Rank

Ranking of attributes in classification or regression datasets.

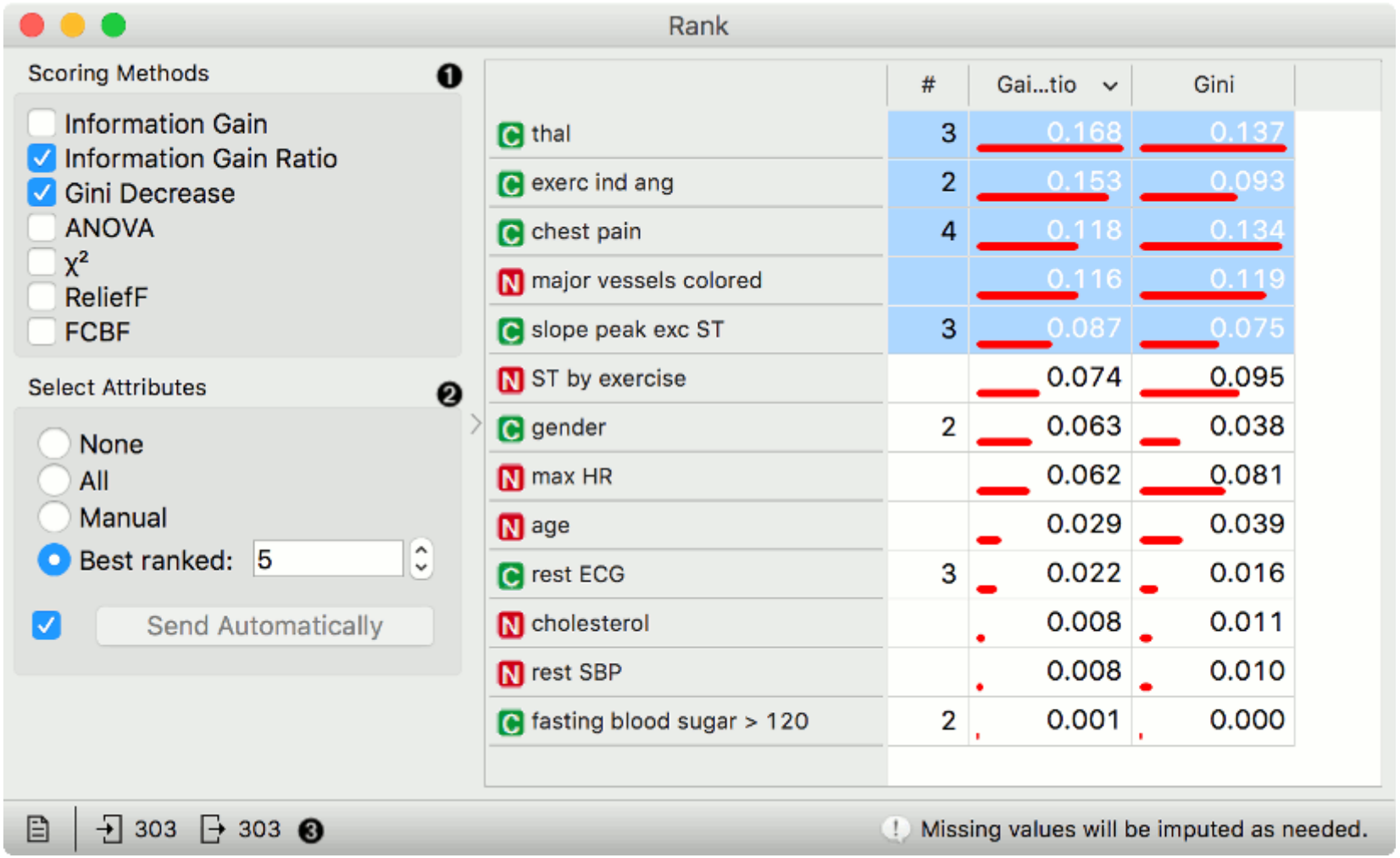
Inputs

- Data: input dataset
- Scorer: models for feature scoring

Outputs

- Reduced Data: dataset with selected attributes
- Scores: data table with feature scores
- Features: list of attributes

The **Rank** widget scores variables according to their correlation with discrete or numeric target variable, based on applicable internal scorers (like information gain, chi-square and linear regression) and any connected external models that supports scoring, such as linear regression, logistic regression, random forest, SGD, etc. The widget can also handle unsupervised data, but only by external scorers, such as PCA.



