

# Visualization

## with KNIME

# KNIME : create workflow

KNIME - /Users/nistor/knime-workspace

Quick Access

**KNIME Explorer**

- 02\_SocialNetworkAnalysis
- 050018\_ChurnPrediction
- ChurnPredictionDeployment
- ChurnPredictionTraining
- 050019\_TwitterAnalysis
- 002005\_PMMI\_Examples
- 002006\_PMMI\_Ensembles\_blog
- 050004\_Lastfm\_Recommendations
- 050005\_Social\_Media\_Clustering
- Dim Reduction Techniques
- Example Workflow
- KNIME\_project

**Favorite Nodes**

- Personal favorite nodes
- Most frequently used nodes
- Last used nodes

**Node Repository**

- Views
  - JFreeChart
    - Scatter Plot (JFreeChart)
    - Scatter Matrix
    - Scatter Plot
- Scripting
  - R
    - Meta Nodes
      - Grouped ScatterPlot
- KNIME Labs
  - Interactive Views
    - JavaScript Scatter Plot

**Select a wizard**

This wizard creates a new KNIME workflow project.

We are happy to help you get started emails!

If you do want regular emails, you haven't done so already.

If you are new to KNIME, you can also start here. You can also send us an email if you have any questions or comments about KNIME. We are always happy to help you get started with KNIME.

Wizards:

- New KNIME Workflow
- New KNIME Workflow Group
- General
- Business Intelligence and Reporting Tools
- Connection Profiles
- Eclipse Modeling Framework
- Java
- Java Emitter Templates
- KNIME
- Plug-in Development
- SQL Development

Show All Wizards.

**Outline**

An outline is not available.

**KNIME Console**

```
WARN Scatter Plot 2:9
WARN Scatter Plot 2:9
WARN Scatter Plot 2:9
WARN Scatter Plot 2:9
WARN Color Manager 2:11
WARN Color Manager 2:11
WARN Hierarchical Clustering 2:7
WARN Scatter Plot 2:9
WARN KnimeRemoteFileSystem
WARN KnimeRemoteFileSystem
WARN Color Manager 0:67
WARN Color Manager 0:67
WARN Color Manager 0:67
Some columns are ignored: bounds missing.
Some columns are ignored: bounds missing.
Some columns are ignored: bounds missing.
Only the first 2500 rows are displayed.
Column "Cluster" has no nominal values set: execute predecessor or add Bi...
Column "Cluster" has no nominal values set: execute predecessor or add Bi...
Execution canceled
Some columns are ignored: bounds missing.
Connecting to server "http://publicserver.knime.org:80/tomee/ejb" failed
Connecting to server "http://publicserver.knime.org:80/tomee/ejb" failed
Column "Churn" has no nominal values set: execute predecessor or add Bi...
Column "Churn" has no nominal values set: execute predecessor or add Bi...
Column "Churn" has no nominal values set: execute predecessor or add Bi...
```

**Node Description**

**Scatter Plot:**  
Creates a scatterplot of two selected attributes.

**PCA:**  
Principal component analysis

**Color Manager:**  
Assigns colors to a selected nominal or numeric column.

**File Reader:**  
Flexible reader for ASCII files.

**k-Means:**  
Creates a crisp center based clustering.

**Hierarchical Clustering:**  
Performs Hierarchical Clustering.

file:///var/folders/j1/s3vnwxd913q7vdkndc1qhbb0000gn/T/intro7231400188949890018.html

# Read data

The screenshot shows the KNIME Analytics Platform interface with the following components:

- KNIME Explorer**: Shows project structure with nodes like "02\_SocialNetworkAnalysis", "050018\_ChurnPrediction", and "050019\_TwitterAnalysis".
- Node Repository**: Shows the "Read" category under "IO", which includes "File Reader", "ARFF Reader", "CSV Reader", "Line Reader", "Table Reader", "PMML Reader", "Model Reader", "Fixed Width File Reader", "List Files", "Read Images", "Read XLS Sheet Names", and "XLS Reader".
- Workflow Area**: Displays a single node, "File Reader", labeled "Node 1".
- Console**: Shows KNIME Console output with several WARN messages related to scatter plots and color managers.
- Node Description**: A detailed description of the "File Reader" node, explaining its purpose and configuration options.

**File Reader Node Description**

This node can be used to read data from an ASCII file or URL location. It can be configured to read various formats. When you open the node's configuration dialog and provide a filename, it tries to guess the reader's settings by analyzing the content of the file. Check the results of these settings in the preview table. If the data shown is not correct or an error is reported, you can adjust the settings manually (see below).

The file analysis runs in the background and can be cut short by clicking the "Quick scan", which shows if the analysis takes longer. In this case the file is not analyzed completely, but only the first fifty lines are taken into account. It could happen then, that the preview appears looking fine, but the execution of the File Reader fails, when it reads the lines it didn't analyze. Thus it is recommended you check the settings, when you cut an analysis short.

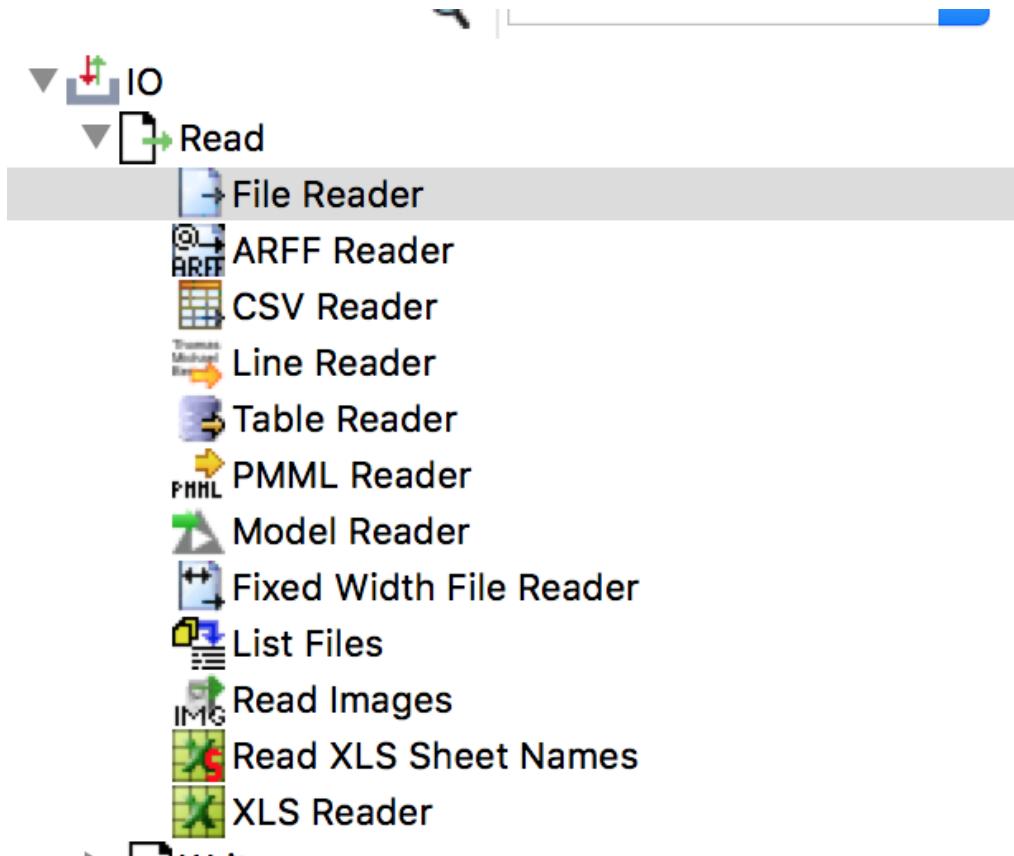
**Dialog Options**

**ASCII file location**

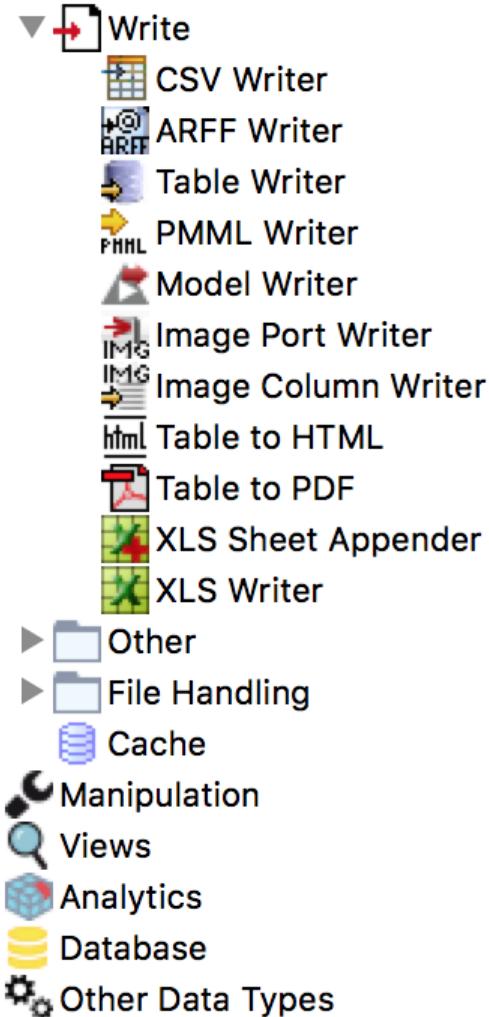
Enter a valid file name or URL. When you press ENTER, the file is analyzed and the settings pre-set. You can also choose a previously read file from the drop-down list, or select a file from the "Browse..."

```
WARN Scatter Plot      2:9      Only the first 2500 rows are displayed.
WARN Color Manager    2:11
WARN Color Manager    2:11
WARN Hierarchical Clustering 2:7
WARN Scatter Plot      2:9      Column "Cluster" has no nominal values set: execute predecessor or add Binr
WARN Color Manager    0:67      Column "Cluster" has no nominal values set: execute predecessor or add Binr
WARN Color Manager    0:67      Execution canceled
WARN KnimeRemoteFileSystem
WARN KnimeRemoteFileSystem
WARN Color Manager    0:67      Some columns are ignored: bounds missing.
WARN Color Manager    0:67      Connecting to server "http://publicserver.knime.org:80/tomee/ejb" failed
WARN Color Manager    0:67      Connecting to server "http://publicserver.knime.org:80/tomee/ejb" failed
WARN Color Manager    0:67      Column "Churn" has no nominal values set: execute predecessor or add Binr
WARN Color Manager    0:67      Column "Churn" has no nominal values set: execute predecessor or add Binr
WARN File Reader       3:1      Column "Churn" has no nominal values set: execute predecessor or add Binr
WARN File Reader       3:1      No Settings available.
```

