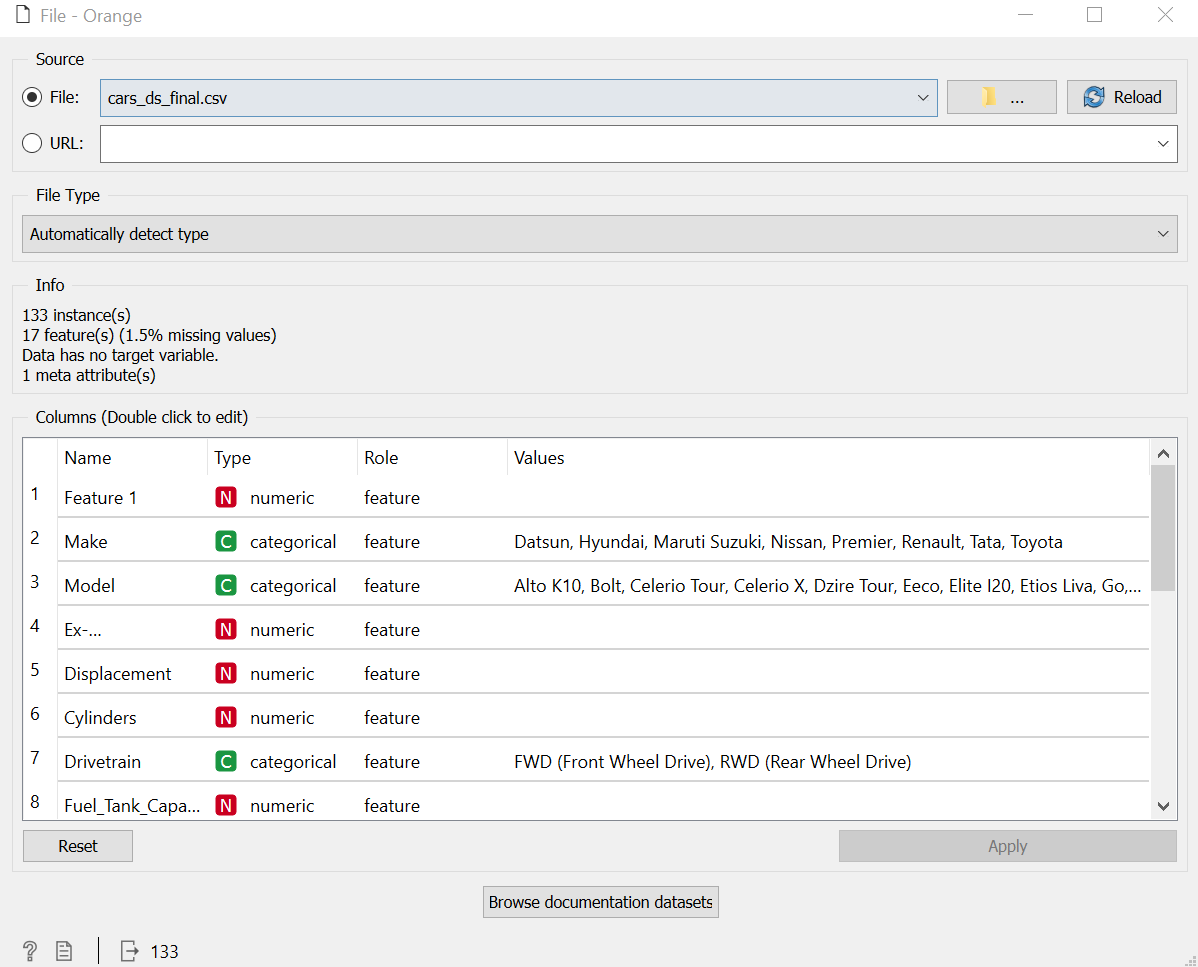
* Using CSV File Import widget, we can import datasets from csv-formatted files.

**4. File**

Need of File:

* This widget is needed to read input data file and sends the dataset to its output channel. The history of most recently opened files is maintained in the widget. The widget also comes with pre-installed datasets.





Interpretation:

* Using File widget, we can import datasets from local machine or use the pre-installed datasets.

**5. Paint Data**

Need of Paint Data:

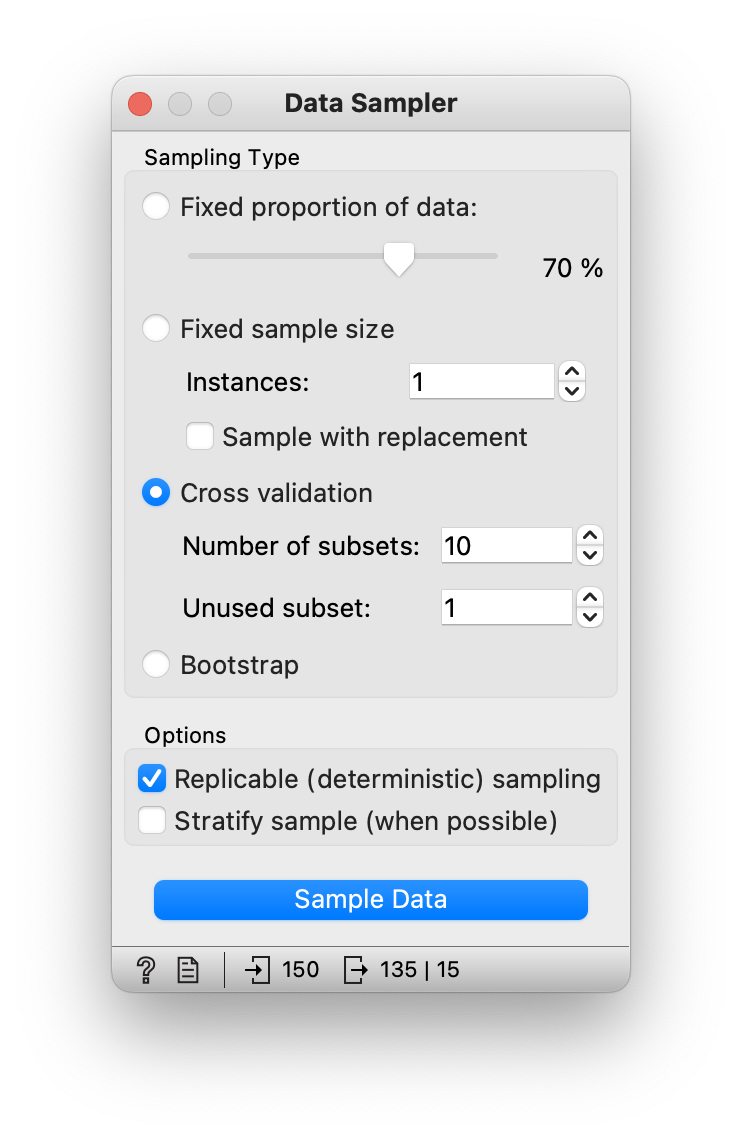
* To create new data instances

**Transform Category**

**1. Data Sampler**

Need of Data Sampler:

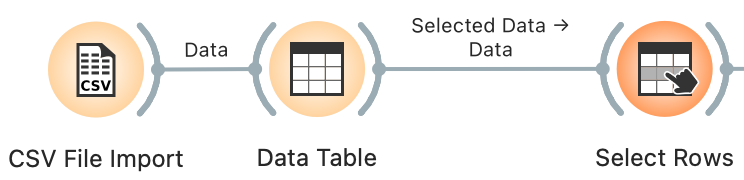
* This widget is needed to implement several data sampling methods to give an output of training and testing datasets which could be used for training and testing machine learning models.

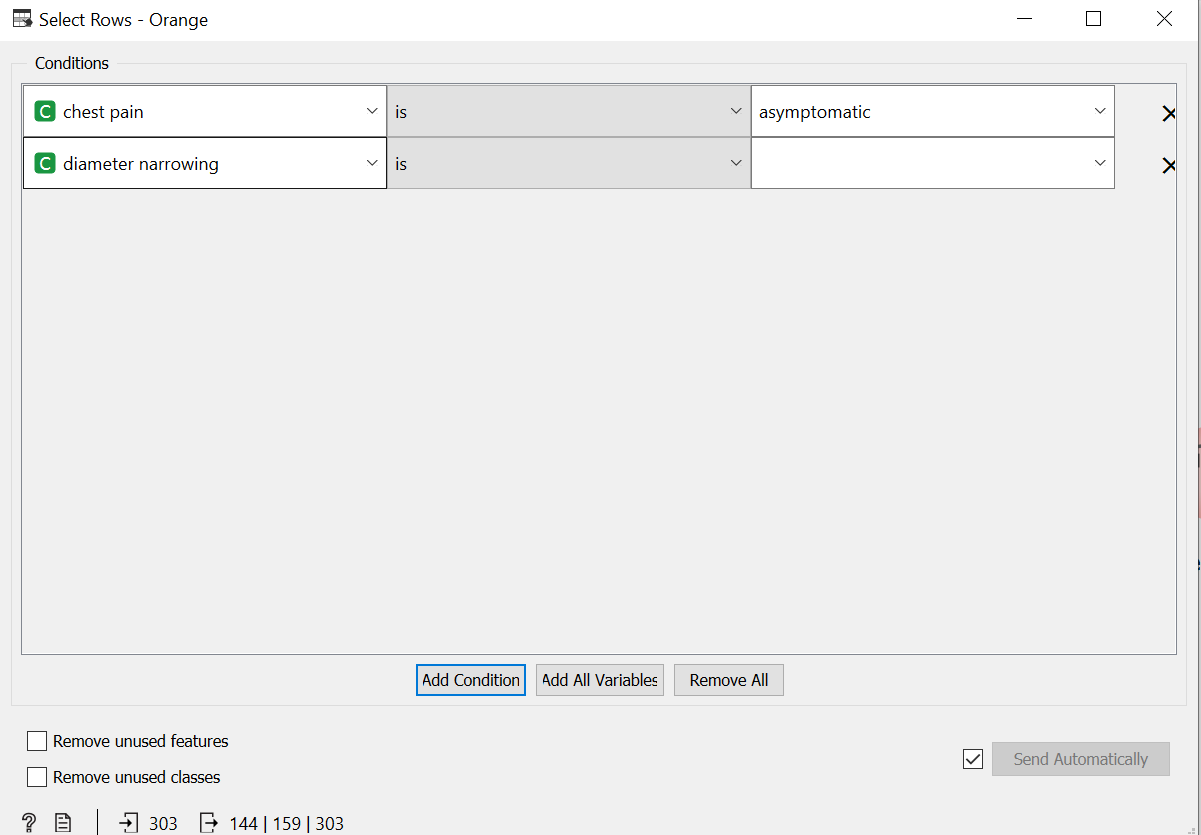


**2. Select Rows**

Need of Select Rows:

* This widget is needed when the user wants a subset of a dataset, based on user-defined conditions. Instances that match the selection rule are placed in the output *Matching Data* channel.