1. Select scoring methods. See the options for classification, regression and unsupervised data in the **Scoring methods** section.
2. Select attributes to output. *None* won't output any attributes, while *All* will output all of them. With manual selection, select the attributes from the table on the right. *Best ranked* will output n best ranked attributes. If *Send Automatically* is ticked, the widget automatically communicates changes to other widgets.
3. Status bar. Produce a report by clicking on the file icon. Observe input and output of the widget. On the right, warnings and errors are shown.

Scoring methods (classification)

1. Information Gain: the expected amount of information (reduction of entropy)

2. Gain Ratio: a ratio of the information gain and the attribute's intrinsic information, which reduces the bias towards multivalued features that occurs in information gain
3. Gini: the inequality among values of a frequency distribution
4. ANOVA: the difference between average vaules of the feature in different classes
5. Chi2: dependence between the feature and the class as measure by the chi-square statistic
6. ReliefF: the ability of an attribute to distinguish between classes on similar data instances
7. FCBF (Fast Correlation Based Filter): entropy-based measure, which also identifies redundancy due to pairwise correlations between features

Additionally, you can connect certain learners that enable scoring the features according to how important they are in models that the learners build (e.g. Logistic Regression, Random Forest, SGD). Please note that the data is normalized before ranking.

Scoring methods (regression)

1. Univariate Regression: linear regression for a single variable
2. RReliefF: relative distance between the predicted (class) values of the two instances.

Additionally, you can connect regression learners (e.g. Linear Regression, Random Forest, SGD). Please note that the data is normalized before ranking.

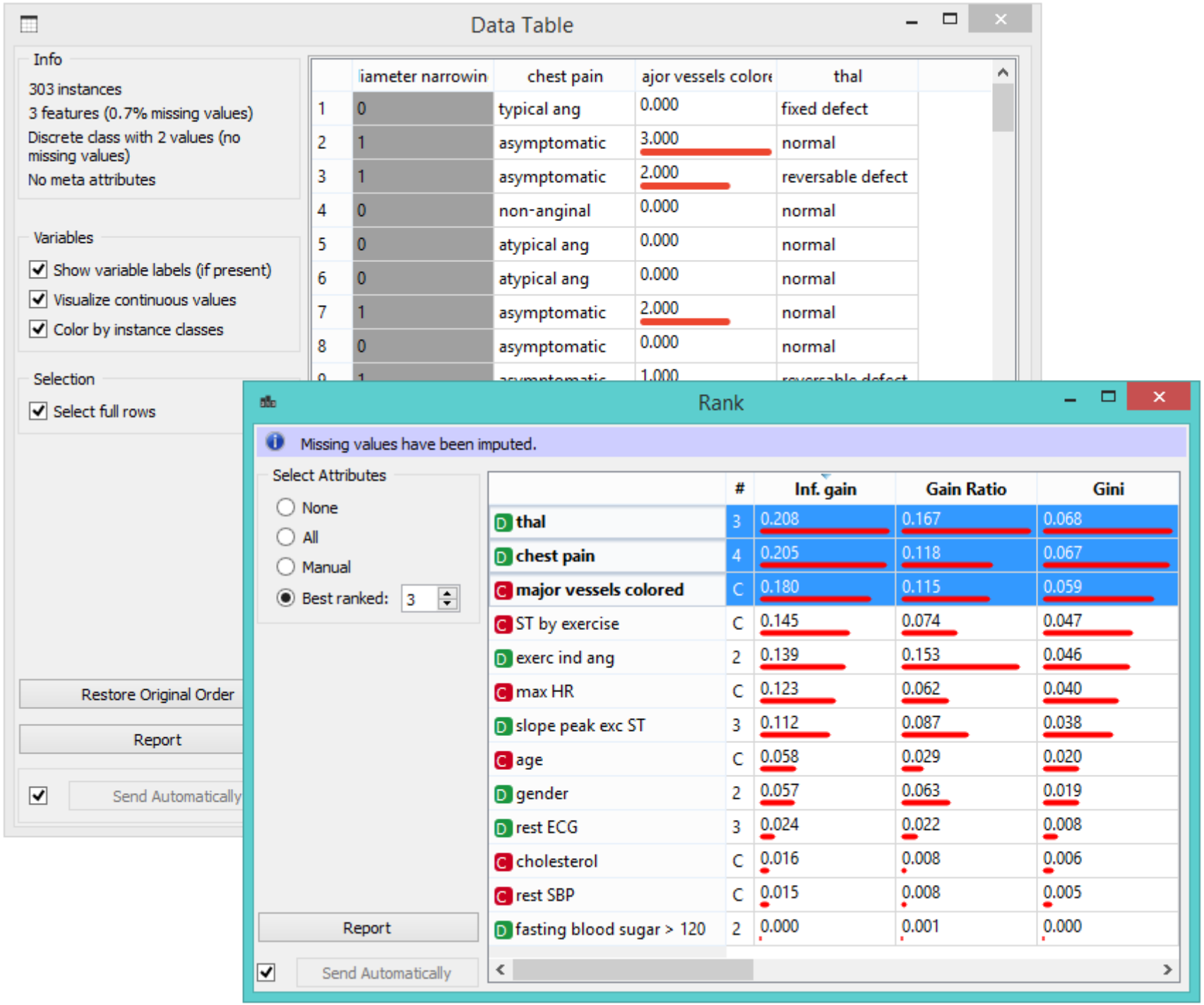
Scoring method (unsupervised)