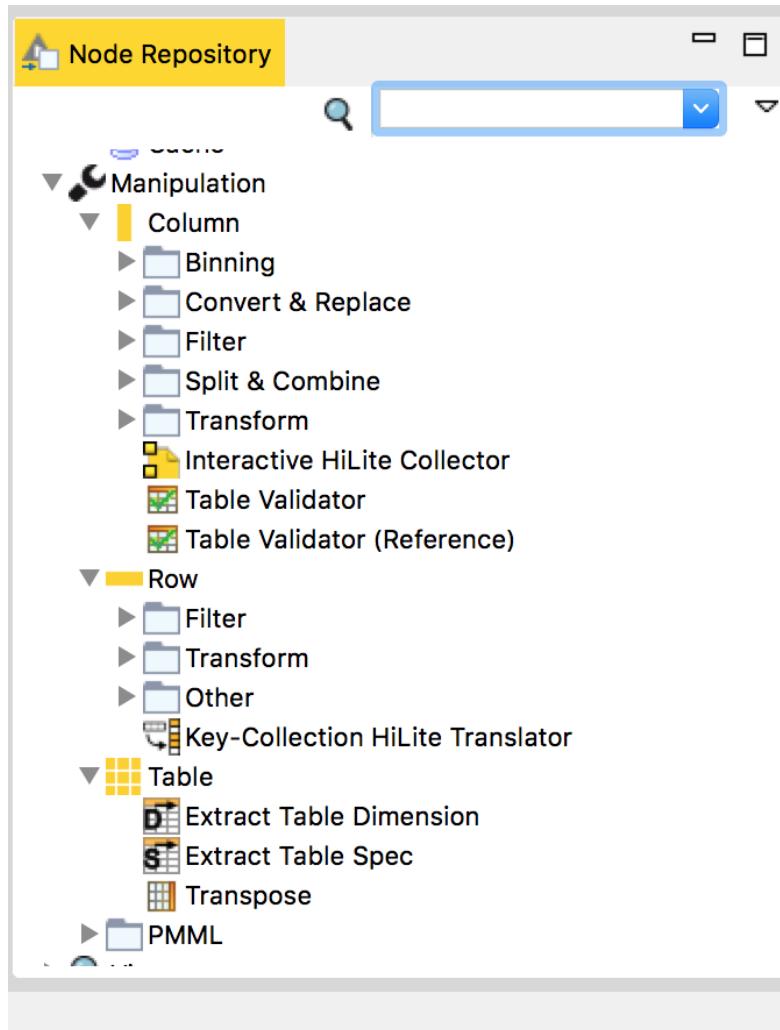
# Several types of importing



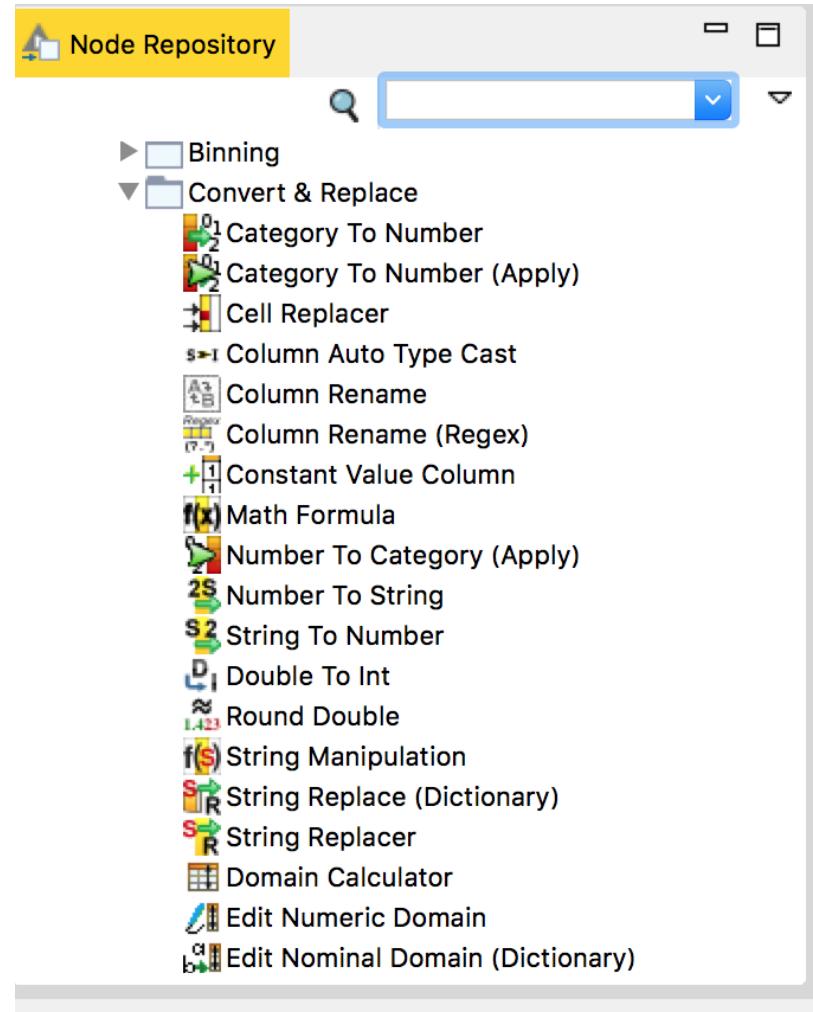
# Write data & results



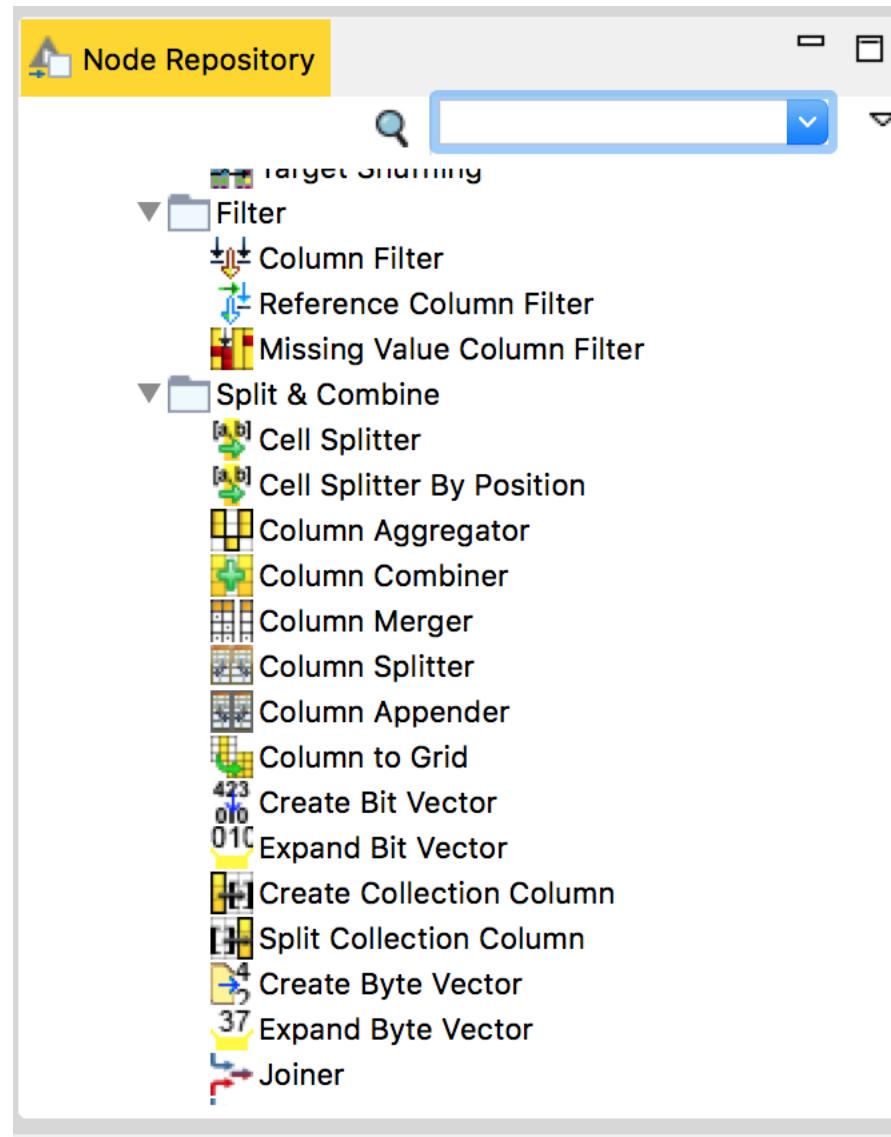
# Data manipulation



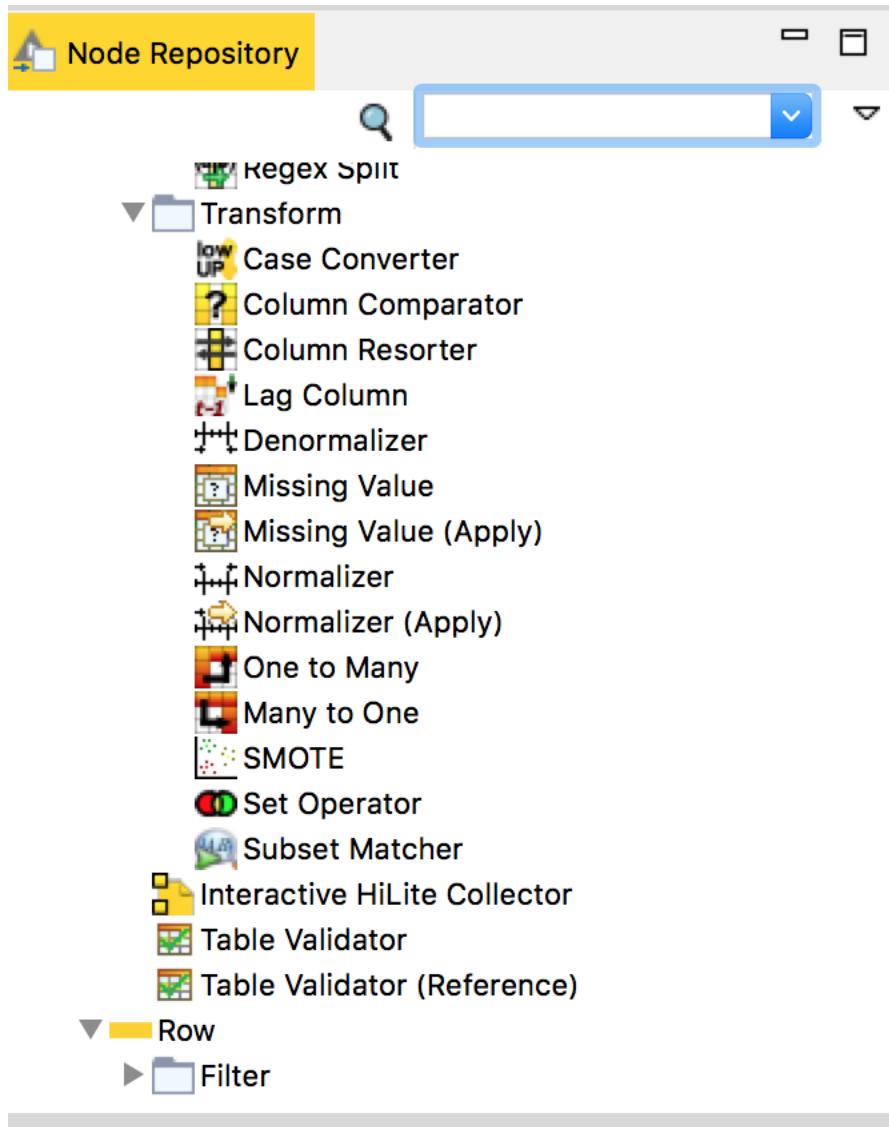
# Data : convert & replace



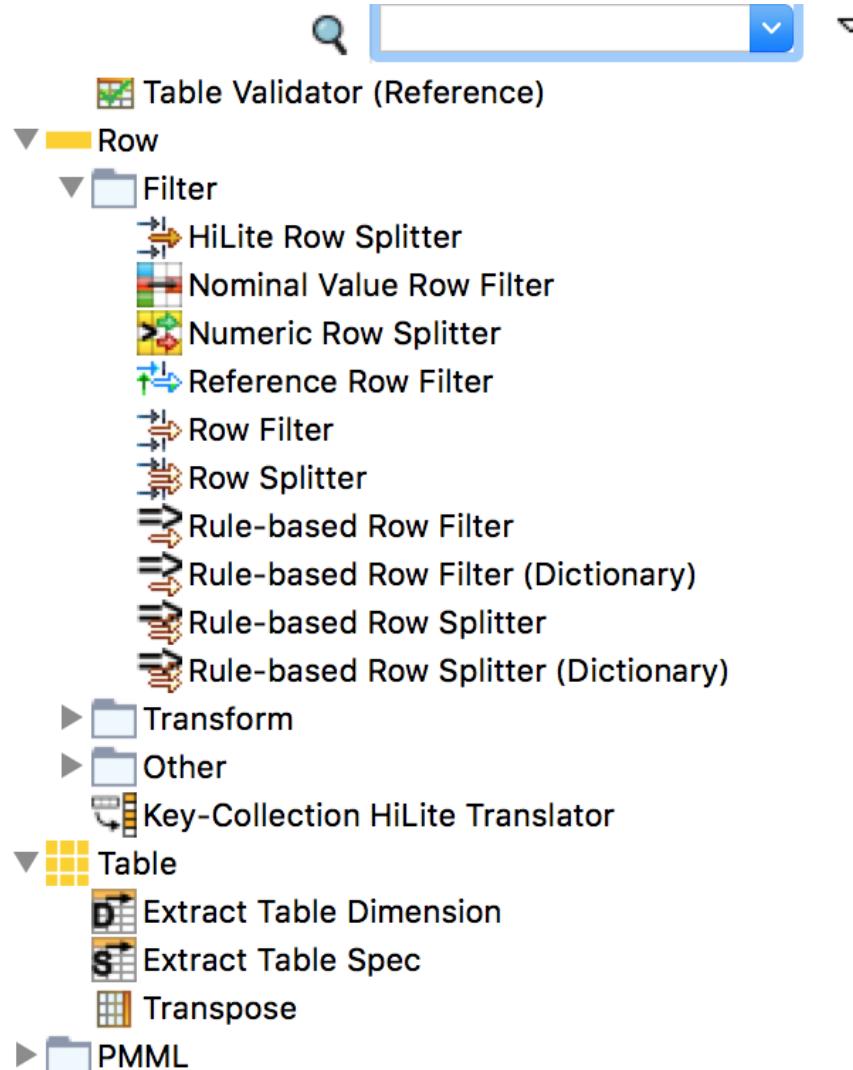
# Split & Combine data



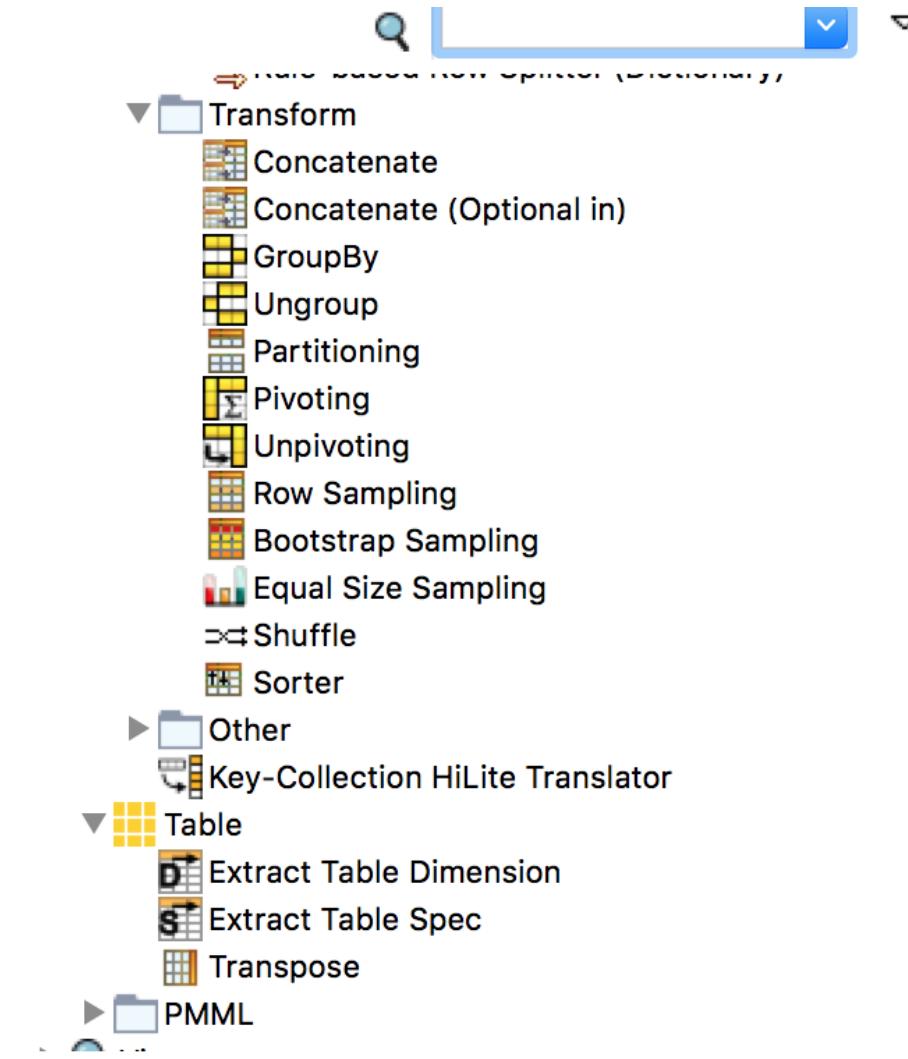
# Data transformation



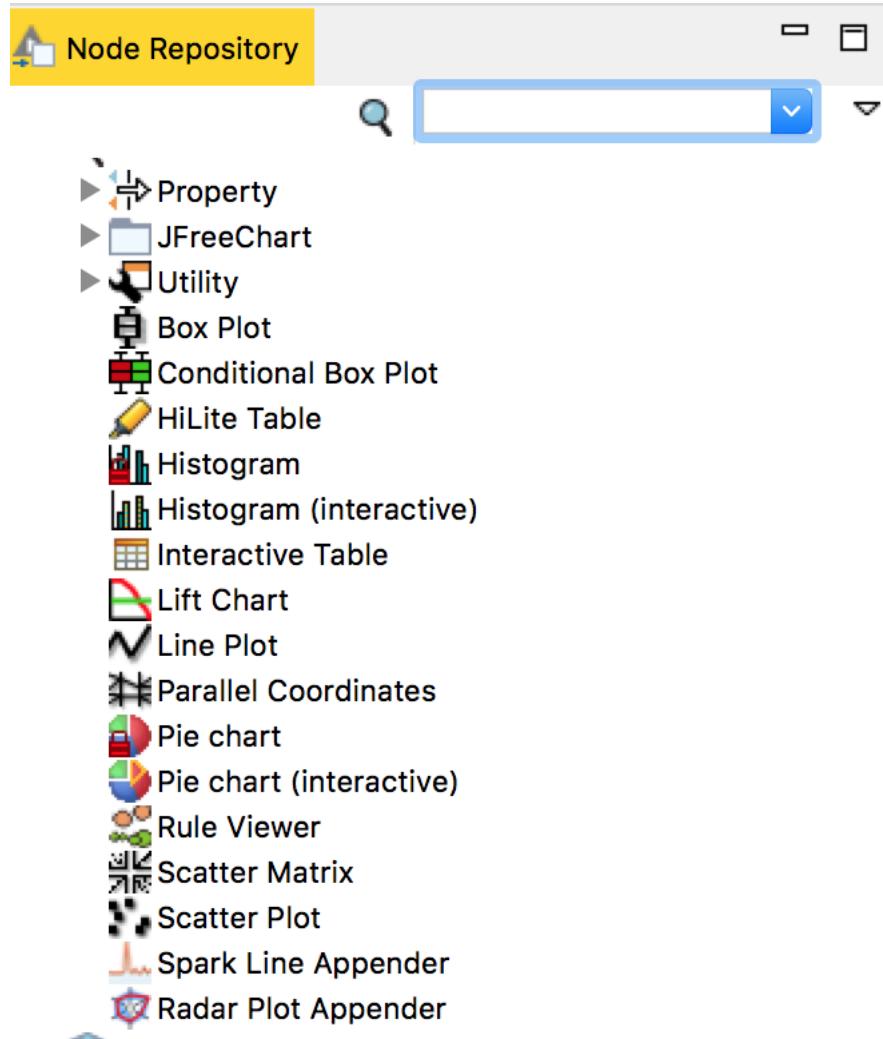
# Transform data : by Row



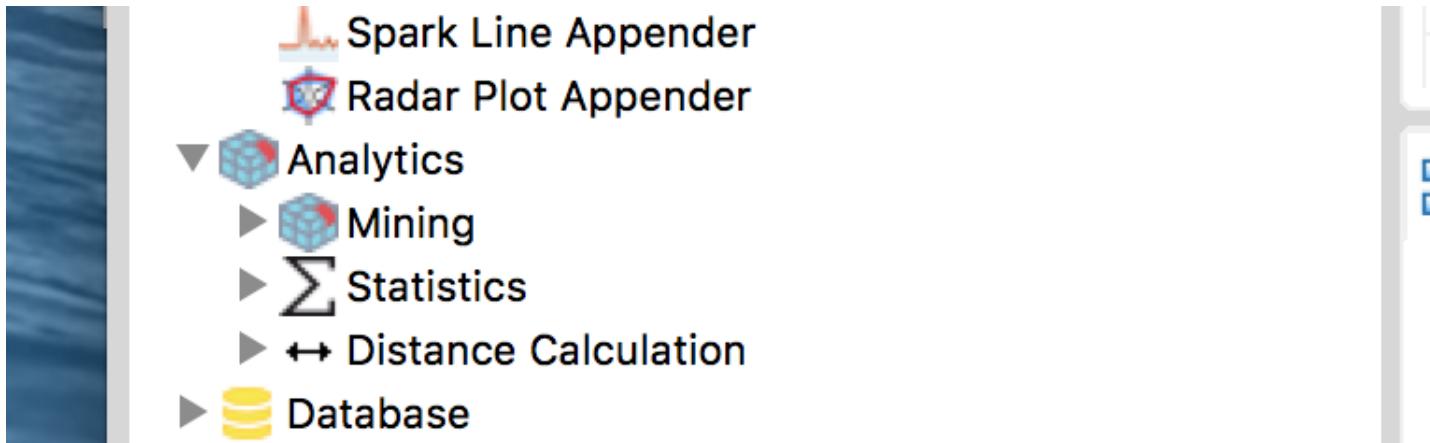
# Transform data: by Column



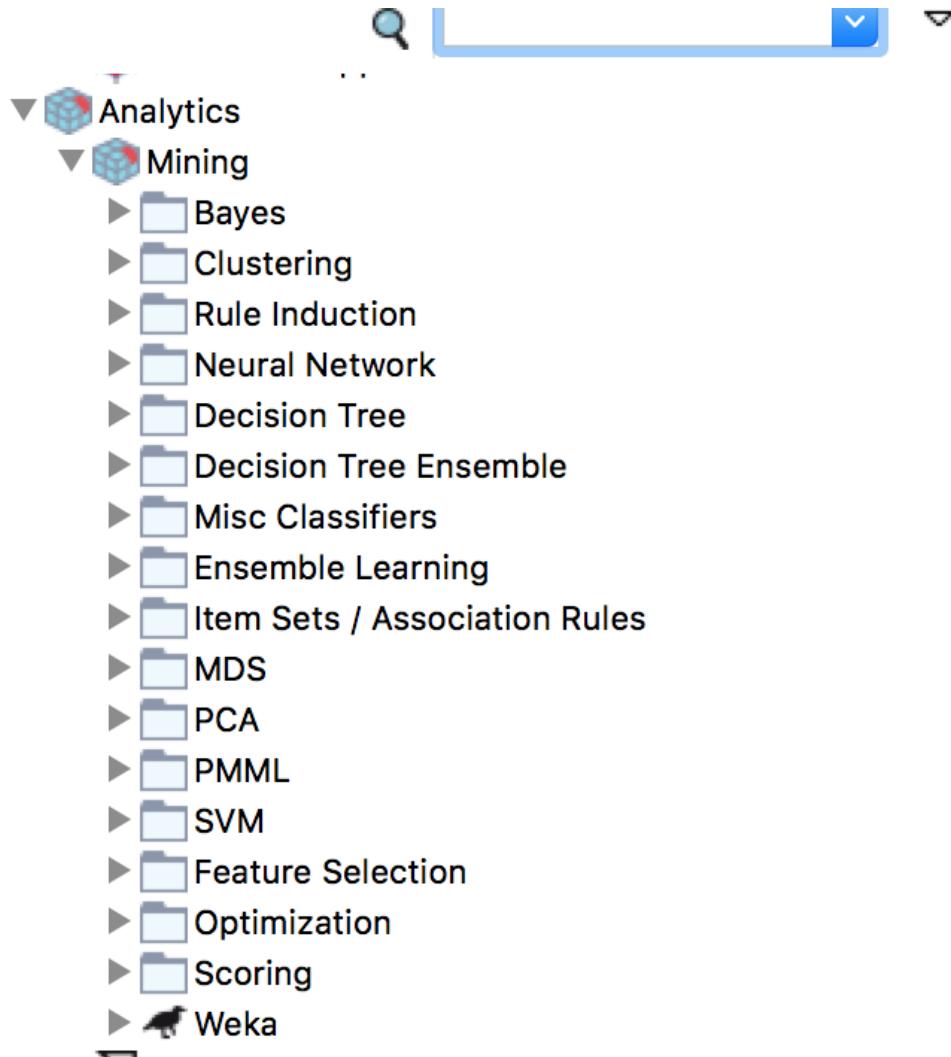
# Data visualization



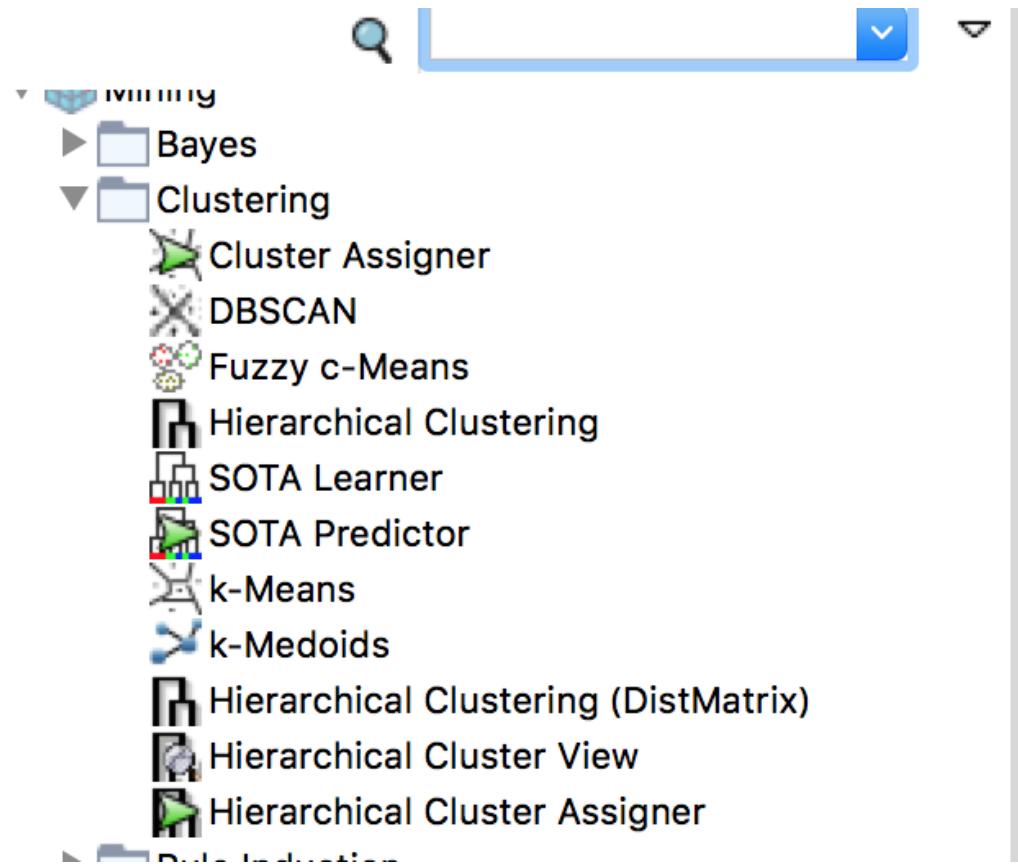