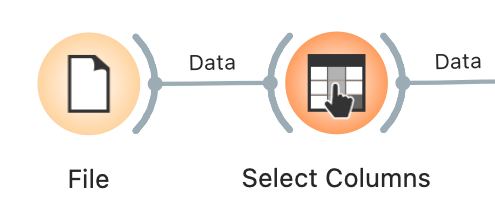
Interpretation:

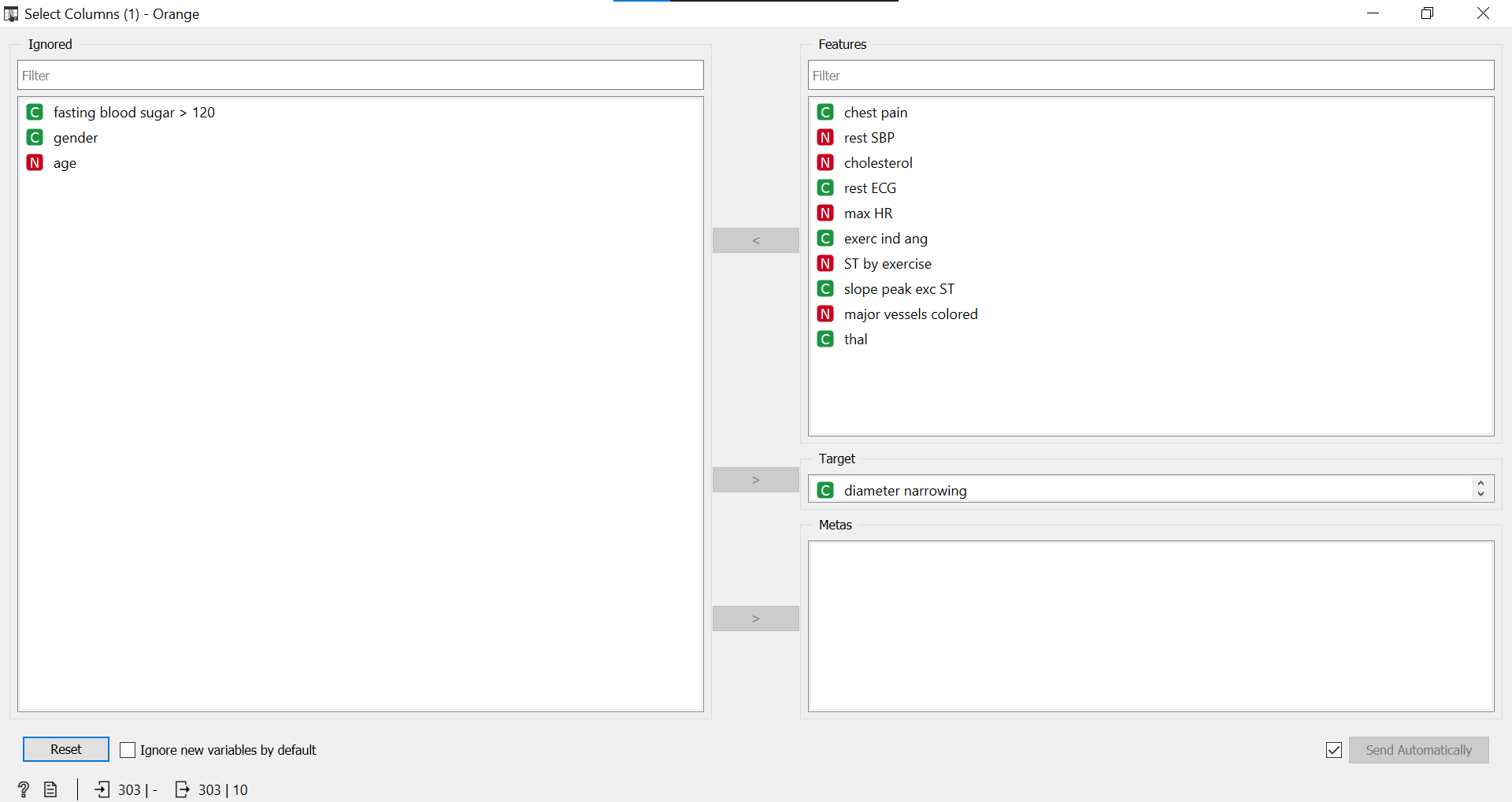
* Using Select Rows, we can fetch specific data instances using user-defined conditions.

**3. Select Columns**

Need of Select Columns:

* It is needed to get rid of unnecessary features and to select the target feature to improve the quality of the dataset.





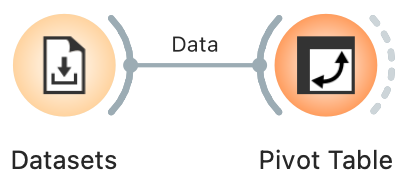
Interpretation:

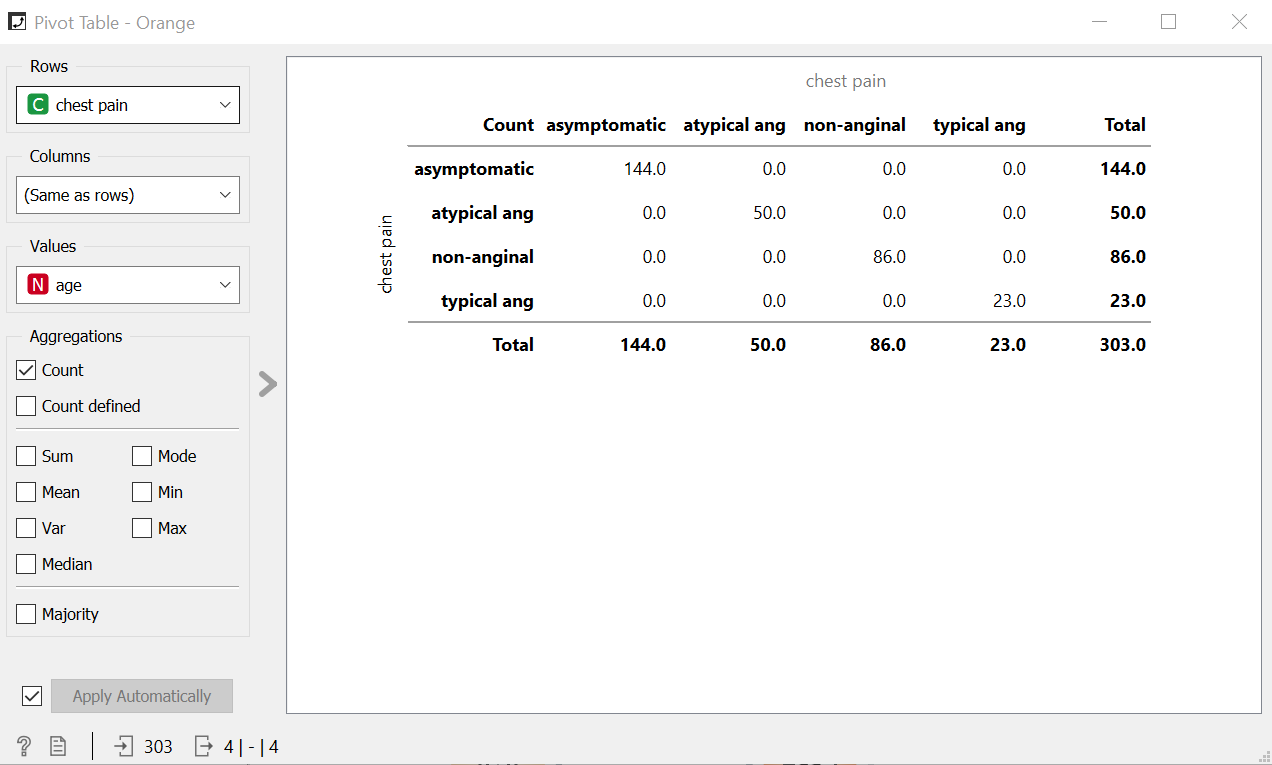
* Using Select Columns, we can ignore features based on the user’s requirement and select a target feature.

**4. Pivot Table**

Need of Pivot Table:

* We get all the statistical details such as mean, median, sum, etc. of the selected feature in a tabular form.

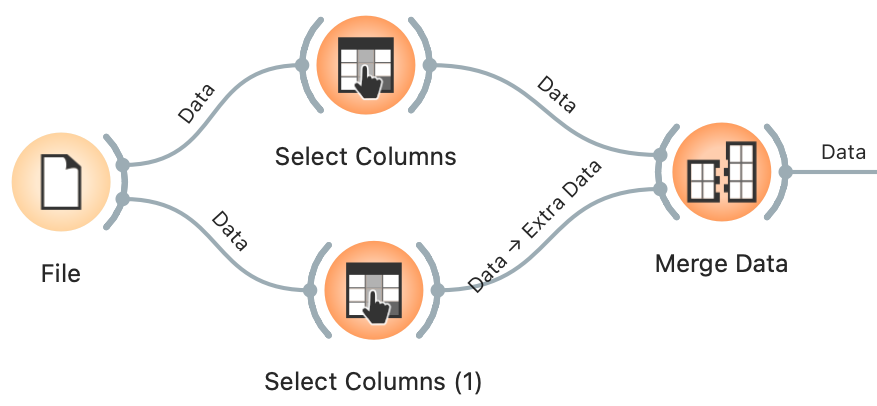




Interpretation:

* Using the pivot table, we get a statistical summary (mean, mode, median, etc.) of selected columns and rows.