Currently, only PCA is supported for unsupervised data. Connect PCA to Rank to obtain the scores. The scores correspond to the correlation of a variable with the individual principal component.

Example: Attribute Ranking and Selection

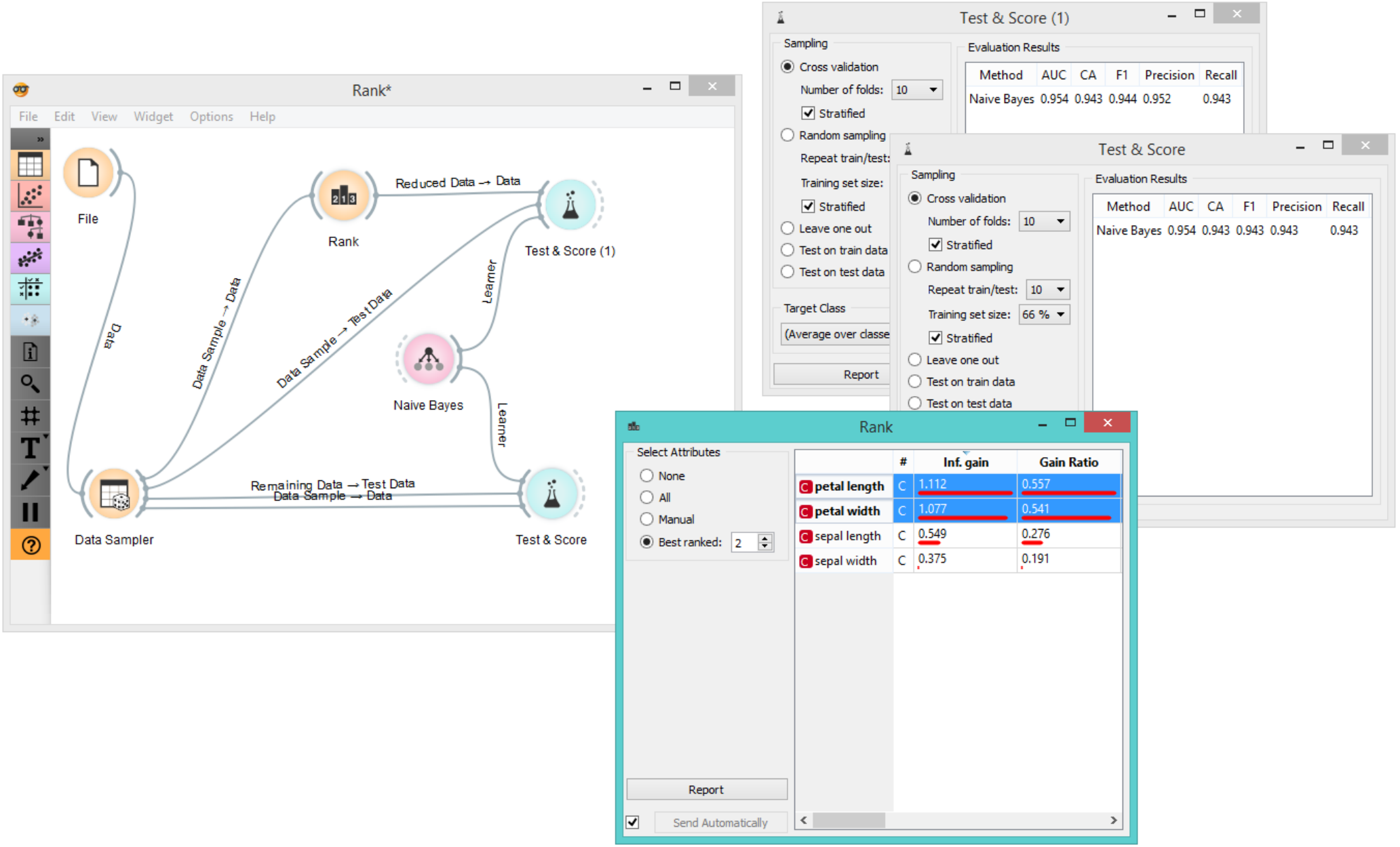
Below, we have used the **Rank** widget immediately after the File widget to reduce the set of data attributes and include only the most informative ones:

Notice how the widget outputs a dataset that includes only the best-scored attributes:


Data Table
Info
303 instances
3 features (0.7% missing values)
Discrete class with 2 values (no missing values)
No meta attributes
Variables
Show variable labels (if present)
Visualize continuous values
Color by instance classes
Selection
Select full rows
Restore Original Order
Report
Send Automatically
iameter narrowin
chest pain
ajor vessels colore
thal
1 0 typical ang 0.000 fixed defect
2 1 asymptomatic 3.000 normal
3 1 asymptomatic 2.000 reversable defect
4 0 non-anginal 0.000 normal
5 0 atypical ang 0.000 normal
6 0 atypical ang 0.000 normal
7 1 asymptomatic 2.000 normal
8 0 asymptomatic 0.000 normal
Rank
Missing values have been imputed.
Select Attributes
None
All
Manual
Best ranked: 3
Report
Send Automatically
Inf. gain Gain Ratio Gini
thal 3 0.208 0.167 0.068
chest pain 4 0.205 0.118 0.067
major vessels colored C 0.180 0.115 0.059
ST by exercise C 0.145 0.074 0.047
exerc ind ang 2 0.139 0.153 0.046
max HR C 0.123 0.062 0.040
slope peak exc ST 3 0.112 0.087 0.038
age C 0.058 0.029 0.020
gender 2 0.057 0.063 0.019
rest ECG 3 0.024 0.022 0.008
cholesterol C 0.016 0.008 0.006
rest SBP C 0.015 0.008 0.005
fasting blood sugar > 120 2 0.000 0.001 0.000

Example: Feature Subset Selection for Machine Learning

What follows is a bit more complicated example. In the workflow below, we first split the data into a training set and a test set. In the upper branch, the training data passes through the **Rank** widget to select the most informative attributes, while in the lower branch there is no feature selection. Both feature selected and original datasets are passed to their own Test & Score widgets, which develop a *Naive Bayes* classifier and score it on a test set.



For datasets with many features, a naive Bayesian classifier feature selection, as shown above, would often yield a better predictive accuracy.