# Other functions



# Knime: Analytics & Data Mining



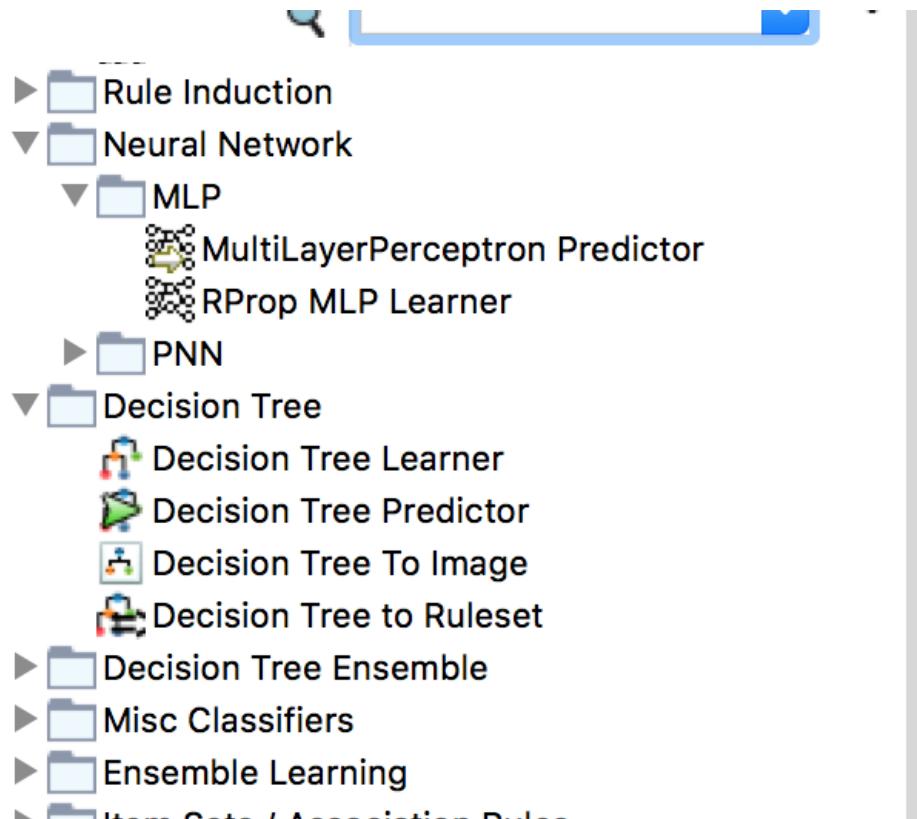
# Clustering data



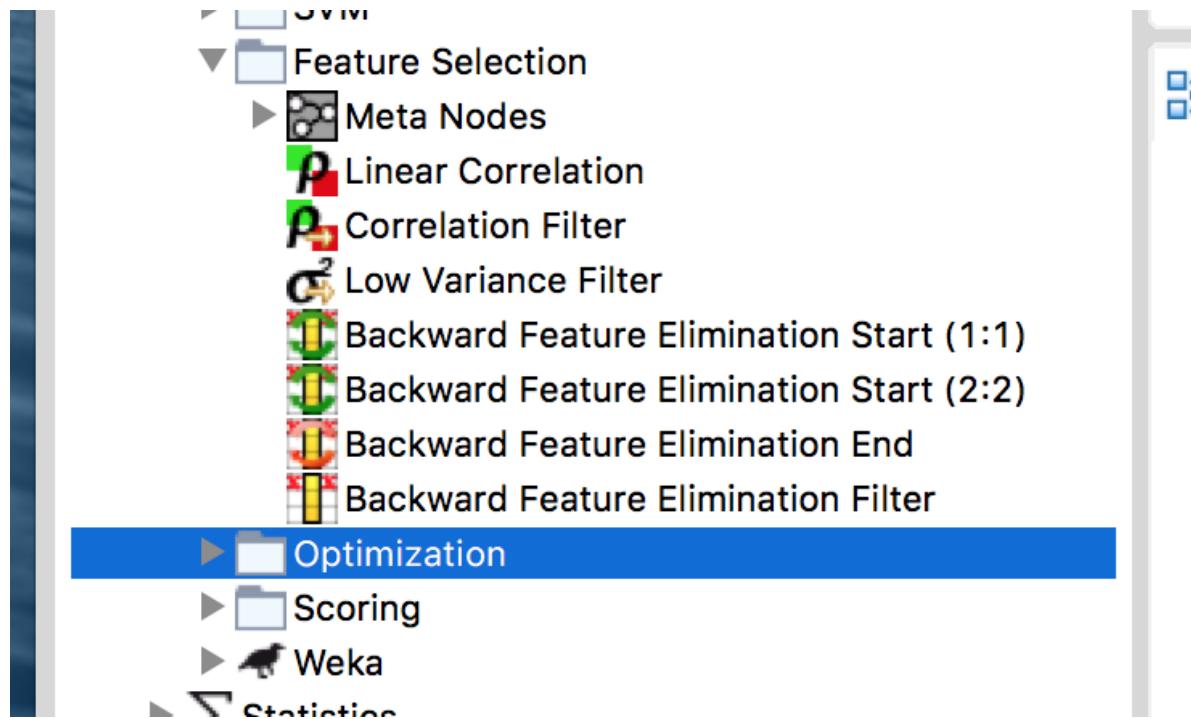
# Data projections methods

- ▶  Item Sets / Association Rules
- ▼  MDS
  -  MDS
  -  MDS (DistMatrix)
  -  MDS Projection
  -  MDS Projection (DistMatrix)
- ▼  PCA
  -  PCA
  -  PCA Compute
  -  PCA Apply
  -  PCA Inversion
- ▶  PMML
- ▶  CSV

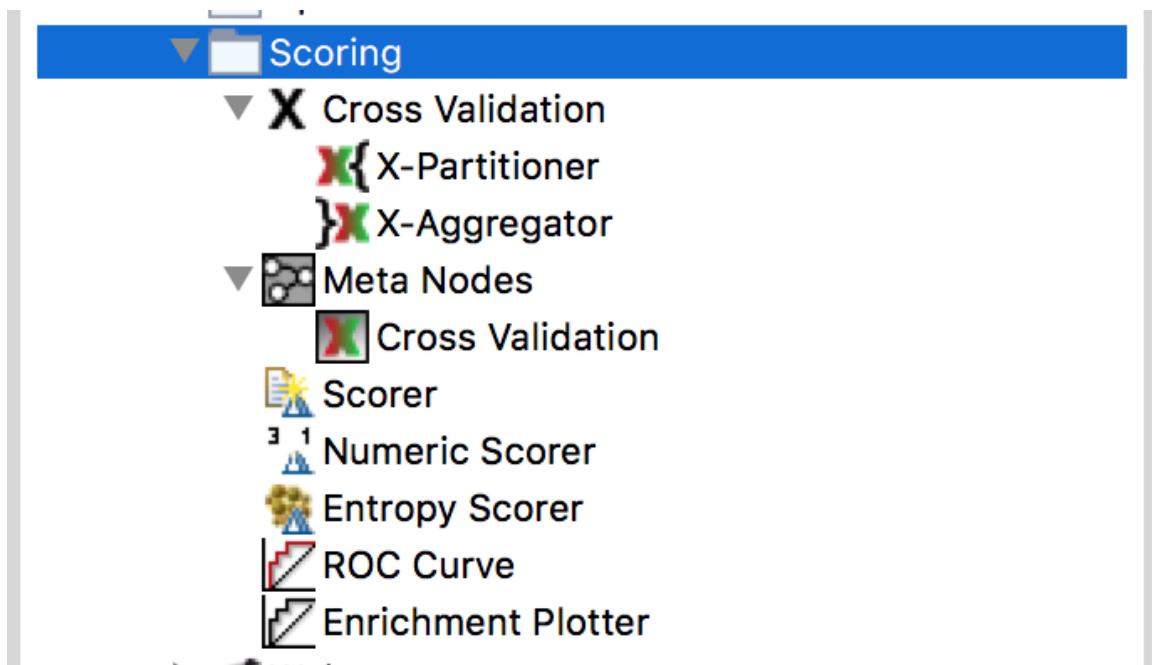
# Predictions models



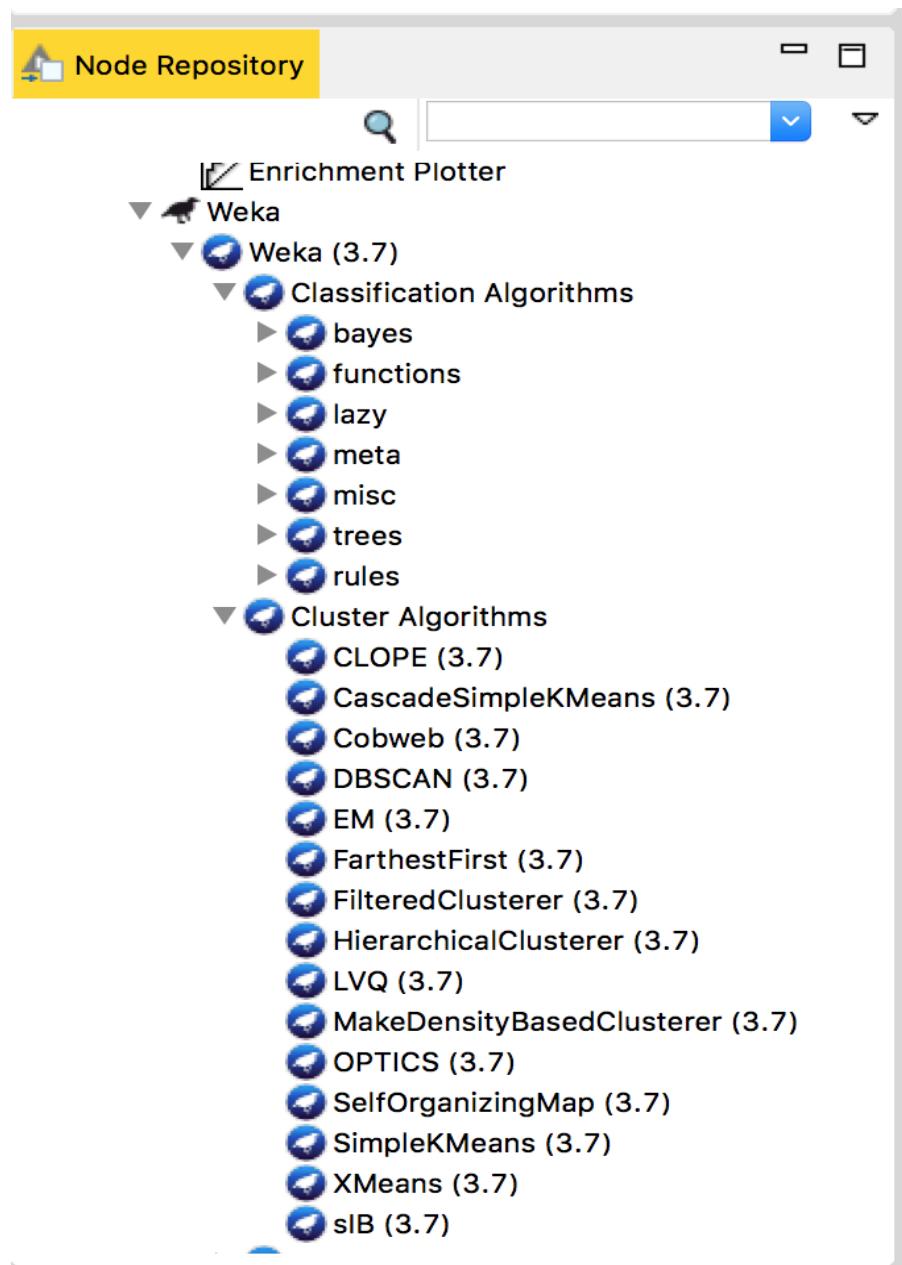
# Features selections approaches



# Validation & Scoring



# Weka



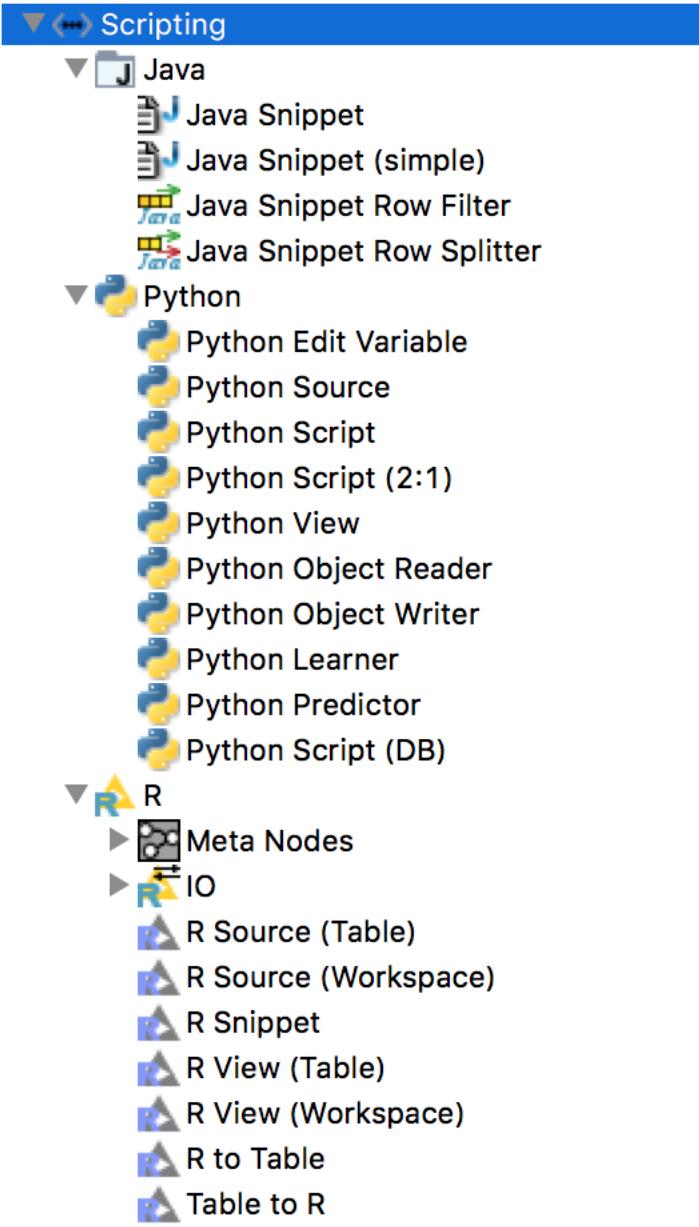
- ▼  Statistics
  -  Hypothesis Testing
  - ▼  Regression
    -  Linear Regression Learner
    -  Polynomial Regression Learner
    -  Logistic Regression Learner
    -  Regression Predictor
  -  Linear Correlation
  -  Statistics
  -  Crosstab
  -  Value Counter
- ▼  Distance Calculation
- ▼  Distance Functions
  -  Numeric Distances
  -  String Distances
  -  Bit Vector Distances
  -  Byte Vector Distances
  -  Mahalanobis Distance
  -  Matrix Distance
  -  Aggregated Distance
  -  Java Distance
- ▼  Distance Matrix
  -  Distance Matrix Reader
  -  Distance Matrix Writer
  -  Distance Matrix Calculate
  -  Distance Matrix Pair Extractor
  -  Similarity Search