**5. Merge Data**

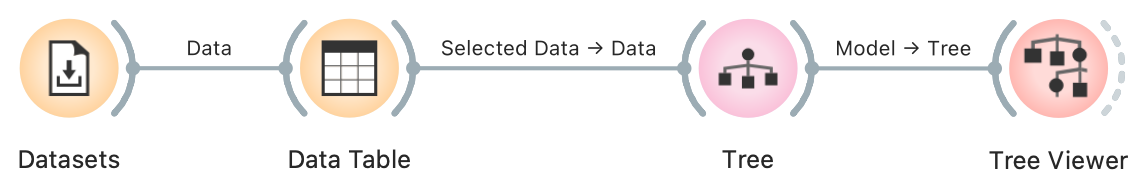


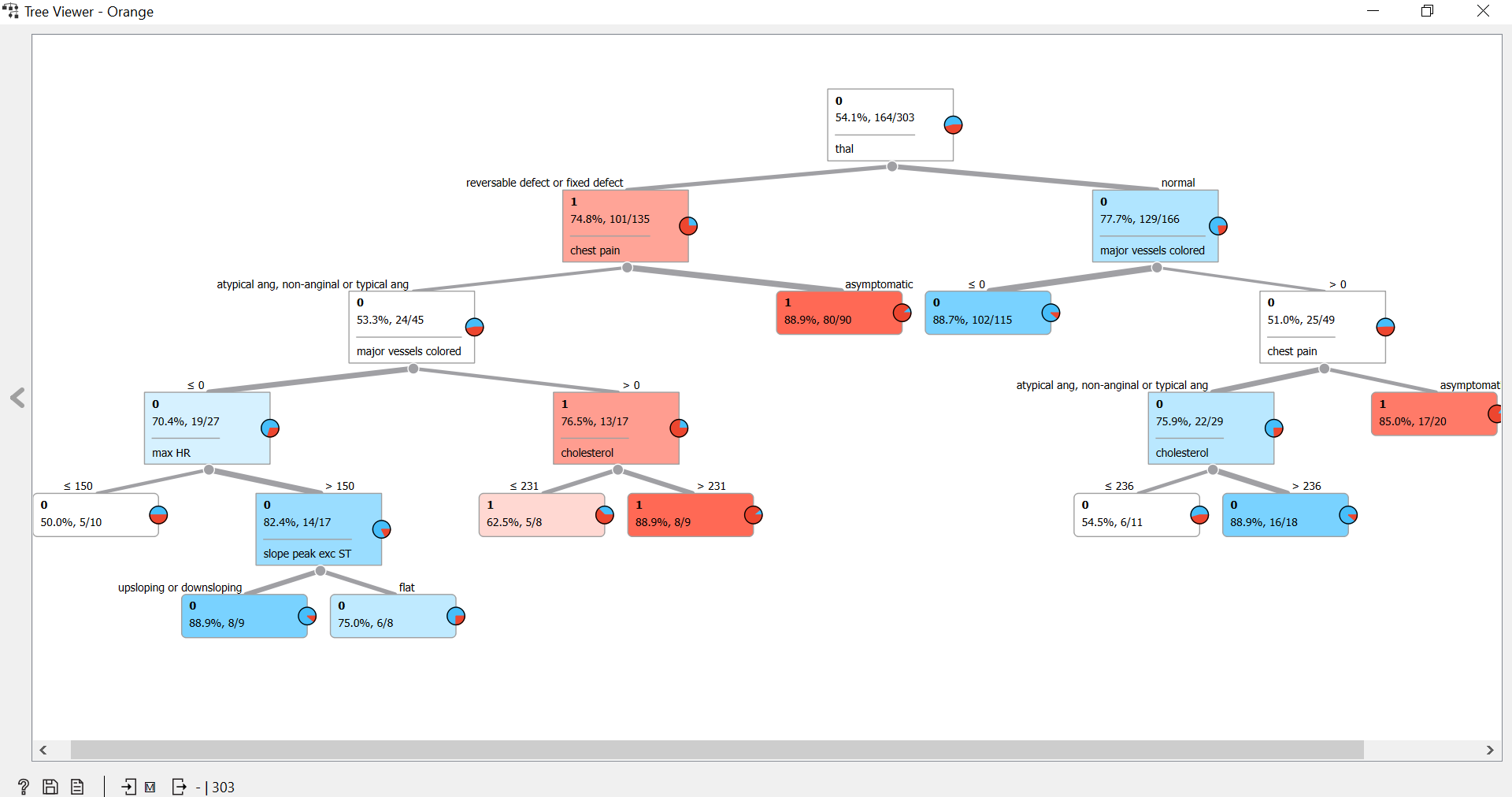
**Visualize Category**

**1. Tree Viewer**

Need of Tree Viewer:

* This widget is needed for 2D visualization of classification and regression trees.
* For the explorative data analysis, the user has to select the node and instruct the widget to output the data associated with the node.





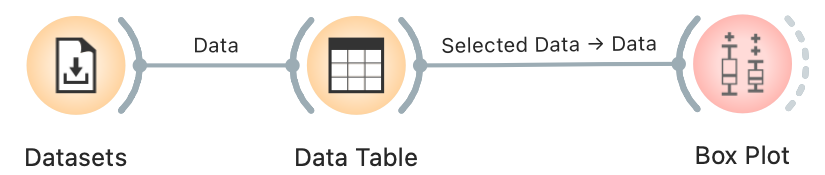
Interpretation:

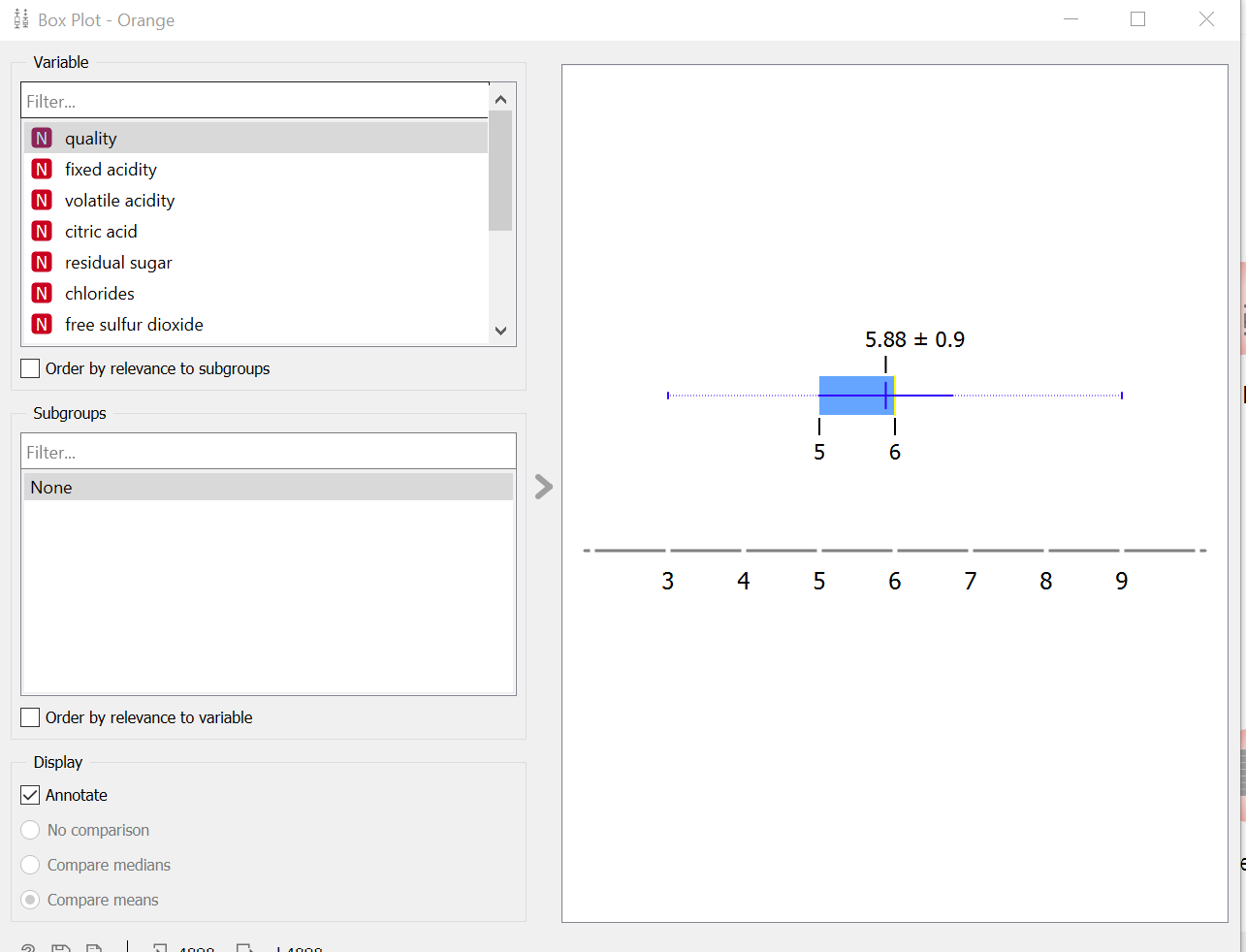
* Starting from the root node, you go to the next nodes and the edges tell you which subset you are looking at. Once you reach the leaf node, the node tells you the predicted outcome.

**2. Box Plot**

Need of Box Plot:

* This widget shows the distributions of attribute values. It is a good practice to check any new data with this widget to quickly discover any anomalies, such as duplicated values, outliers, and alike. Bars can be selected – for example, values for categorical data or the quantile range for numeric data.





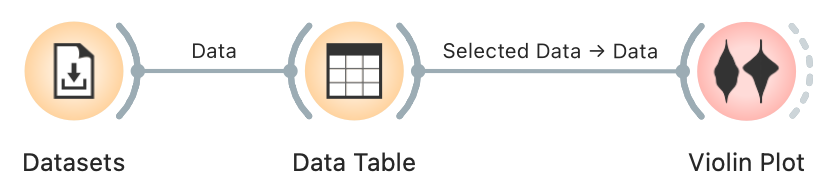
Interpretation:

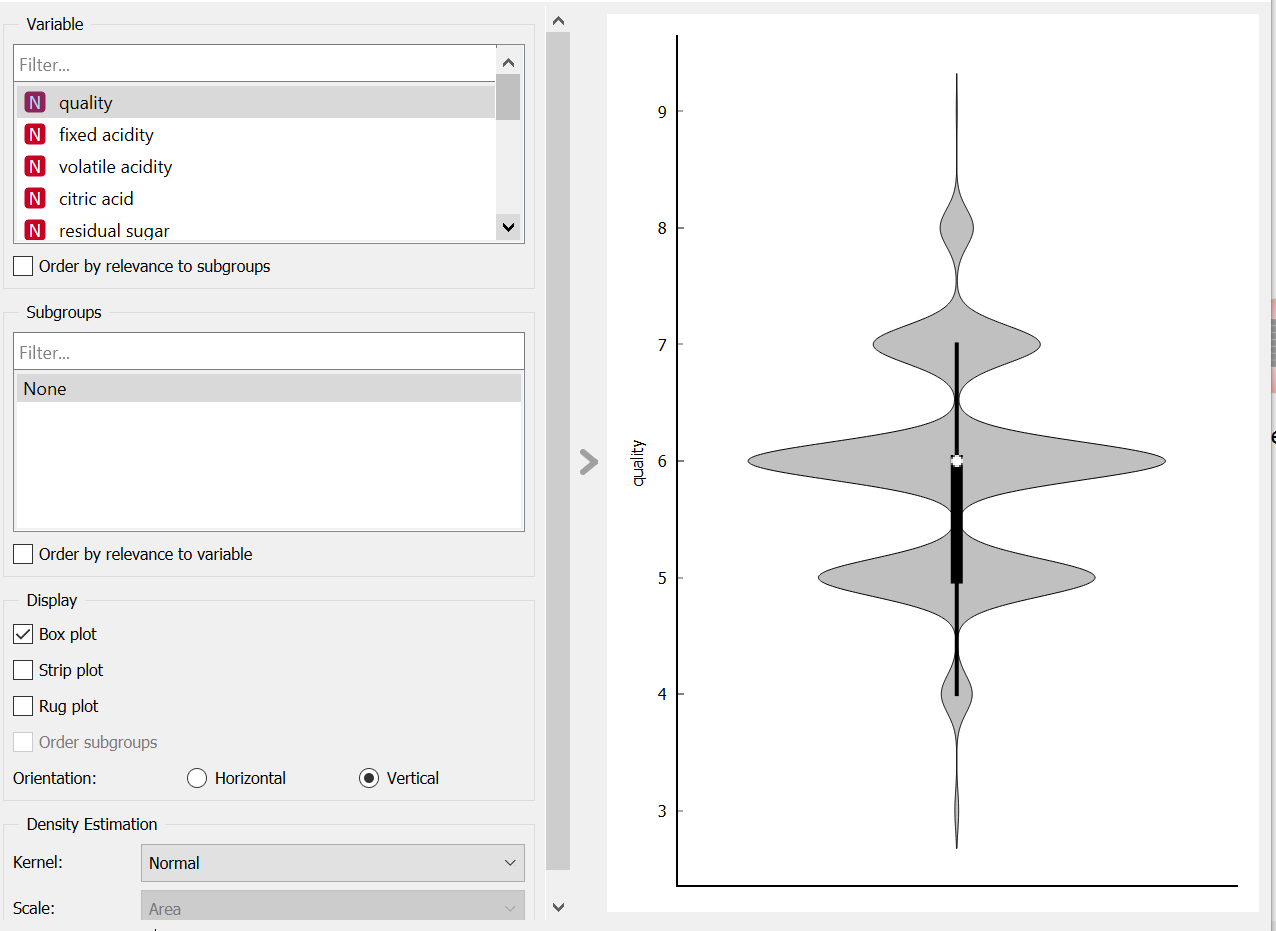
* Using the Box Plot widget, we can see the range of values for each feature along with the mean, median and standard deviation.
* We can also see the range of values for each feature that are overlapping.