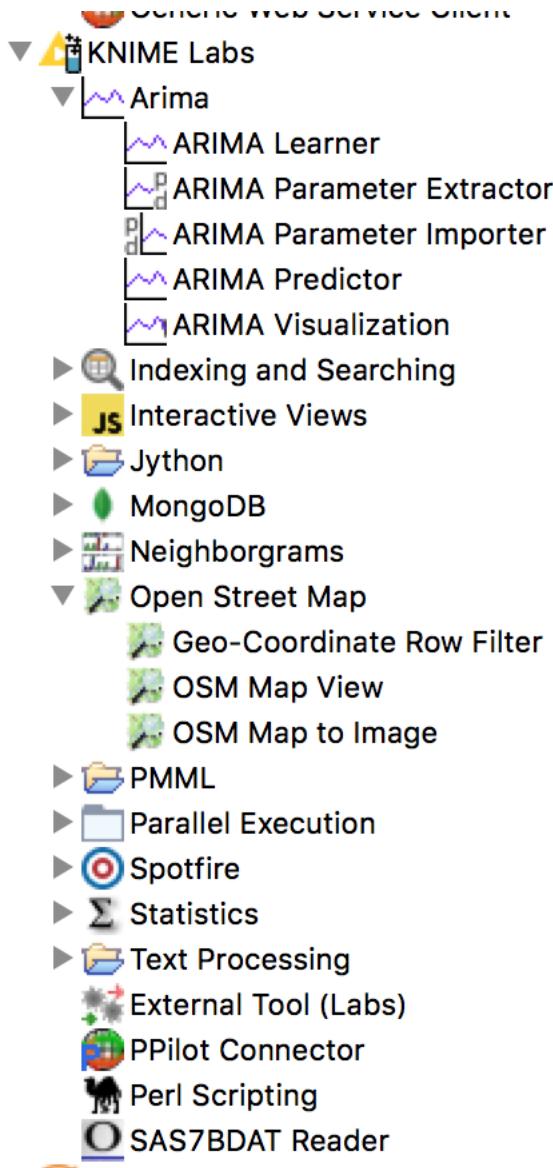
# Statistics

## Distance computing

## Distance matrix

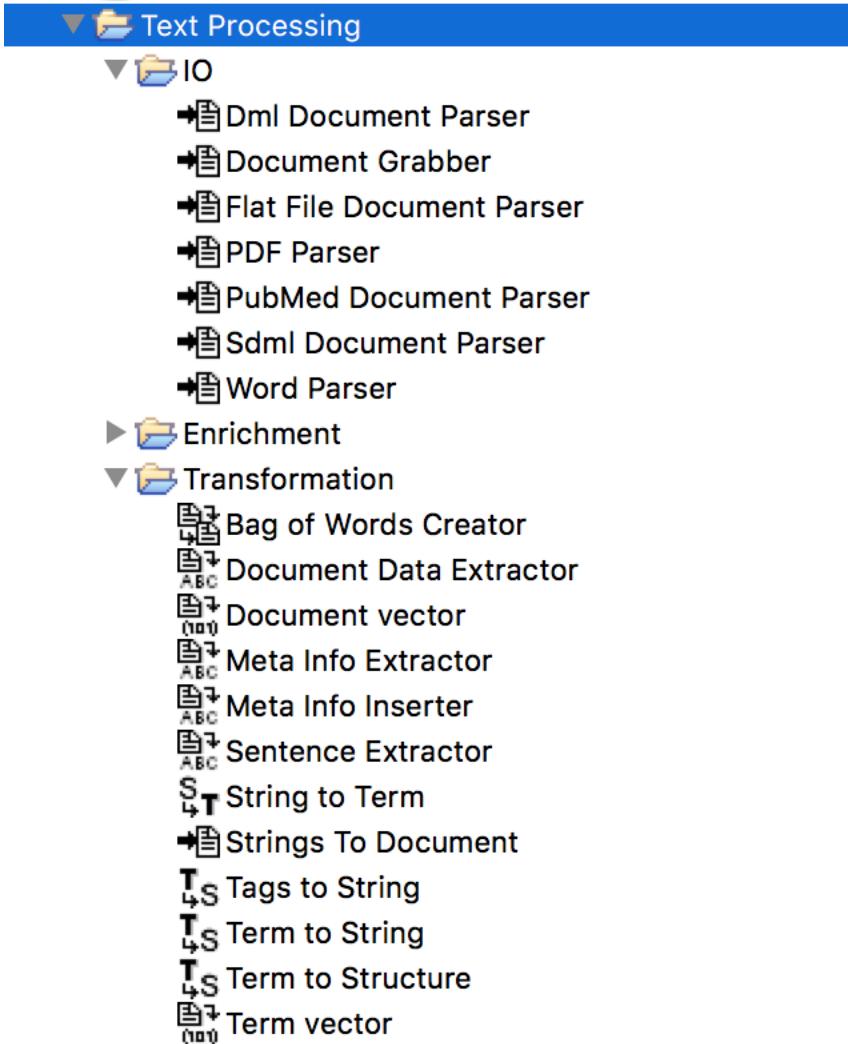
# Scripting





# Knime Labs

# Text Processing (1)



# Text Processing (2)

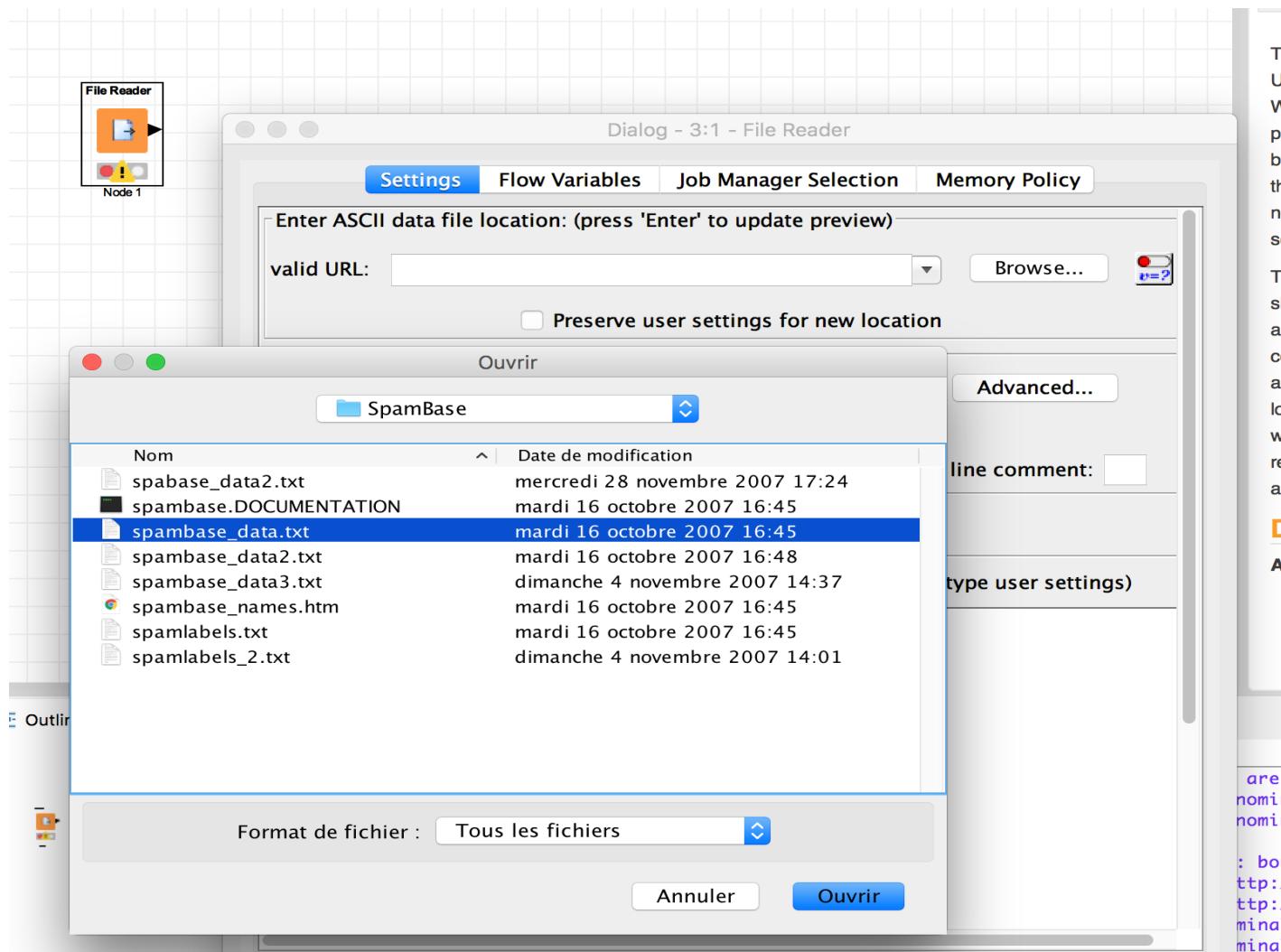
The screenshot shows a software interface with a navigation tree on the left side. The tree is organized into several main categories:

- Text Processing** (selected, highlighted in blue):
  - IO**:
    - Dml Document Parser
    - Document Grabber
    - Flat File Document Parser
    - PDF Parser
    - PubMed Document Parser
    - Sdml Document Parser
    - Word Parser
  - Enrichment**
  - Transformation**:
    - Bag of Words Creator
    - Document Data Extractor
    - Document vector
    - Meta Info Extractor
    - Meta Info Inserter
    - Sentence Extractor
    - String to Term
    - Strings To Document
    - Tags to String
    - Term to String
    - Term to Structure
    - Term vector
  - Preprocessing**
  - Frequencies**
  - Mining**
  - Misc**

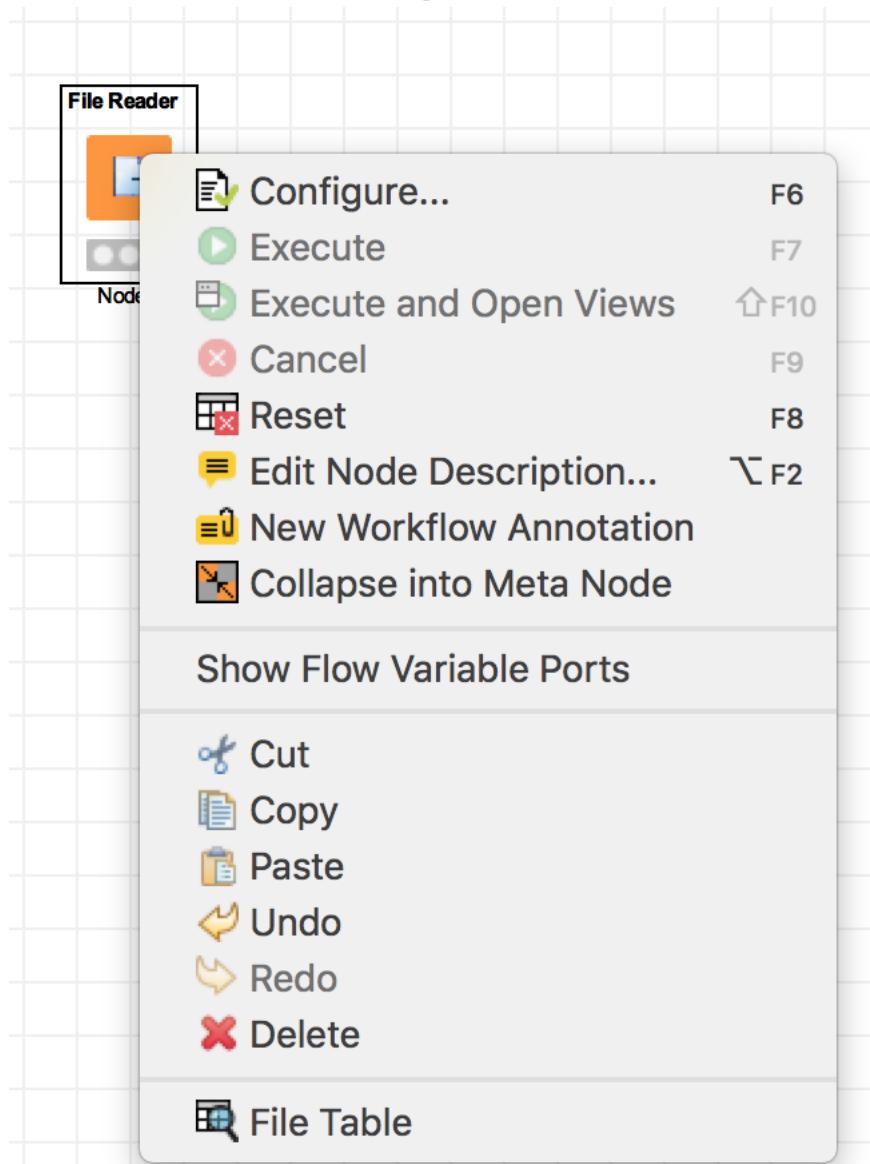
# Other utilities

-  External Tool (Labs)
-  PPilot Connector
-  Perl Scripting
-  SAS7BDAT Reader
-  Workflow Control
  -  Automation
    -  Wait...
    -  Save Workflow
    -  Timer Info
    -  Global Timer Info
  -  Quickforms
  -  Variables
  -  Loop Support
  -  Switches
  -  Error Handling
  -  Meta Nodes
-  Social Media
  -  Google API
  -  Twitter API
-  Reporting
  -  Data to Report
  -  Image to Report
-  Chemistry
  -  I/O
  -  Mining
  -  Misc
  -  Translators

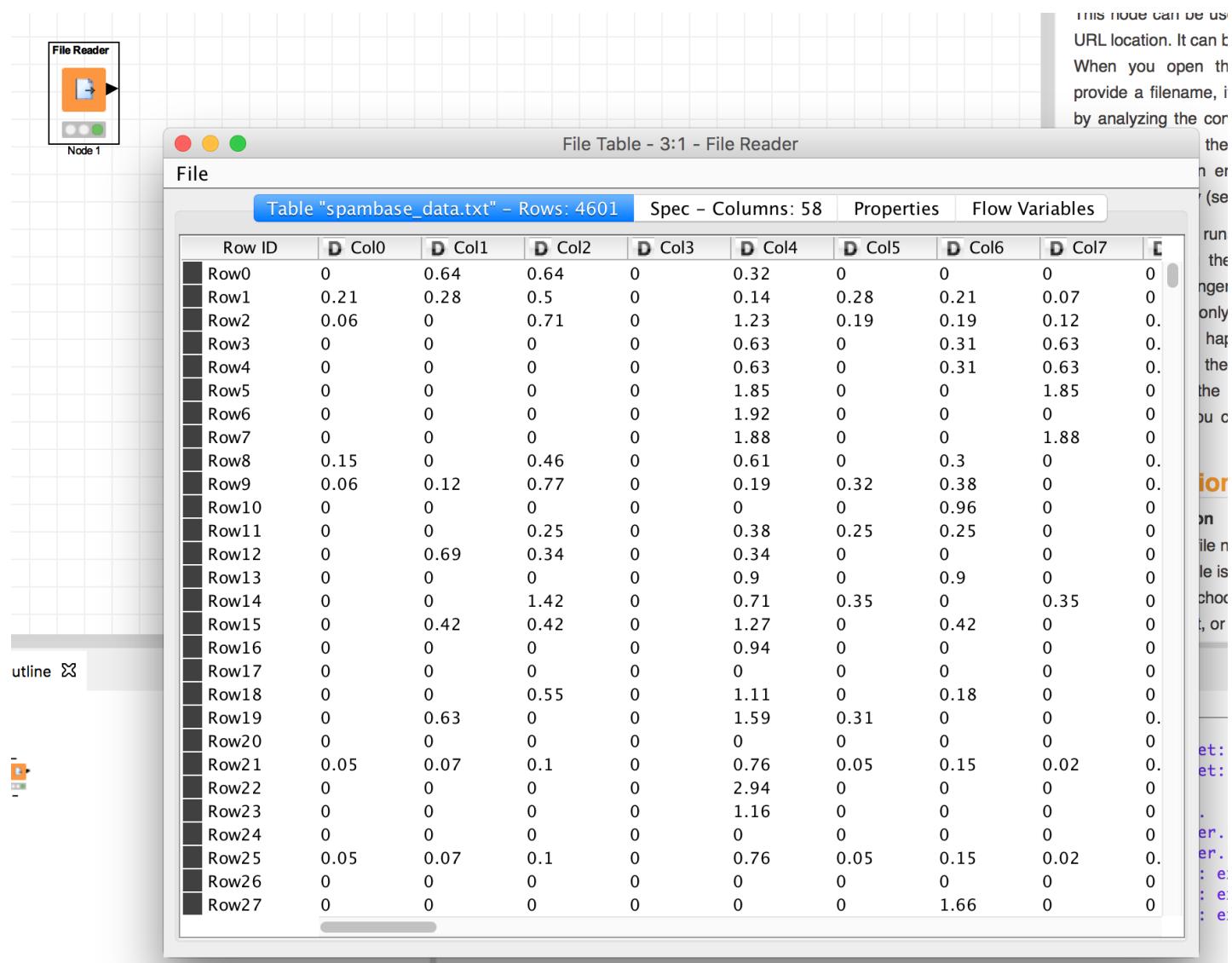
# File Reader



# File Reader configuration



# Table in File Reader



# Statistics view

Analysis  
Prediction

nDeployment  
nTraining

nodes

File

statistics

ing  
e t-test  
groups t-te

DVA

ession Learne  
regression L  
esson Lear  
redictor  
on

onbach Alp  
1  
tations

8:0 - Cross ...   Welcome to K...   \*2: KNIME\_pr...   \*0: ChurnPre...   \*3: KNIME\_pr...   16   Node Description

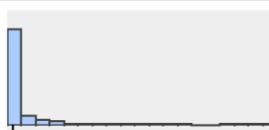
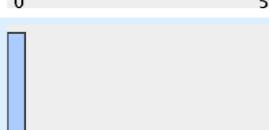
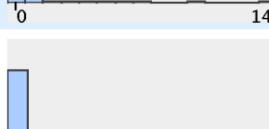
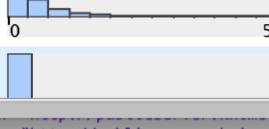
File Reader → Statistics

Statistics

This node calculates statistical moments such as minimum, maximum, mean, standard deviation, median, overall sum, number of missing values, count across all numeric columns, and occurrences of values together with their occurrences. The output is a table with histograms.

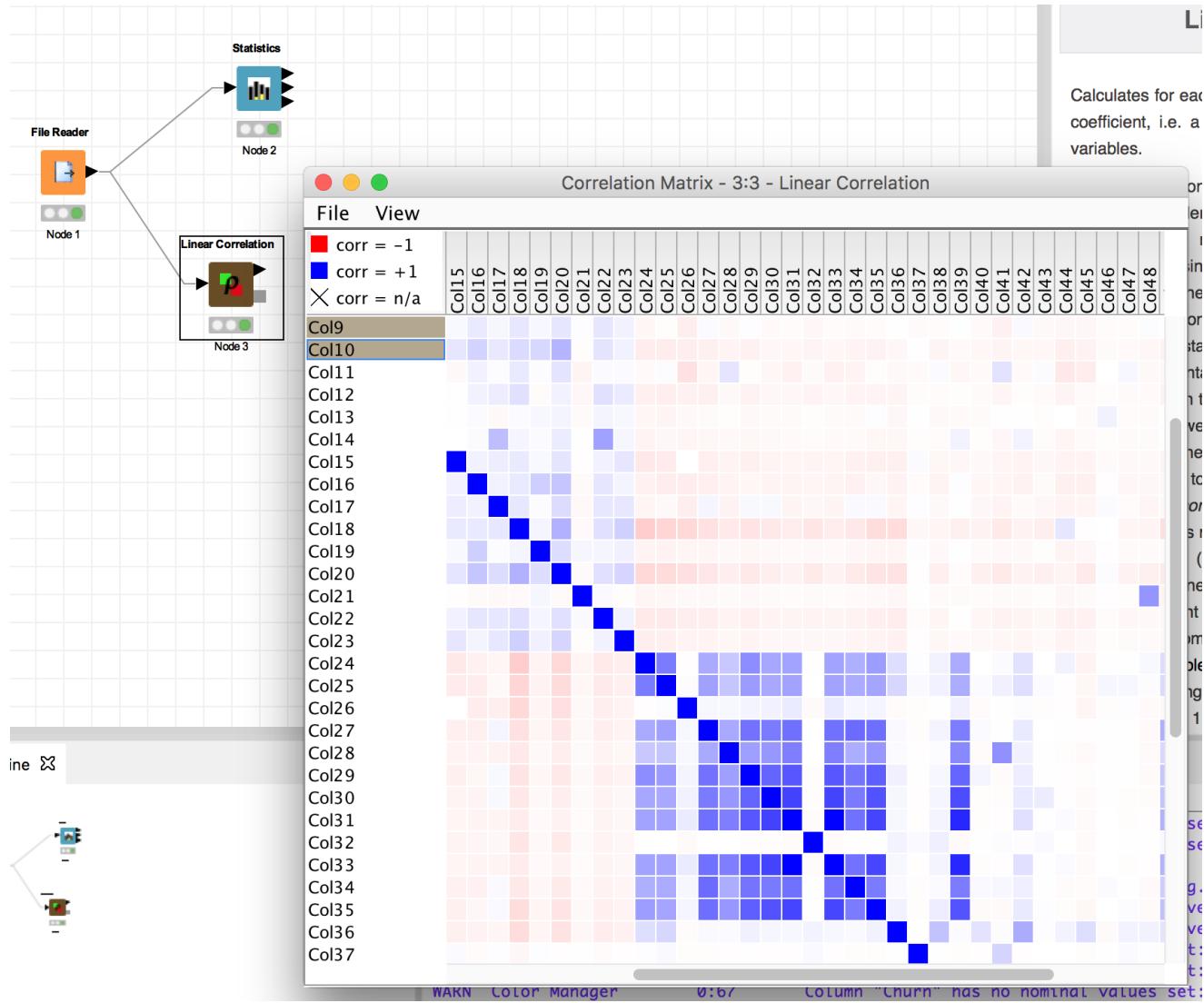
Statistics View - 3:2 - Statistics

Numeric   Nominal   Top/bottom

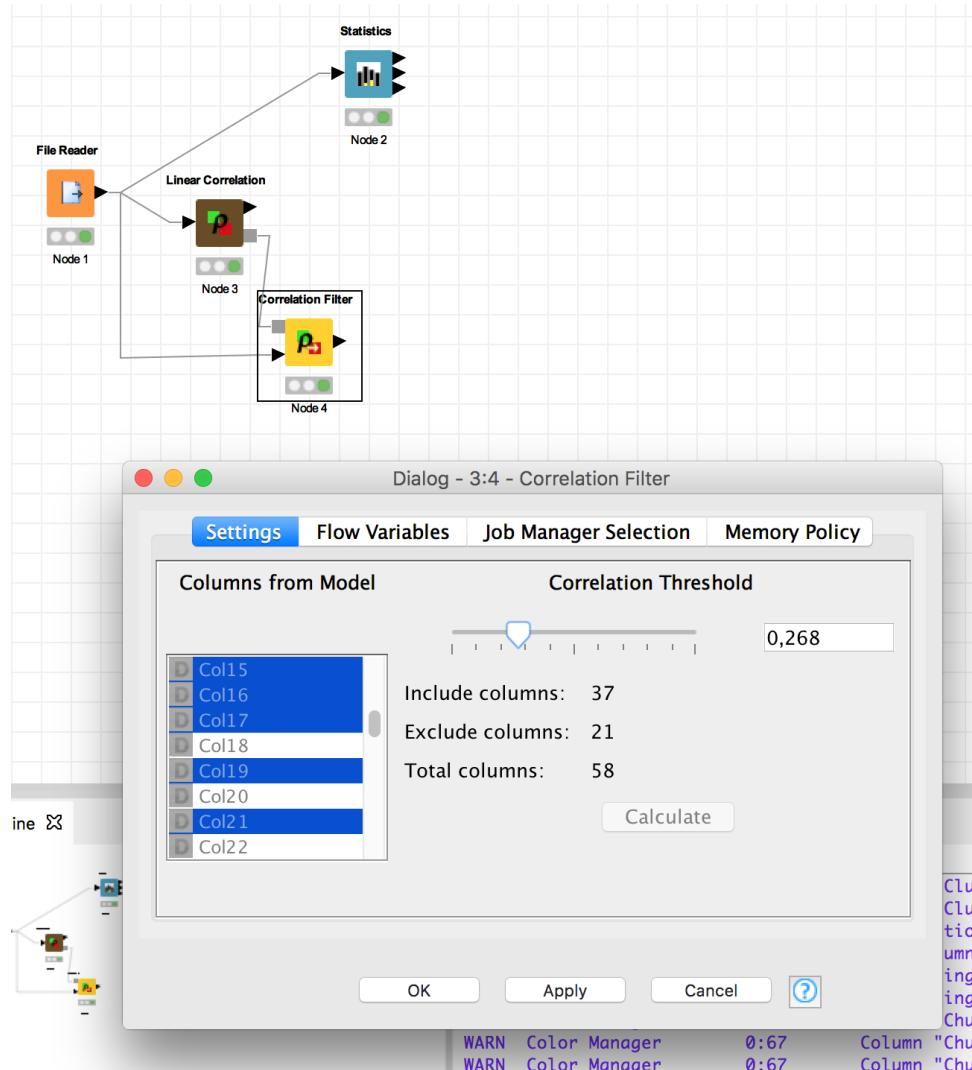
| Column | Min | Mean   | Median | Max   | Std. Dev. | Skewness | Kurtosis | No. Missing | No. +∞ | No. -∞ | Histogram   |
|--------|-----|--------|--------|-------|-----------|----------|----------|-------------|--------|--------|---|
| Col0   | 0.0 | 0.1046 | ?      | 4.54  | 0.3054    | 5.6756   | 49.3051  | 0           | 0      | 0      |    |
| Col1   | 0.0 | 0.213  | ?      | 14.28 | 1.2906    | 10.0868  | 105.6475 | 0           | 0      | 0      |   |
| Col2   | 0.0 | 0.2807 | ?      | 5.1   | 0.5041    | 3.0092   | 13.3087  | 0           | 0      | 0      |  |
| Col3   | 0.0 | 0.0654 | ?      | 42.81 | 1.3952    | 26.2277  | 726.4515 | 0           | 0      | 0      |  |

WARN KnimeRemoteFileSystem  
WARN Cell Manager  
WARN Cell Manager  
Connecting to server "http://publicserver.knime.org:80/tomcat" "Churn" has been loaded