**3. Violin Plot**

Need of Violin Plot:

* Violin plots are used when you want to observe the distribution of numeric data, and are especially useful when you want to make a comparison of distributions between multiple groups.





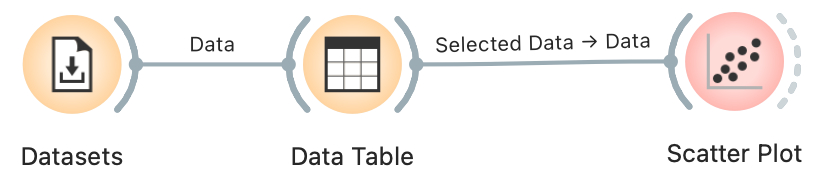
Interpretation:

* Using theViolin Plot, we can identify the most occurring values for each feature and category by looking at the x-axis and the range of values between which each feature lies by looking at the y-axis.
* We can also approximate the range of values of different classes that overlap each other in the distribution.

**4. Scatter Plot**

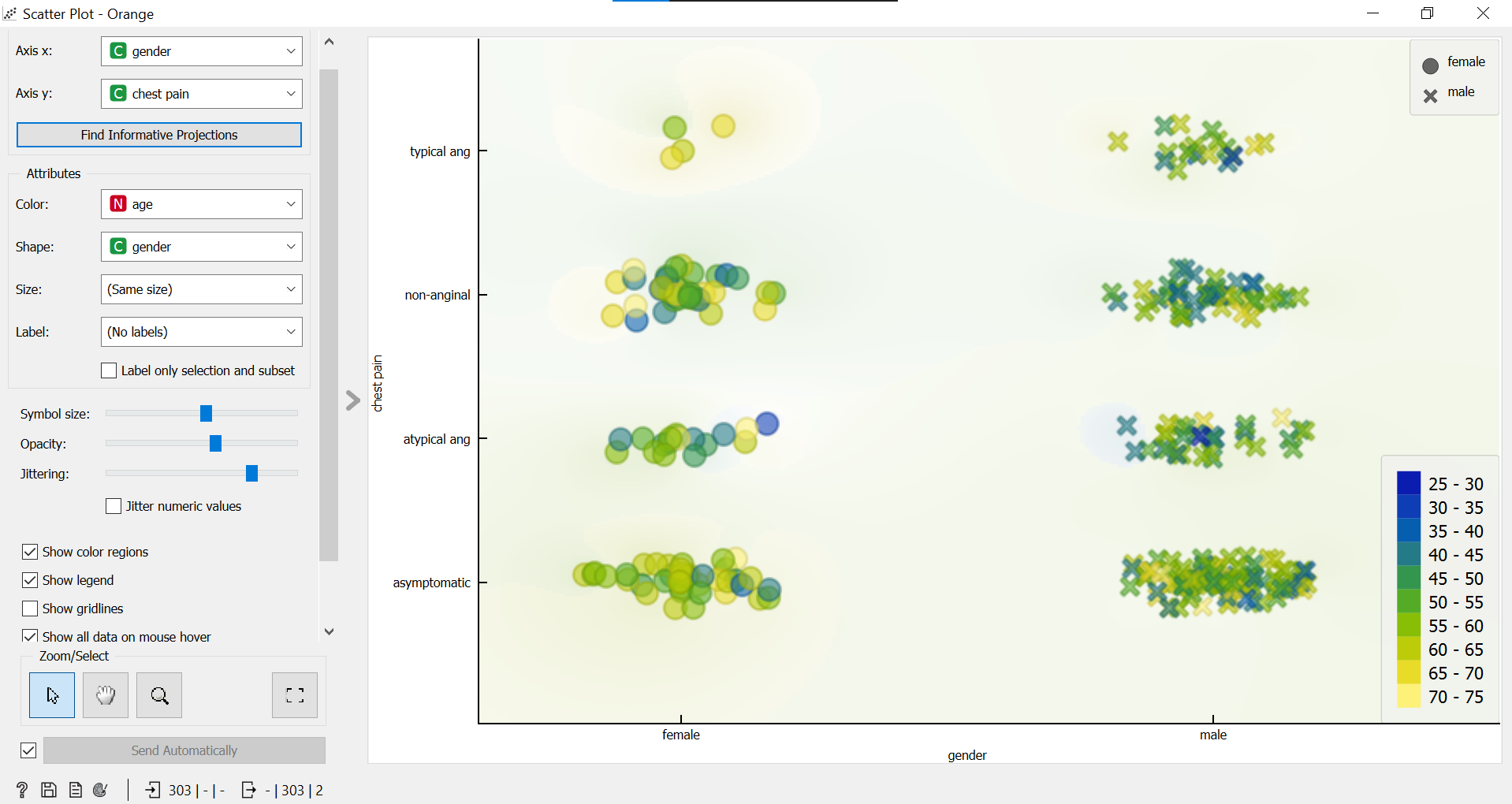
Need of Scatter Plot:

* This widget is needed for 2D scatter plot visualization. The data is displayed as a collection of points, each having the value of the x-axis attribute determining the position on the horizontal axis and the value of the y-axis attribute determining the position on the vertical axis.





**Exerc ing ang and major vessels**



**Male and Female Chest Pain**

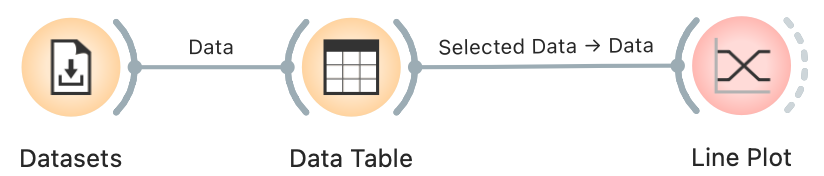
Interpretation:

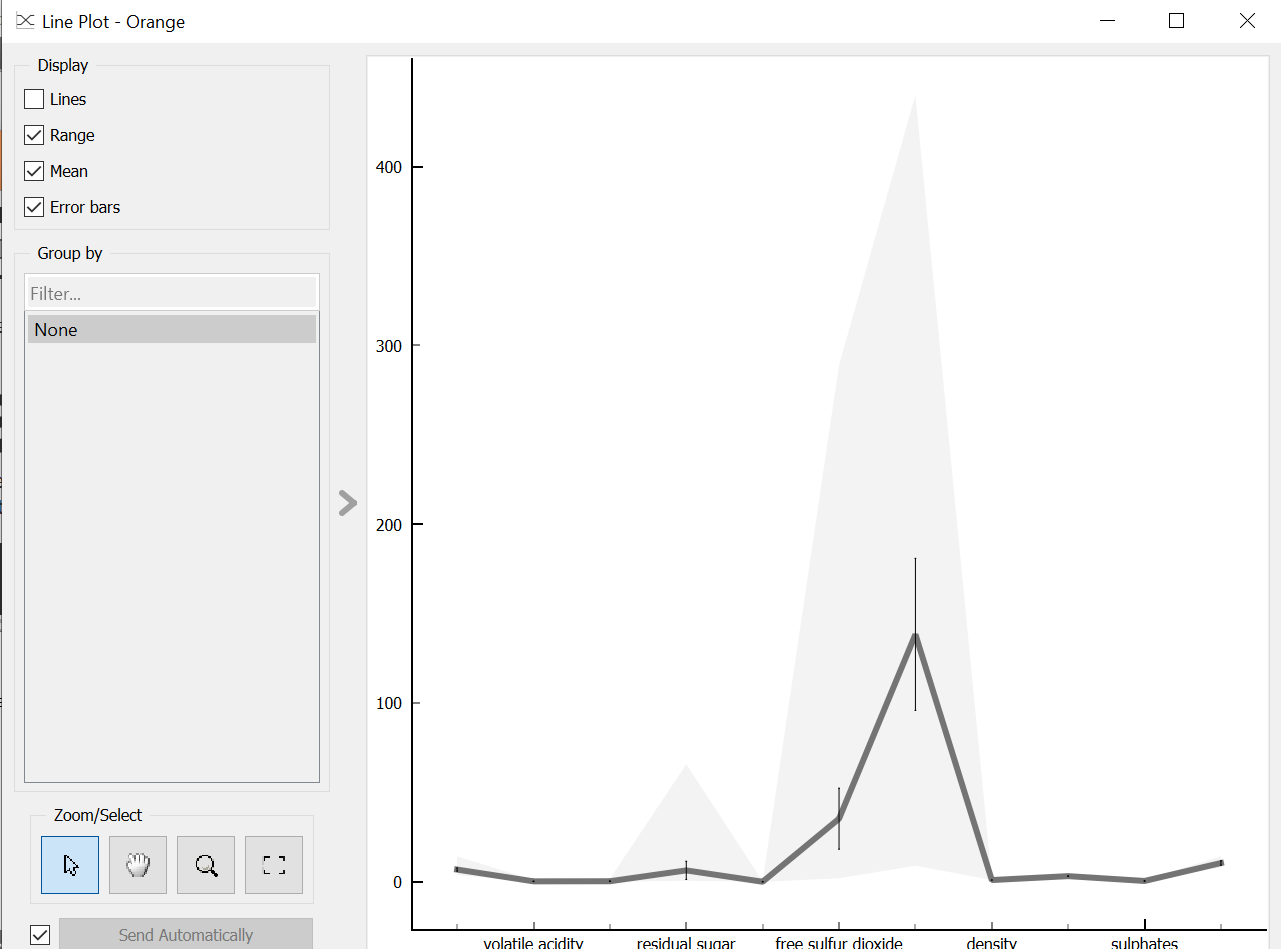
* For visualization of scatter plot, we chose gender as the x-axis and chest pain as the y-axis.
* As we can see, that male with the age bracket of 45-55 are having asymptomatic chest pain.
* We found that both male and female have asymptomatic chest pain.

**5. Line Plot**

Need of Line Plot:

* This widget is needed to display data as a series of points, connected by straight line segments. It only works for numerical data, while categorical data helps with the grouping of the data points.





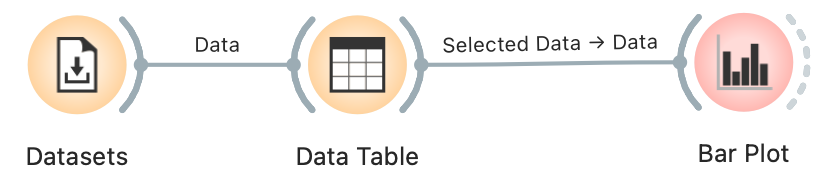
Interpretation:

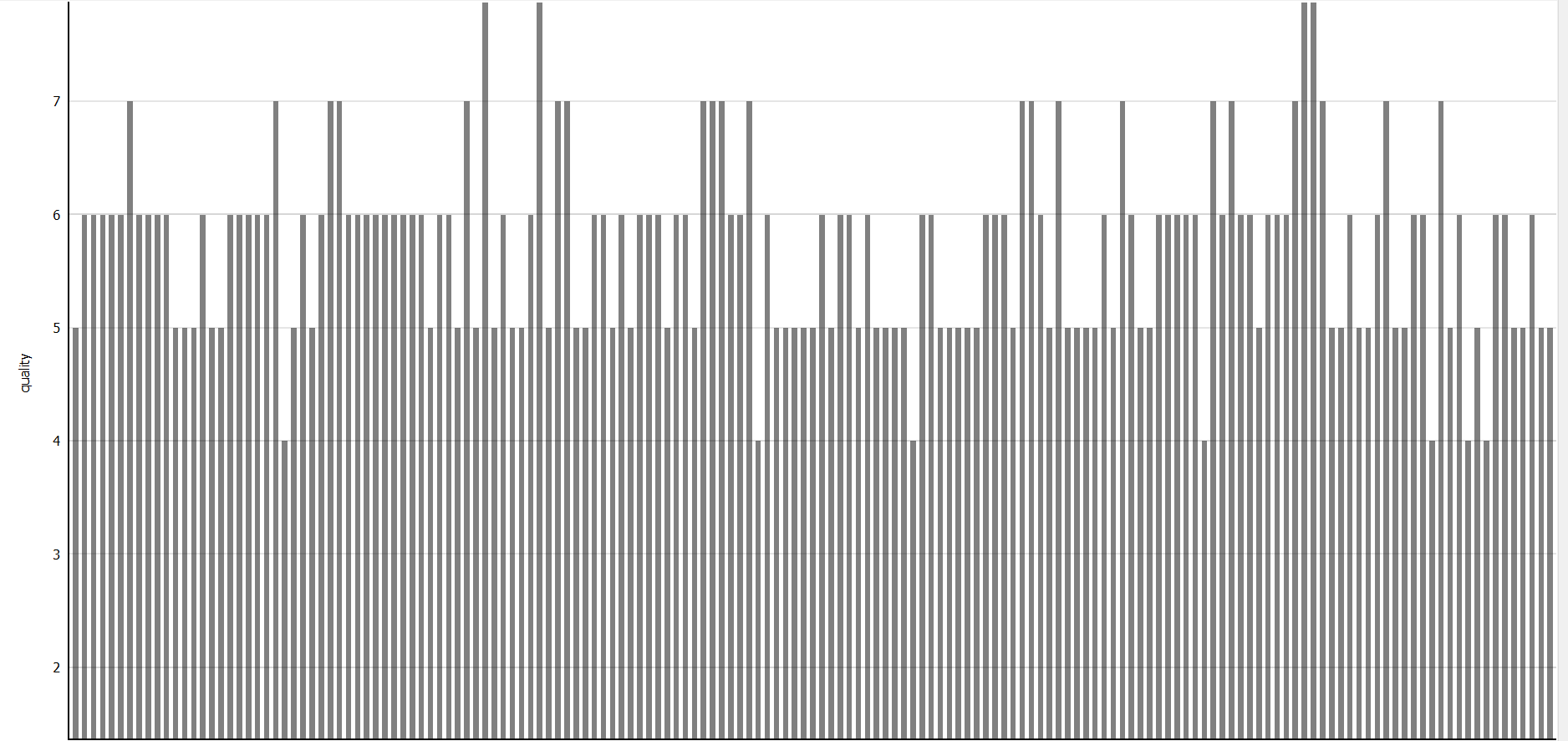
* We can see that free Sulphur oxide is high in this wine. This depicts the quality of wine.

**6. Bar Plot**

Need of Bar Plot:

* This widget is needed to visualize numeric variables and to compare categorical values. It is useful of observing outliers, distributions within groups, and comparing categories.

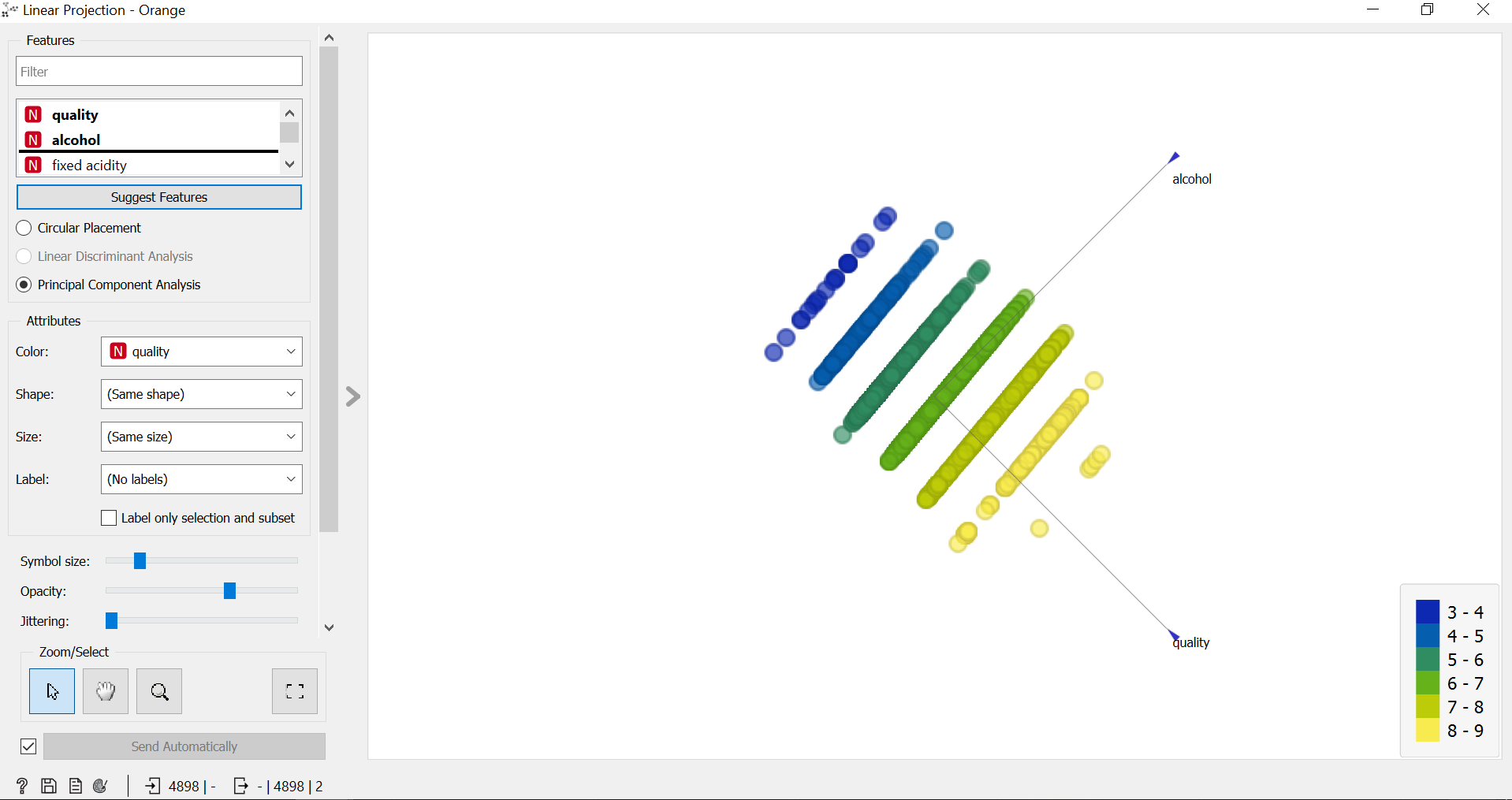




**7. Linear Projection**

Need of Linear Projection:

This widget is used to display linear projections of class-labeled data. It supports various types of projections such as circular, linear discriminant analysis, and principal component analysis.



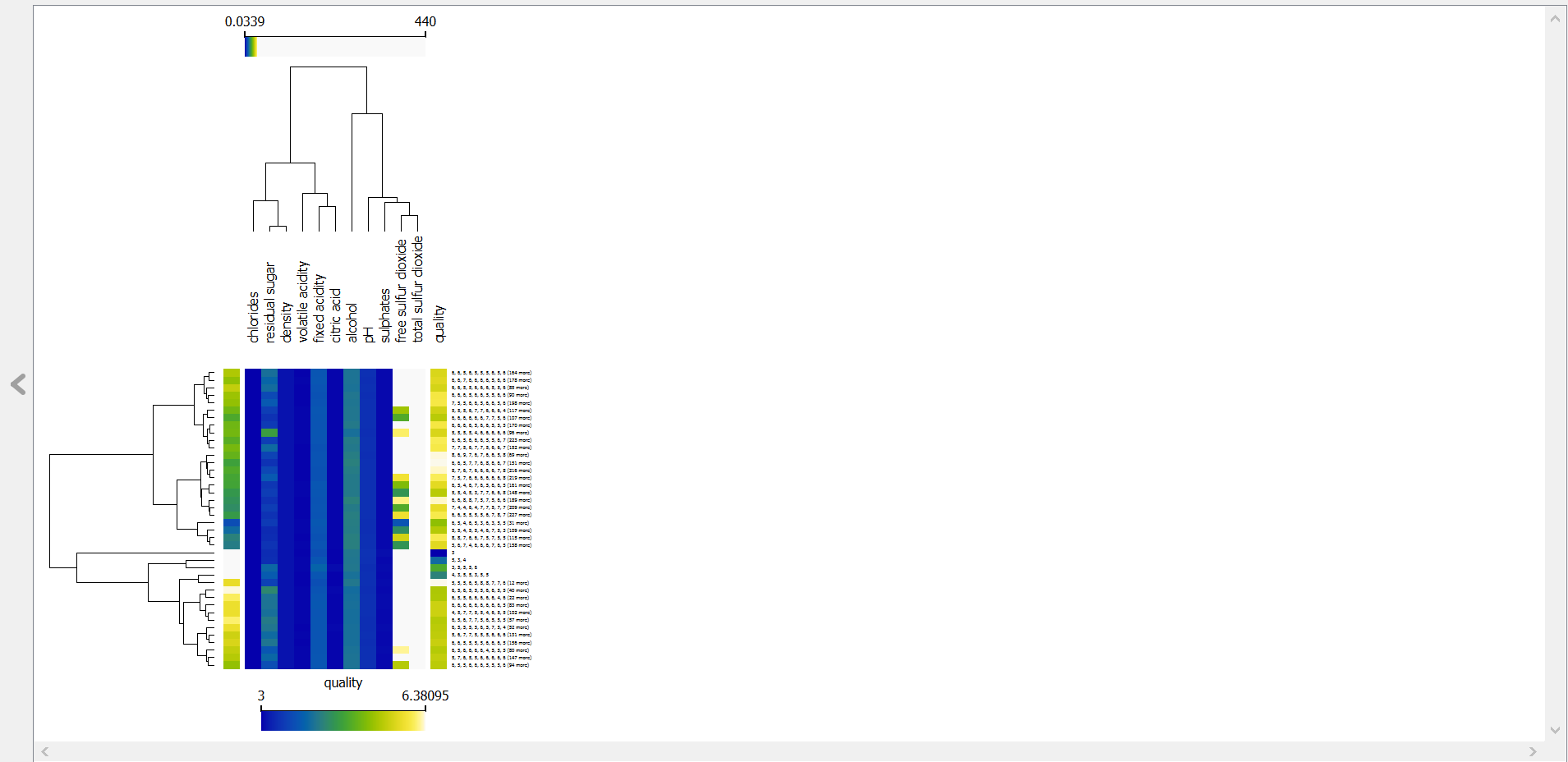
Interpretation:

* This widget takes all the features in different combinations to obtain the optimal combination where clusters can be easily distinguished.
* We can see that the quality of wine is determined by the quality of alcohol. The quality of alcohol comes under the range of 6-7.

**8. Heat Map**

Need of Heat Map:

This widget is needed for graphical visualization of attribute values in a two-way matrix using colors. By combining class variables and attributes on the x-axis and y-axis, we see where the attribute values are the strongest and where they are the weakest.