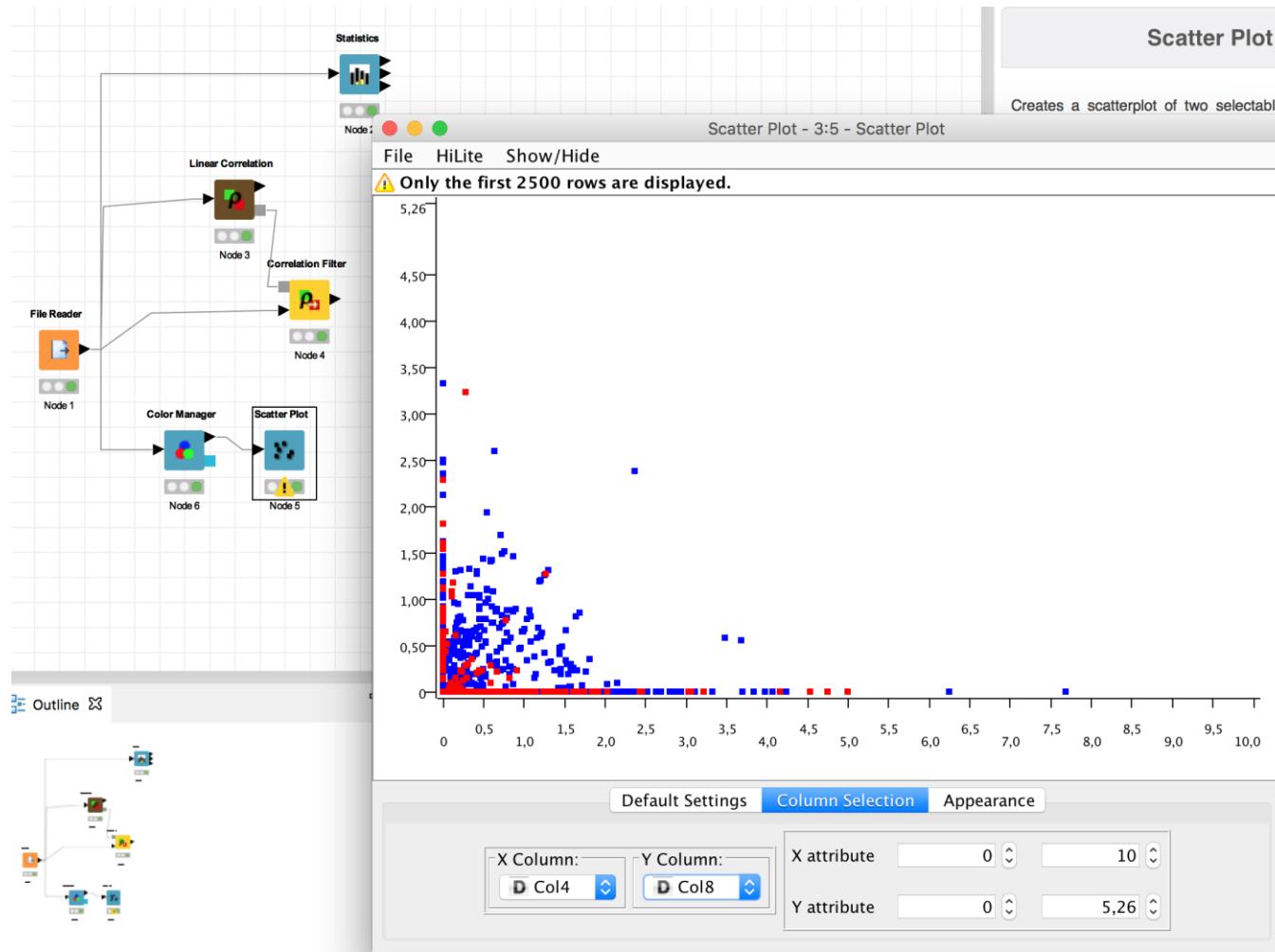
# Correlation view



# Correlation Filter



# Visualization: Scatter Plot



# Principal Component Analysis (PCA)

The screenshot shows a KNIME workflow interface. On the left, a process flow diagram illustrates the data pipeline. It starts with a 'File Reader' node (Node 1) reading data from a file. This data is then processed by a 'Color Manager' node (Node 2). Following this, a 'Linear Correlation' node (Node 3) is used to calculate correlations between variables. The output of Node 3 is then filtered by a 'Correlation Filter' node (Node 4). The resulting data is visualized using 'Scatter Plot' nodes (Nodes 5, 8, and 9). A 'PCA' node (Node 6) is also present in the flow, receiving input from the 'Color Manager' node and the 'Correlation Filter' node. The final output of the PCA node is a 'Statistics' node (Node 7), which displays the results of the principal component analysis.

**PCA Dialog - 3:7 - PCA**

This node performs a principal component analysis (PCA) on the given data. The input data is projected from its original feature space into a space of (possibly) lower dimensions.

**Options** Flow Variables Job Manager Selection Memory Policy

Fail if missing values are encountered (skipped per default)

Dimensions to reduce to

Minimum information fraction to preserve (%)

Replace original data columns

**Exclude** Column(s):  Search  Select all search hits

**Select**

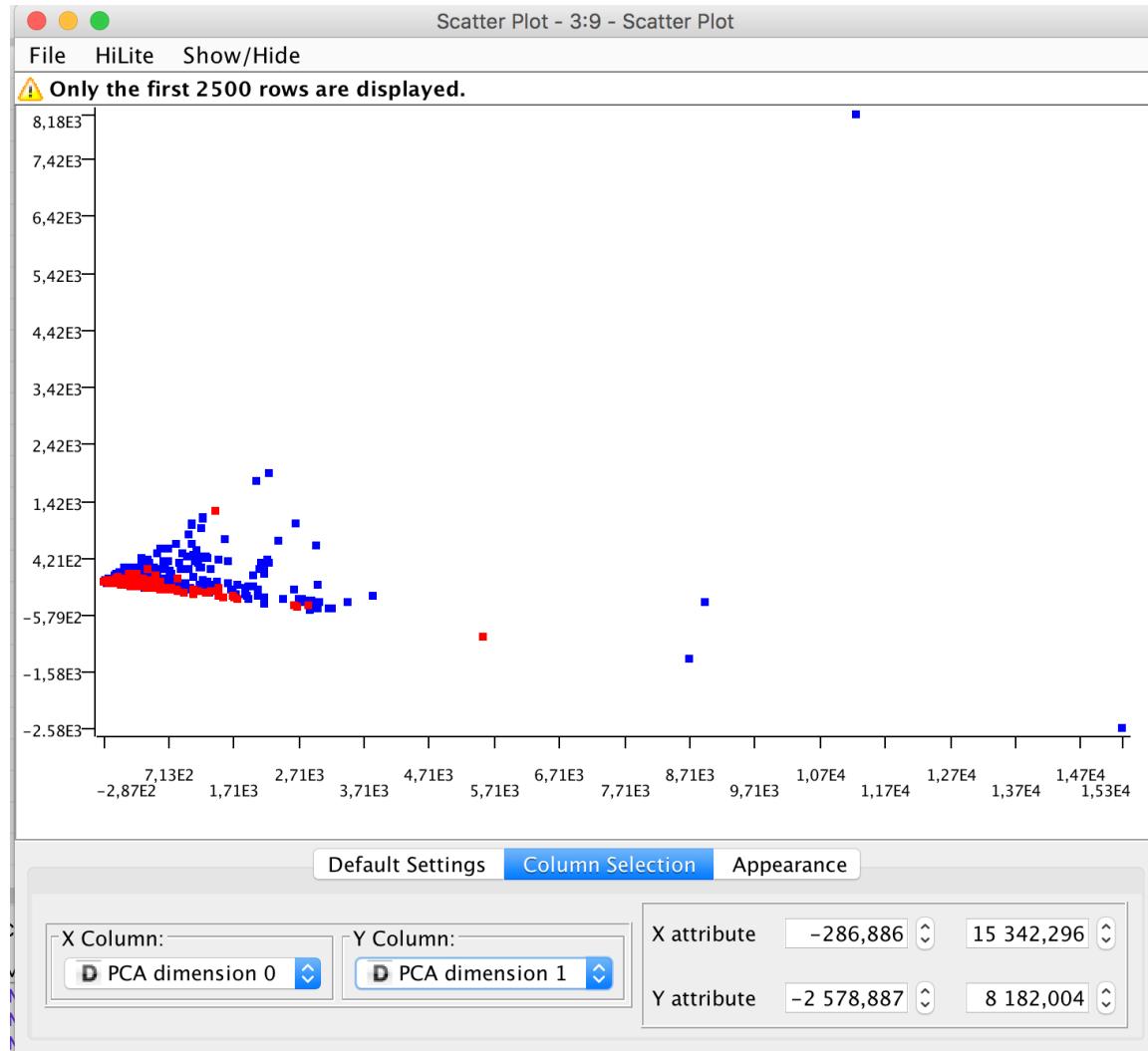
**Include** Column(s):  Search  Select all search hits

D Col0  
D Col1  
D Col2  
D Col3  
D Col4  
D Col5  
D Col6  
D Col7  
D Col8

OK Apply Cancel ?

WARN Color Manager 3/6 No column selected and no categorical column available

# Project the data using PCA



# Data normalization

File Reader → Normalizer

This node normalizes the values of all (numeric) columns. In the dialog, you can choose the columns you want to work on. The following normalization methods are available in the dialog:

**Dialog Options**

Methods Flow Variables Job Manager Selection Memory Policy

Manual Selection  Wildcard/Regex Selection

Exclude  Search  Select all search hits

Enforce exclusion

Select

Include  Search  Select all search hits

Enforce inclusion

D Col0  
D Col1  
D Col2  
D Col3  
D Col4  
D Col5  
D Col6  
D Col7

Settings

Min-Max Normalization      Min: 0.0  
 Z-Score Normalization (Gaussian)  
 Normalization by Decimal Scaling      Max: 1.0

OK Apply Cancel ?

# Data normalization

The image shows a data processing workflow and its resulting statistical summary.

**Flowchart:**

```
graph LR; A[File Reader] --> B[Normalizer]; B --> C[Statistics]
```

Nodes labeled:

- Node 11: File Reader
- Node 12: Normalizer
- Node 13: Statistics

**Statistics View - 3:13 - Statistics**

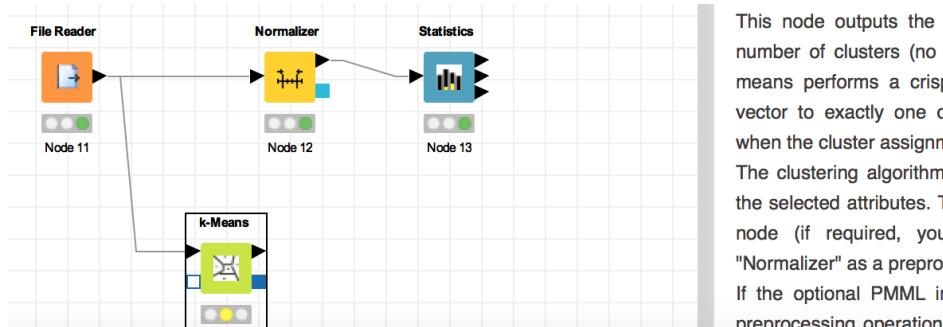
**File**

**Statistics View - 3:13 - Statistics**

**Numeric** **Nominal** **Top/bottom**

| Column | Min     | Mean      | Median | Max     | Std. Dev. | Skewness | Kurtosis | N |
|--------|---------|-----------|--------|---------|-----------|----------|----------|---|
| Col0   | -0,3424 | -2,47E-16 | ?      | 14,5254 | 1         | 5,6756   | 49,3051  |   |
| Col1   | -0,1651 | 1,92E-16  | ?      | 10,8998 | 1         | 10,0868  | 105,6475 |   |
| Col2   | -0,5567 | 1,80E-15  | ?      | 9,5595  | 1         | 3,0092   | 13,3087  |   |
| Col3   | -0,0469 | -4,49E-17 | ?      | 30,6379 | 1         | 26,2277  | 726,4515 |   |
| Col4   | -0,4643 | 9,40E-17  | ?      | 14,4053 | 1         | 4,7471   | 37,9412  |   |

# Clustering



This node outputs the number of clusters (no means performs a crisp vector to exactly one cluster when the cluster assignment). The clustering algorithm uses the selected attributes. This node (if required, you can use "Normalizer" as a preprocessing operation). If the optional PMML is specified, it performs preprocessing operations.

Dialog - 3:14 - k-Means

K-Means Properties Flow Variables Job Manager Selection Memory Policy

number of clusters:

max. number of iterations:

Search hits

Select

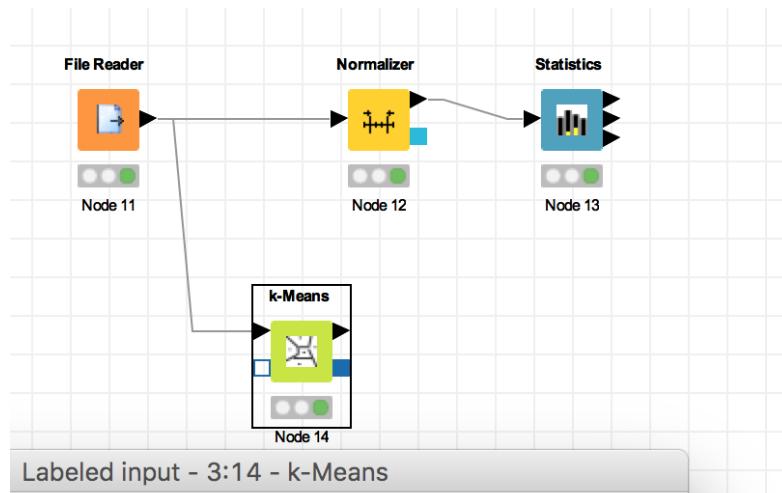
Include

Column(s):