Interpretation:

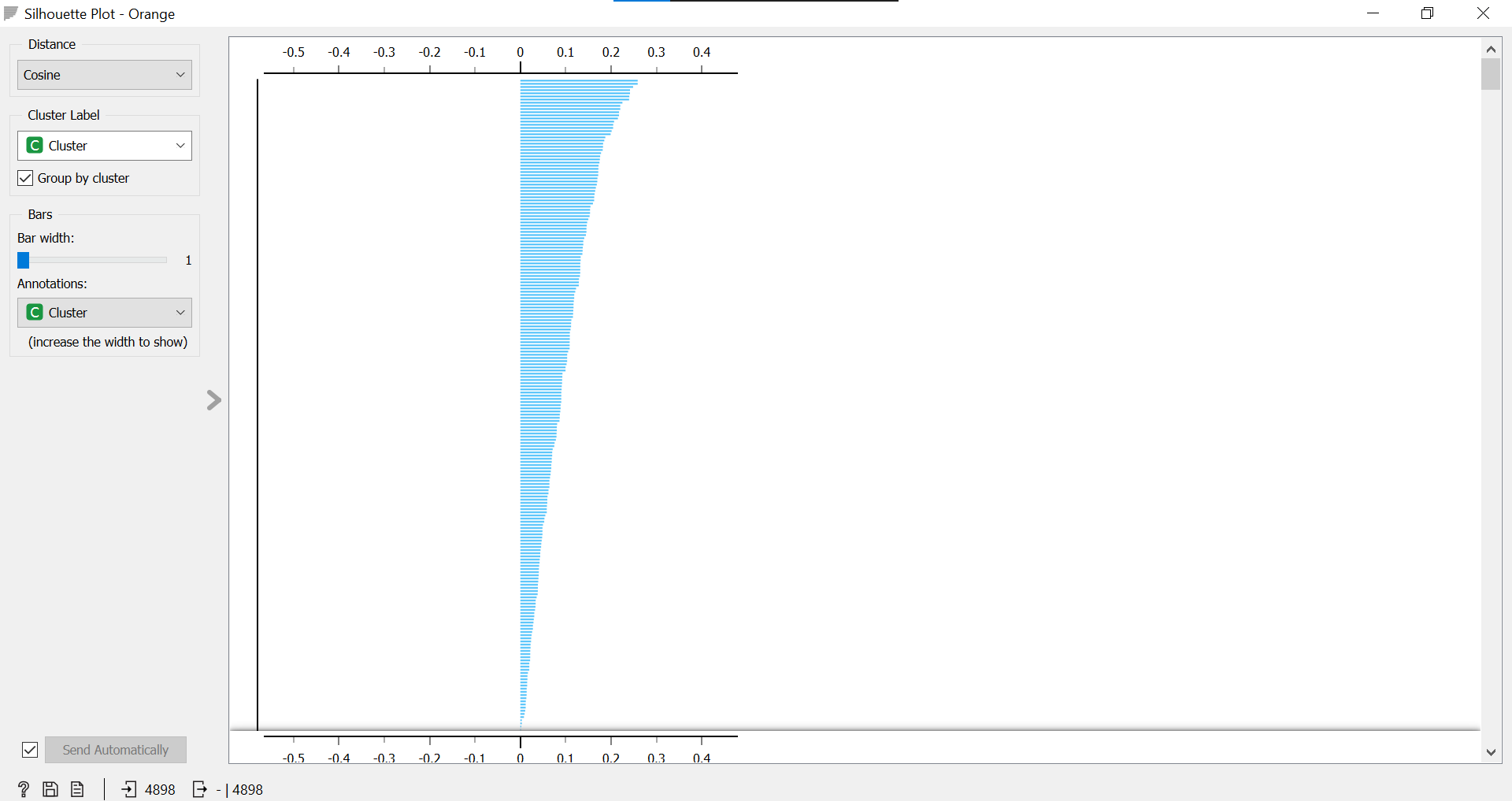
* Since higher color denotes a higher value and vice versa, we can approximate the range of values for each feature by looking at the color variations.
* We can infer that the range of values of quality lies in between 3-6.3.

**9. Silhouette Plot**

Need of Silhouette Plot:

* This widget is need to visualize the uniformity within the cluster of data and it helps the users to know the cluster quality.
* Silhouette score is a measure of how similar an object is to its own cluster in comparison to the other cluster.
* If the score is close to 1, it indicates that the data instance is close the center of it’s cluster, whereas if the score is close to 0, it indicates that the data instance is on the border between two clusters.





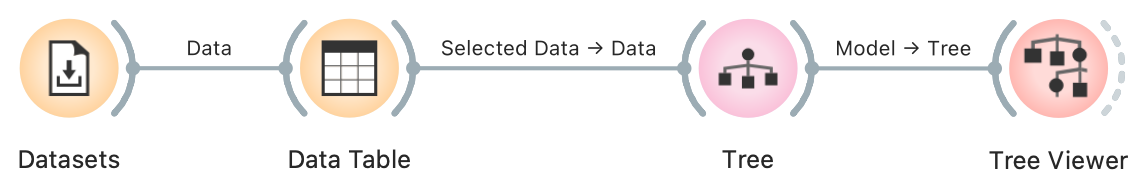
Interpretation:

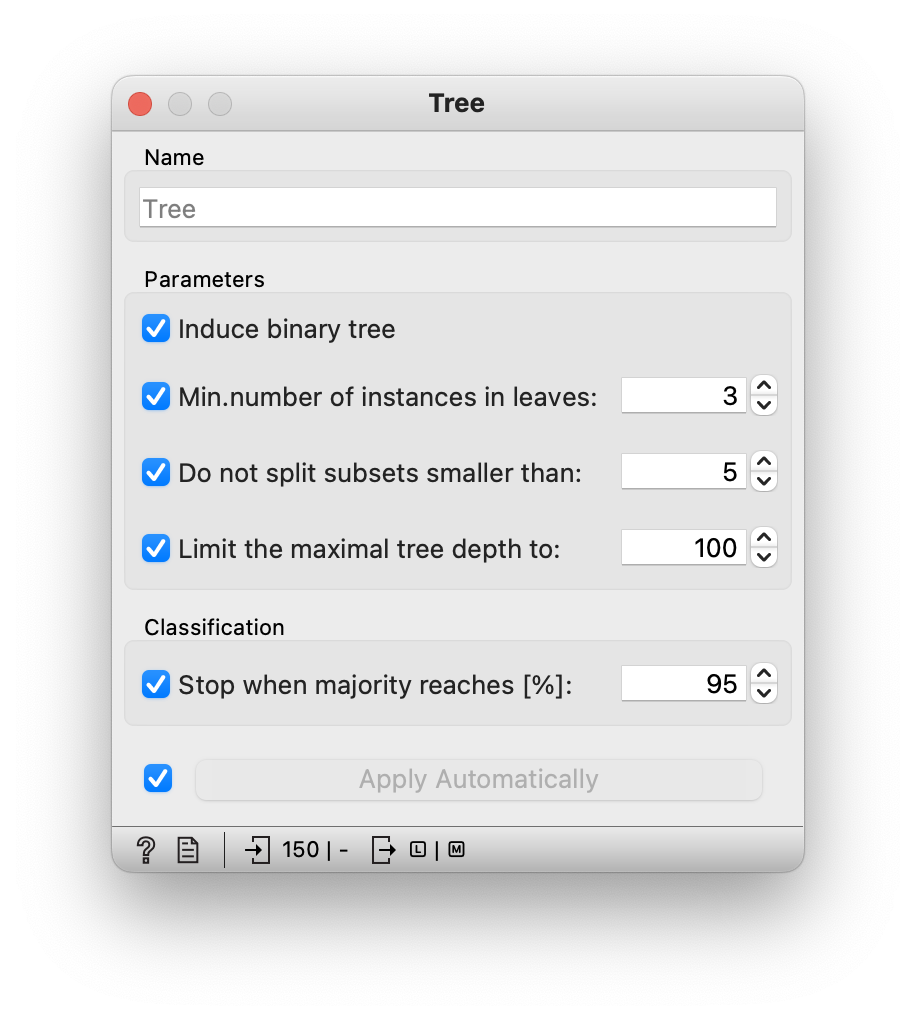
* From the Silhouette Plot, we can infer that most of the data points are close to the center of their respective clusters and they do not overlap with neighboring clusters.

**Model Category**

**1. Tree**

It is a simple algorithm that splits the data into nodes by class purity. It is a precursor to Random Forest. It can handle both categorical and numeric datasets.





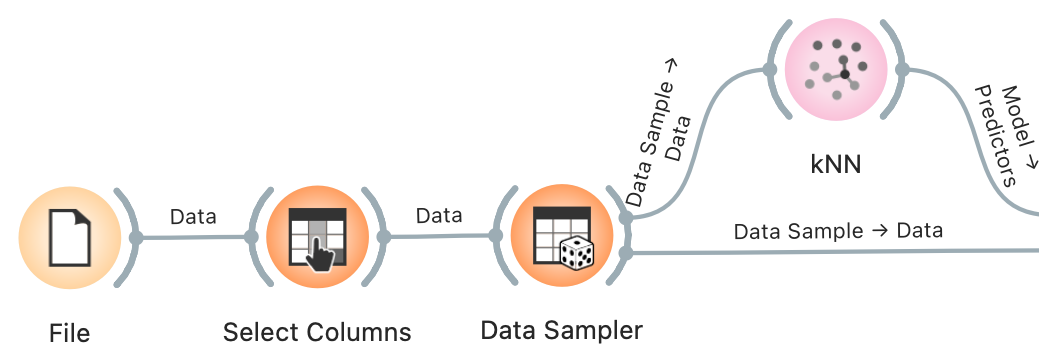
Interpretation:

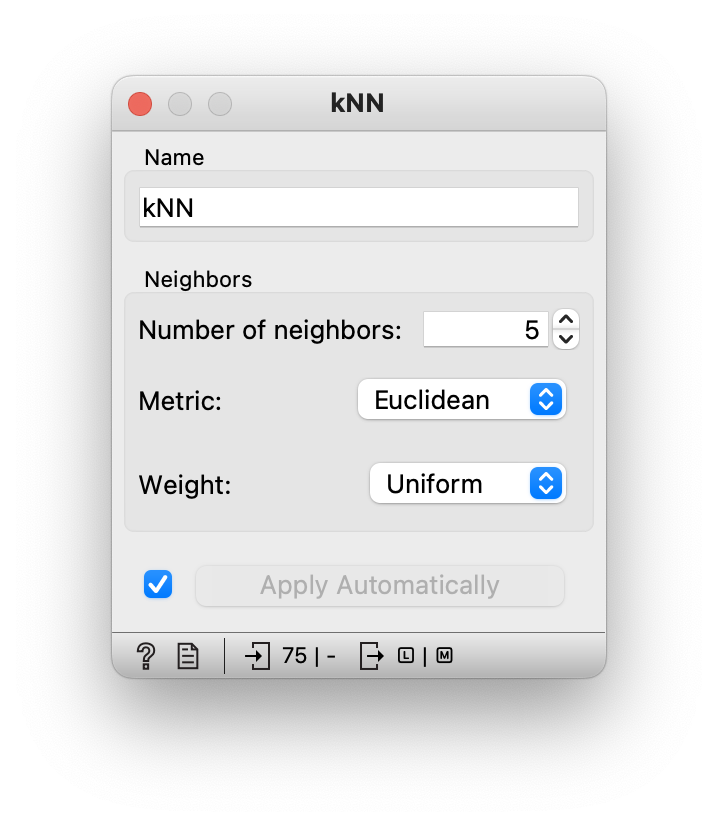
* This widget give an option limit the number of instances in leaf nodes, limit the size of subsets and limit the maximum tree depth.

**2. kNN**

Need of kNN:

The kNN widget uses the kNN algorithm that makes predictions based on the training instances most similar to the data instance being classified.

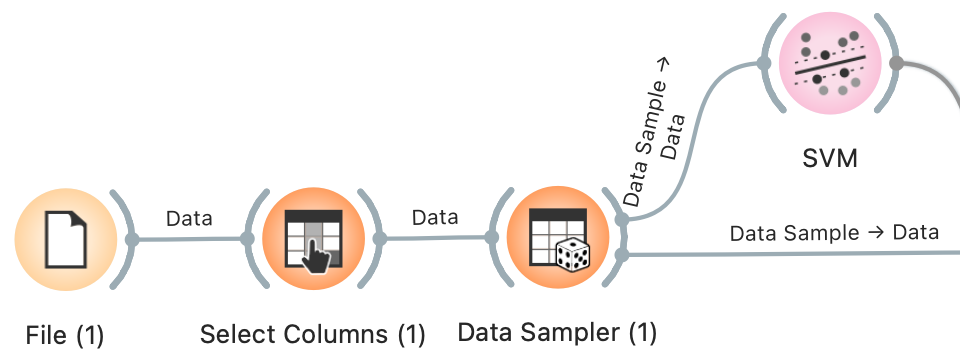


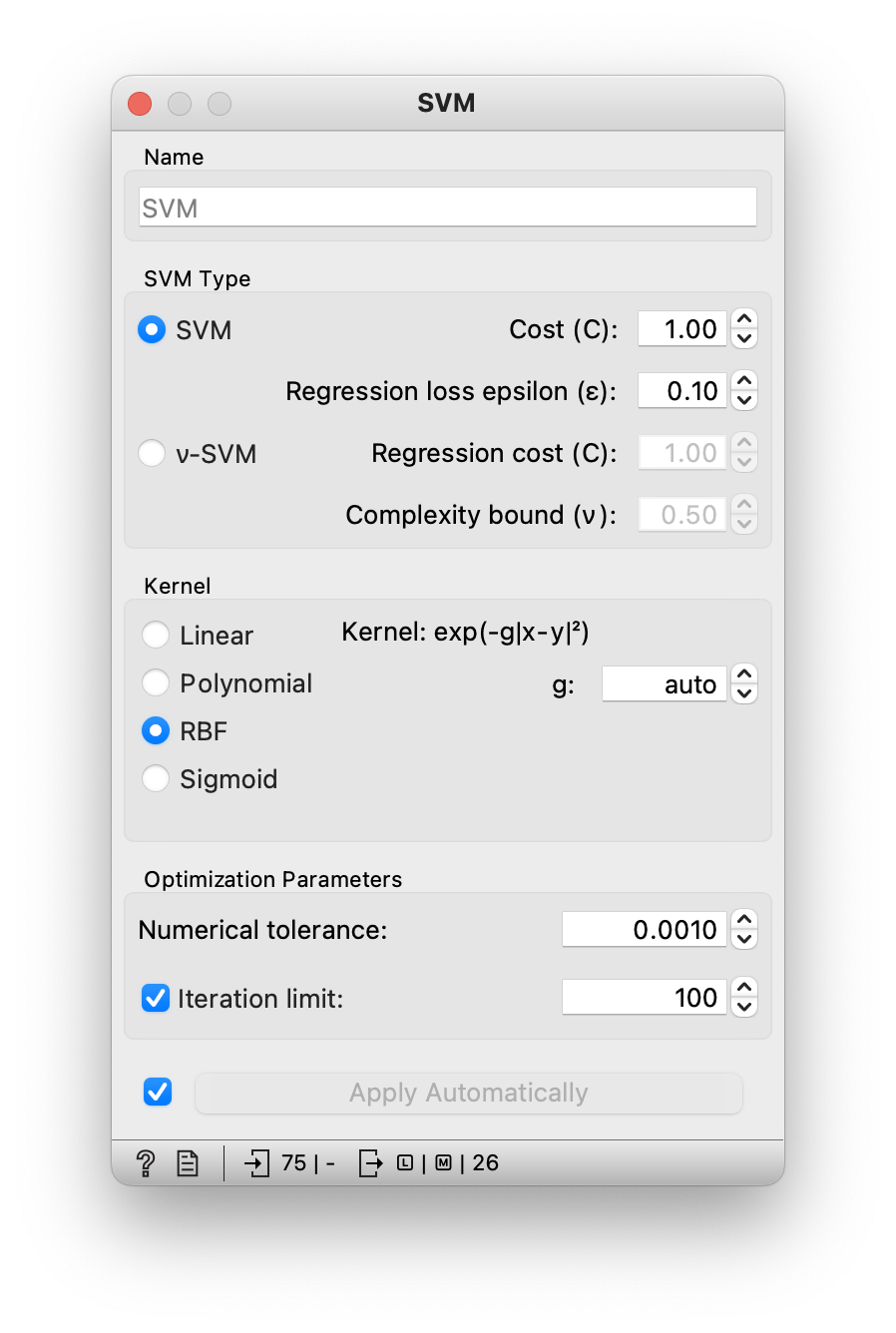


**3. SVM**

Need of SVM:

* This ML technique separates the attribute space with a hyperplane, which maximizes the margin between the instances of different classes. The technique often yields supreme predictive performance results.

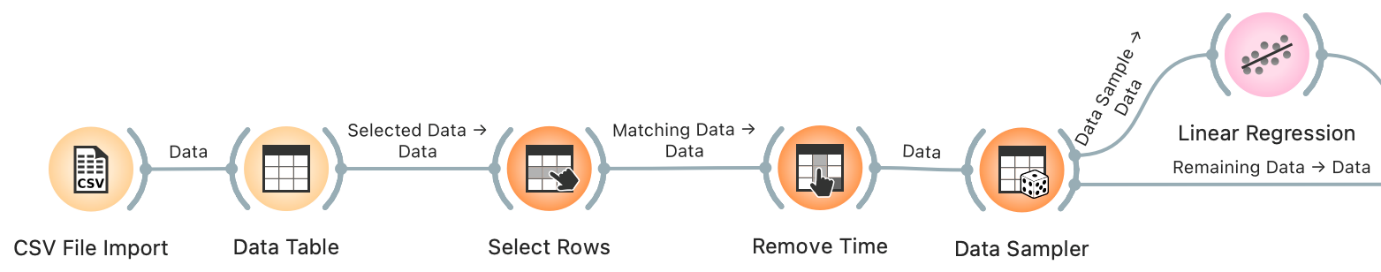


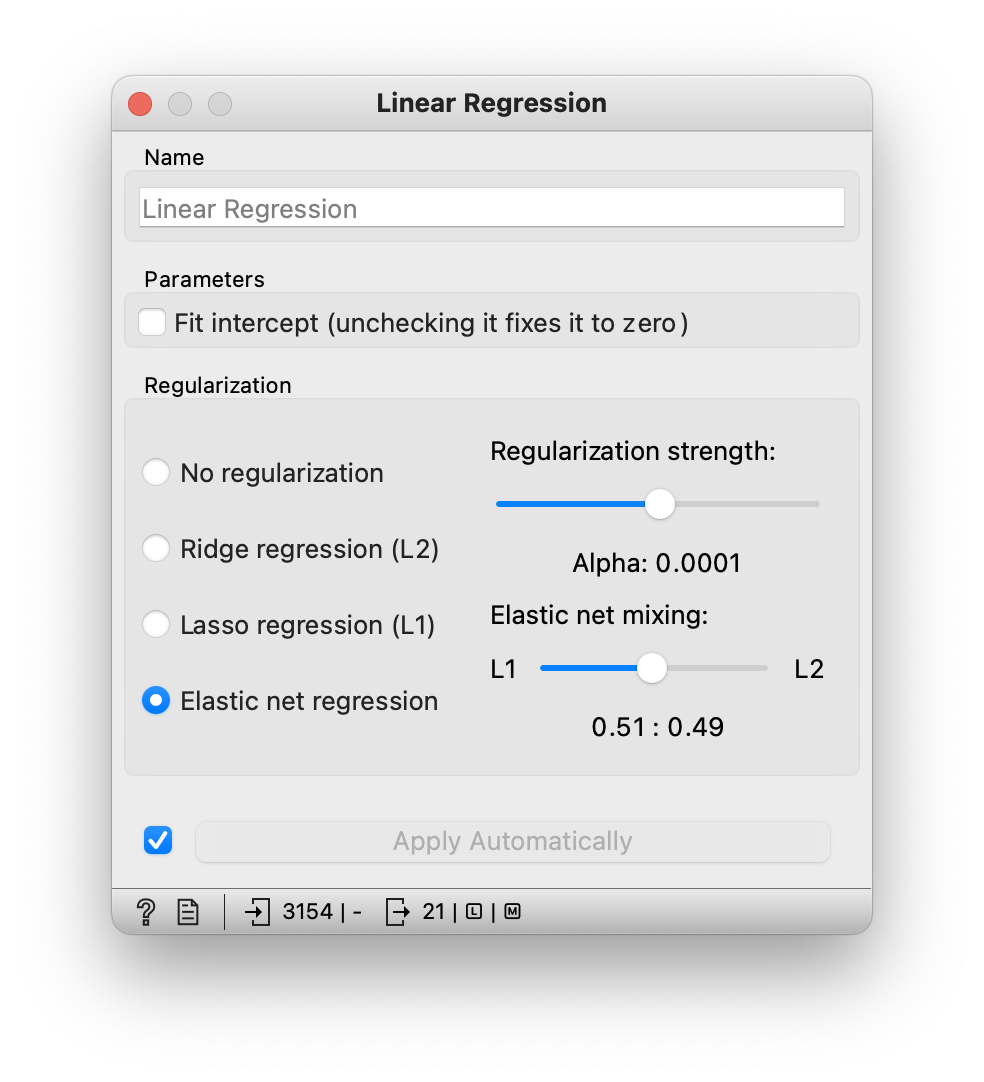


**4. Linear Regression**

Need of Linear Regression:

* Linear Regression is a method that tries to predict a value of a continuous class variable based on the value of several predictors. The model identifies the relationship between a predictor x­i and the response variable y.
* Additionally, Lasso and Ridge regularization parameters can be specified. Lasso regression minimizes a penalized version of the least squares loss function with L1-norm penalty and Ridge regularization with L2-norm penalty.

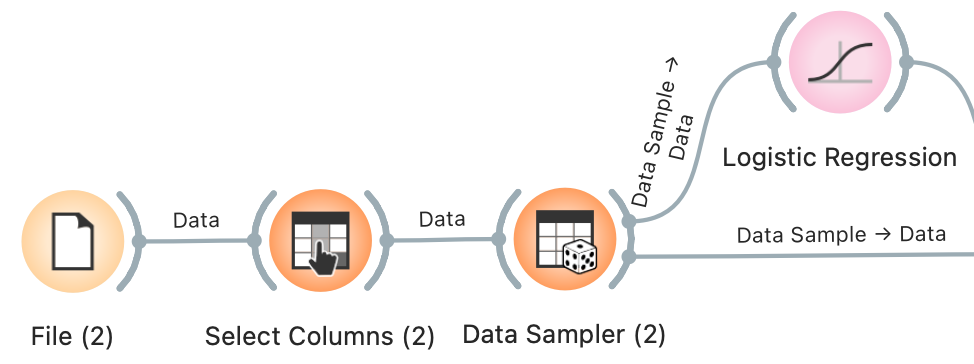


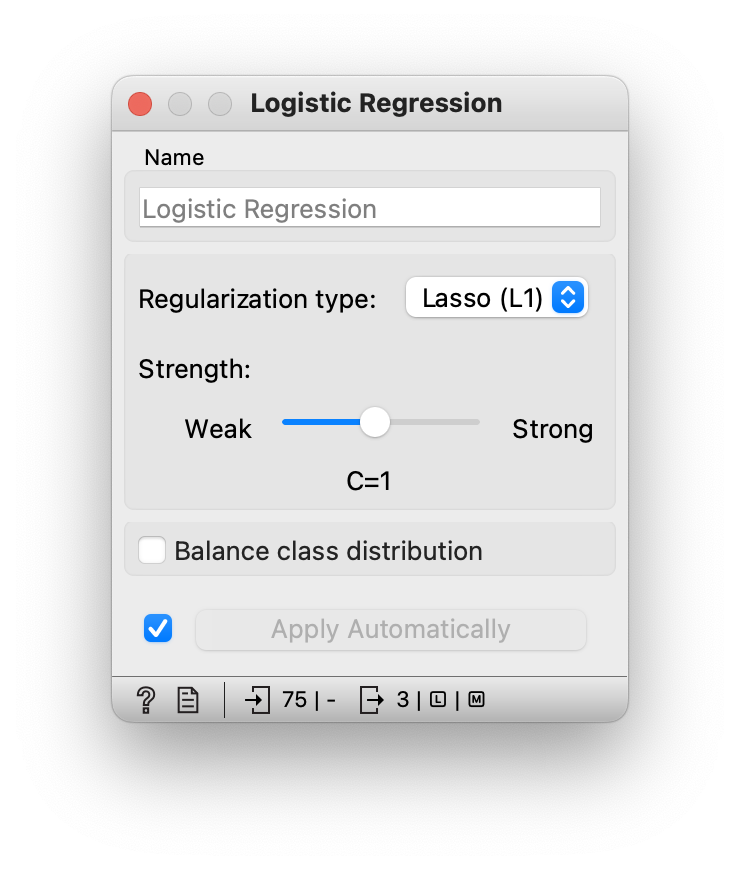


**5. Logistic Regression**

Need of Logistic Regression:

Logistic Regression is a classification technique that fits data into a logistic function. We can improve the process by acquiring sample data or selecting variables in a stepwise manner.

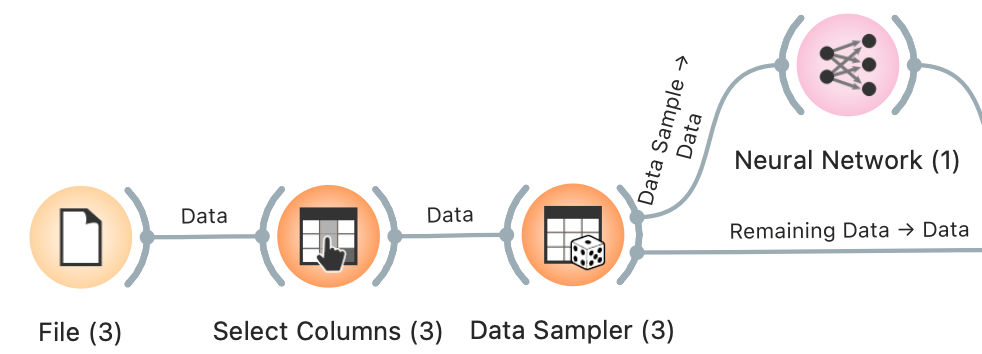


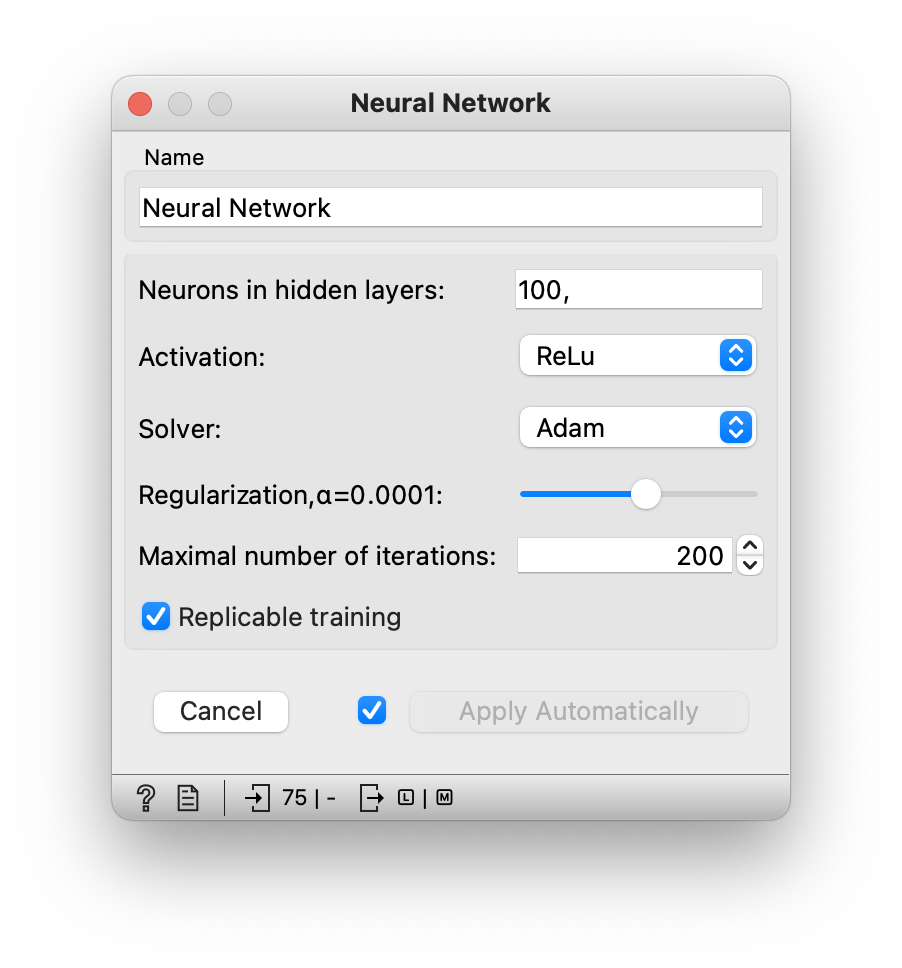


**6. Neural Network**

Need of Neural Network:

* Neural networks can help computers help make intelligent decisions with limited human assistance.
* In Orange, sklearn’s Multi-layer Perceptron algorithm is used to learn linear as well as non-linear models.



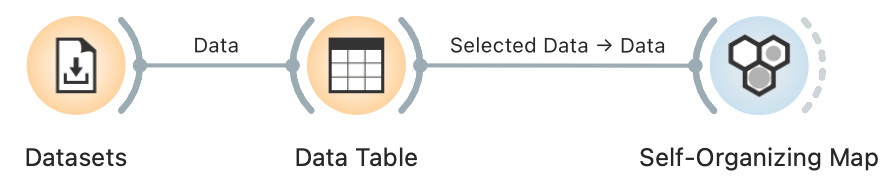


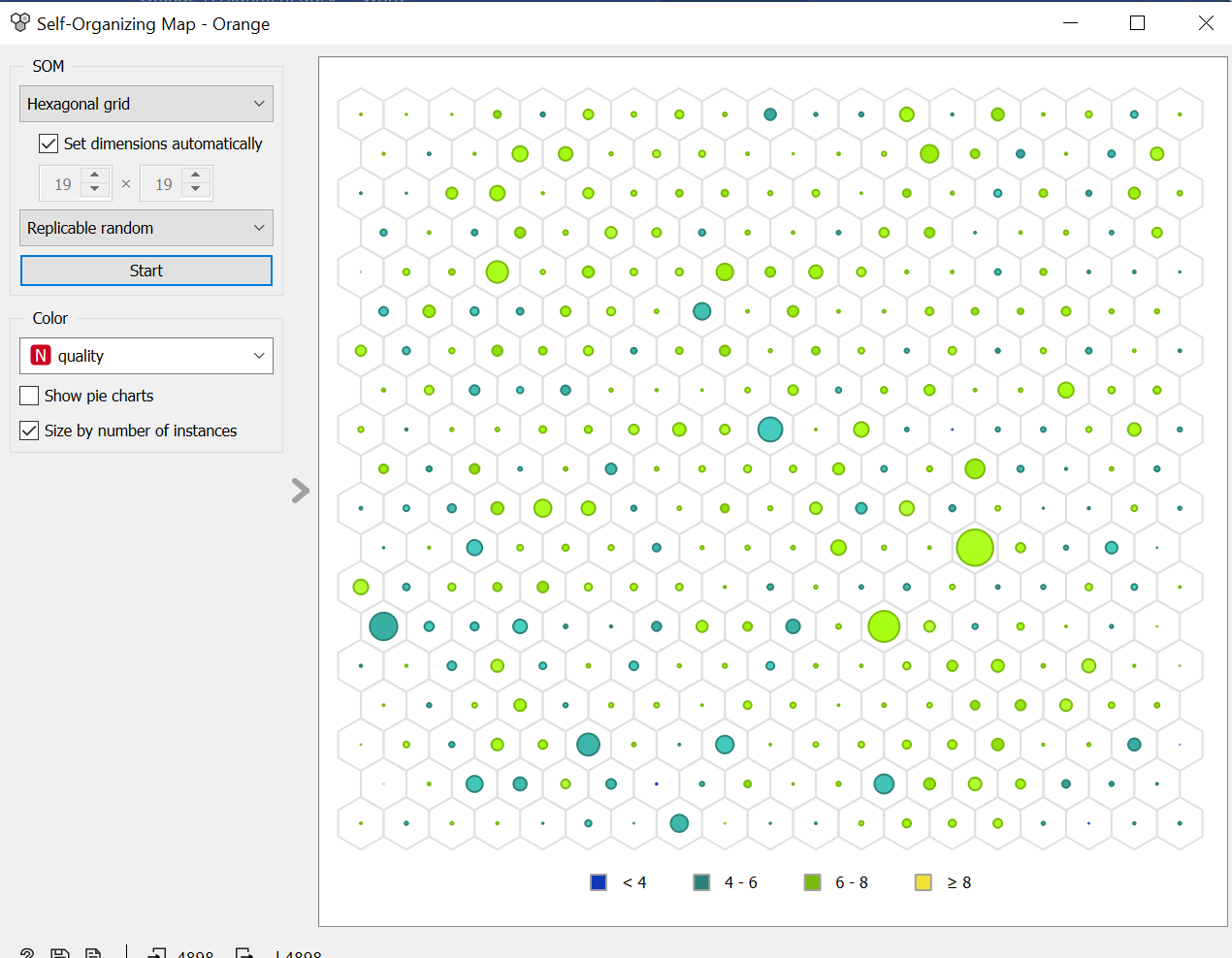
**Unsupervised Category**

**1. Self-Organizing Map**

Need of Self-Organizing Map:

* Self-organizing maps are low-dimensional projections of the input data. It is a type of ANN that is trained using unsupervised learning to produce a two-dimensional, discretized representation of the data. It is a method to do dimensionality reduction.
* It uses PCA (Principal Component Analysis). PCA shows clusters of samples based on their similarity.





Interpretation:

* This widget takes all the features in different combinations to obtain the optimal combination where clusters can be easily distinguished in a two dimensional representation using a Principal Component Analysis (PCA).