D Col0  
D Col1  
D Col2  
D Col3  
D Col4  
D Col5  
D Col6  
D Col7  
D Col8

OK Apply Cancel

# Clustering



Labeled input - 3:14 - k-Means

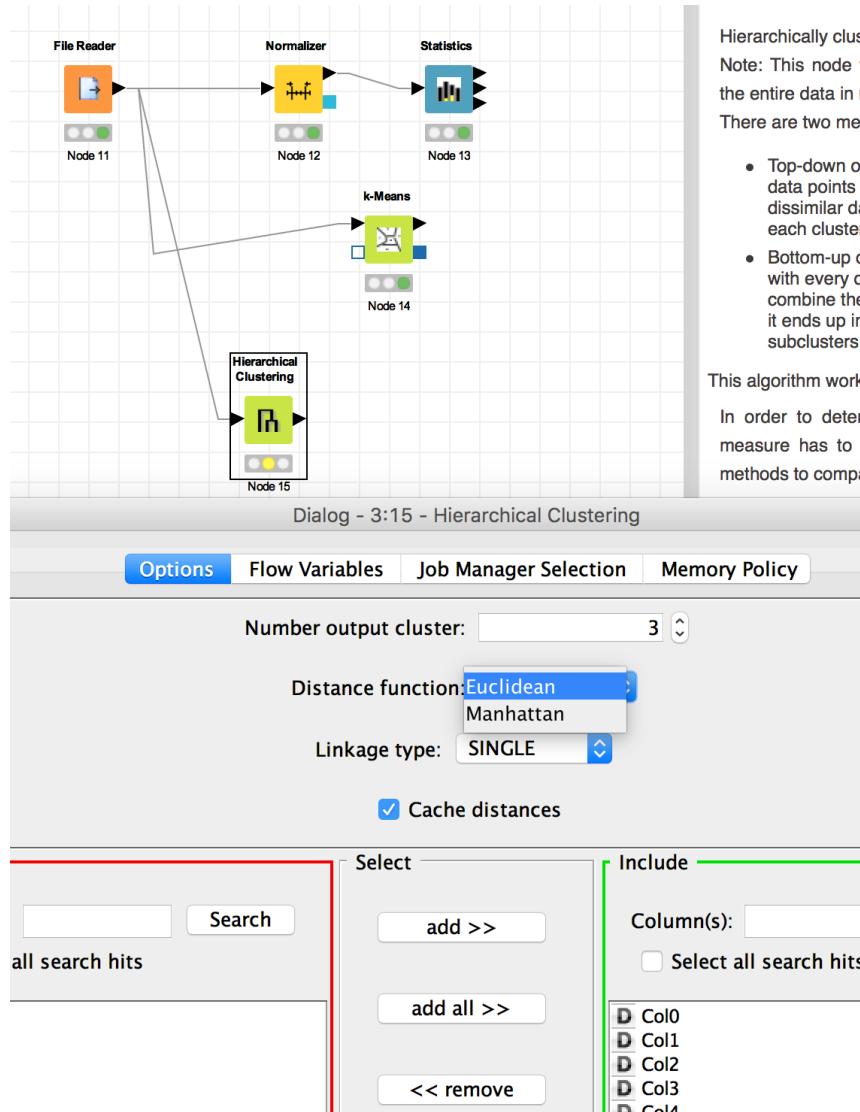
|   | Col55 | Col56 | Col57 | Cluster   |
|---|-------|-------|-------|-----------|
| 4 | 61    | 278   | 1     | cluster_0 |
|   | 101   | 1028  | 1     | cluster_1 |
|   | 485   | 2259  | 1     | cluster_1 |
|   | 40    | 191   | 1     | cluster_0 |
|   | 40    | 191   | 1     | cluster_0 |
|   | 15    | 54    | 1     | cluster_0 |
|   | 4     | 112   | 1     | cluster_0 |
|   | 11    | 49    | 1     | cluster_0 |
|   | 445   | 1257  | 1     | cluster_1 |
|   | 43    | 749   | 1     | cluster_1 |
|   | 6     | 21    | 1     | cluster_0 |
|   | 11    | 184   | 1     | cluster_0 |
|   | 61    | 261   | 1     | cluster_0 |
|   | 7     | 25    | 1     | cluster_0 |
|   | 24    | 205   | 1     | cluster_0 |
|   | 55    | 249   | 1     | cluster_0 |

'spambase\_data.txt' - Rows: 4601

urn" has  
s availa  
guration  
selected  
irst 250  
efault a

WARN Scatter Plot 3:9 Some columns are i

# Clustering with Hierarchical clustering



Hierarchically clust

Note: This node v  
the entire data in n  
There are two met

- Top-down or  
data points i  
dissimilar da  
each cluster
- Bottom-up o  
with every di  
combine the  
it ends up in  
subclusters.

This algorithm work

In order to deter  
measure has to b  
methods to compa

# Machine Learning (1)

Dialog - 3:19 - Partitioning

First partition Flow Variables ►

Choose size of first partition

Absolute 80

Relative[%] 10

Take from top

Linear sampling

Draw randomly

Stratified sampling

Use random seed 1 490 712 151 187

OK Apply Cancel ?

Node 14

File Reader Node 16 → Partitioning Node 19 → Decision Tree Learner Node 17

Decision Tree Predictor Node 18

Reader 3:1 No Settings  
Correlation 3:3 Auto config  
Manager 3:6 No column se  
Plot 3:5 Only the fi  
Plot 3:7 using as de  
Plot 3:9 Some column  
Plot 3:9 Some column  
Plot 3:9 Only the fi  
Normaliz... 3:12 Auto config

```
graph LR; FR[File Reader Node 16] --> P[Partitioning Node 19]; P --> DTL[Decision Tree Learner Node 17]; P --> DTP[Decision Tree Predictor Node 18]
```

# Machine Learning (2)

Dialog - 3:16 - File Reader

Settings Flow Variables Job Manager Selection Memory Policy

Enter ASCII data file location: (press 'Enter' to update preview)

valid URL: file:/Users/nistor/Documents/NistorMAC/podium\_

Preserve user settings for new location

**Basic Settings**

read row IDs Column delimiter: ,   
 read column headers  ignore spaces and tabs

use style comments  single line comment:

**Column Properties**

New settings for column 'Col57'

DON'T include column in output table

Name: Col57

Type: Number (integer)

Number (double precision)  
Number (integer)  
Number (long)  
Smarts  
Smiles  
String  
URI

Preview

Click column

| Row ID | 51    |
|--------|-------|
| Row0   |       |
| Row1   |       |
| Row2   |       |
| Row3   |       |
| Row4   |       |
| Row5   |       |
| Row6   |       |
| Row7   |       |
| Row8   |       |
| Row9   |       |
| Row10  |       |
| Row11  |       |
| Row12  |       |
| Row13  |       |
| Row14  |       |
| Row15  | 0.063 |

| 5    | Col56 | Col57 |
|------|-------|-------|
| 278  | 1     |       |
| 1028 | 1     |       |
| 2259 | 1     |       |
| 191  | 1     |       |
| 191  | 1     |       |
| 54   | 1     |       |
| 112  | 1     |       |
| 49   | 1     |       |
| 1257 | 1     |       |
| 749  | 1     |       |
| 21   | 1     |       |
| 184  | 1     |       |
| 261  | 1     |       |
| 25   | 1     |       |
| 205  | 1     |       |
| 249  | 1     |       |

# Machine Learning (3)

Dialog - 3:17 - Decision Tree Learner

**Options**   **PMMLSettings**   **Flow Variables** ►

**General**

Class column: Col57

Quality measure: Gini index

Pruning method: No pruning

Reduced Error Pruning

Min number records per node: 2

Number records to store for view: 10 000

Average split point

Number threads: 4

Skip nominal columns without domain information

**Binary nominal splits**

Binary nominal splits

Max #nominal: 10

Filter invalid attribute values in child nodes

OK   Apply   Cancel   ?

Statistics → Node 13

k-Means → Node 14

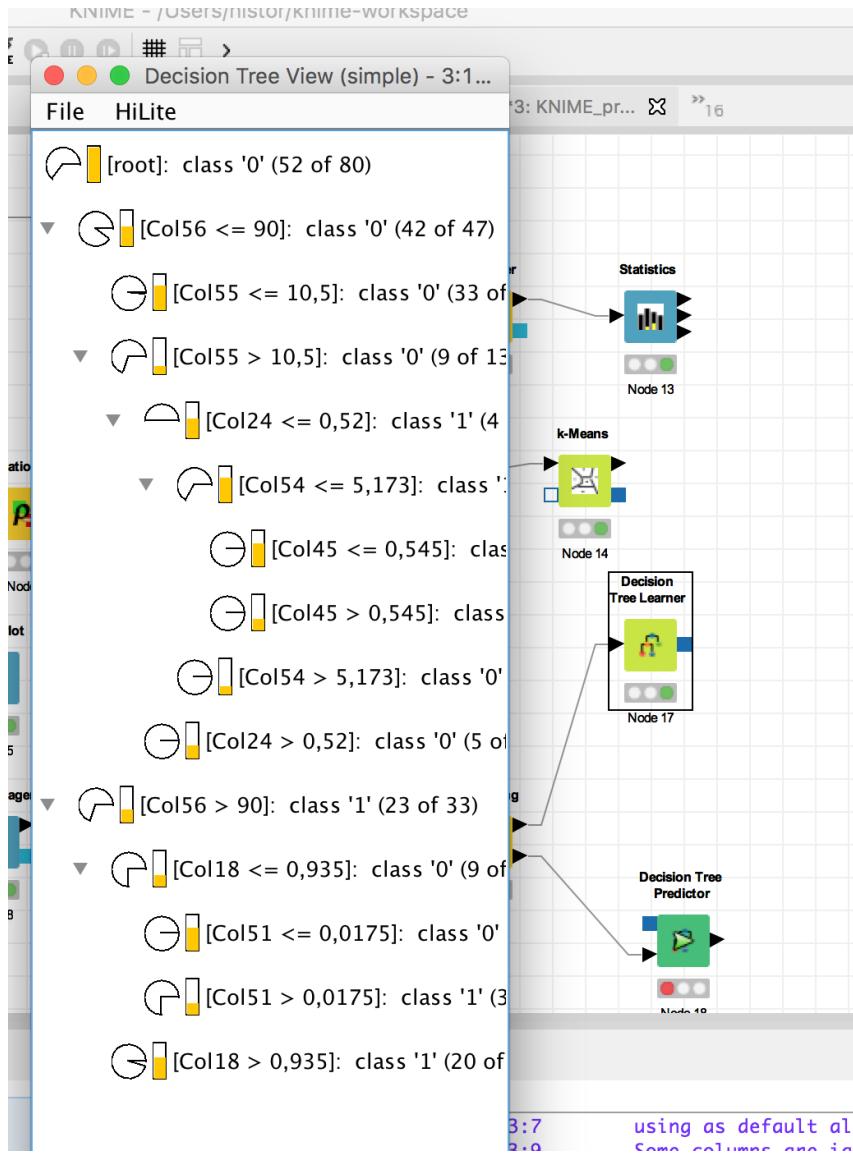
Decision Tree Learner → Node 17

Decision Tree Predictor →

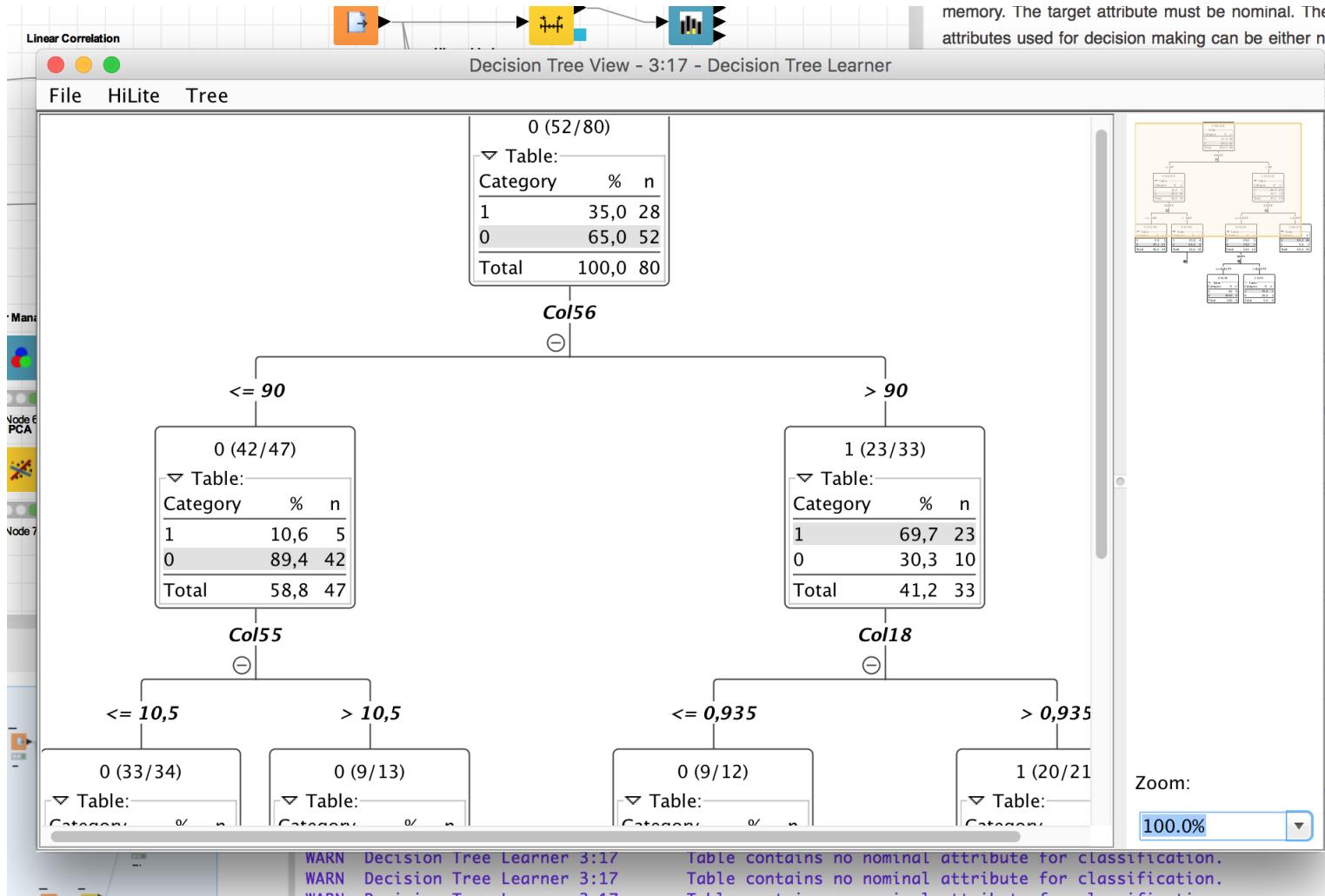
using as default  
Some columns are  
Some columns are  
Only the first 2  
Auto-configure:  
Potential deadlo  
No sampling meth  
Table contains  
Table contains  
Table contains

WARN Decision Tree Learner 3:17  
WARN Decision Tree Learner 3:17  
WARN Decision Tree Learner 3:17

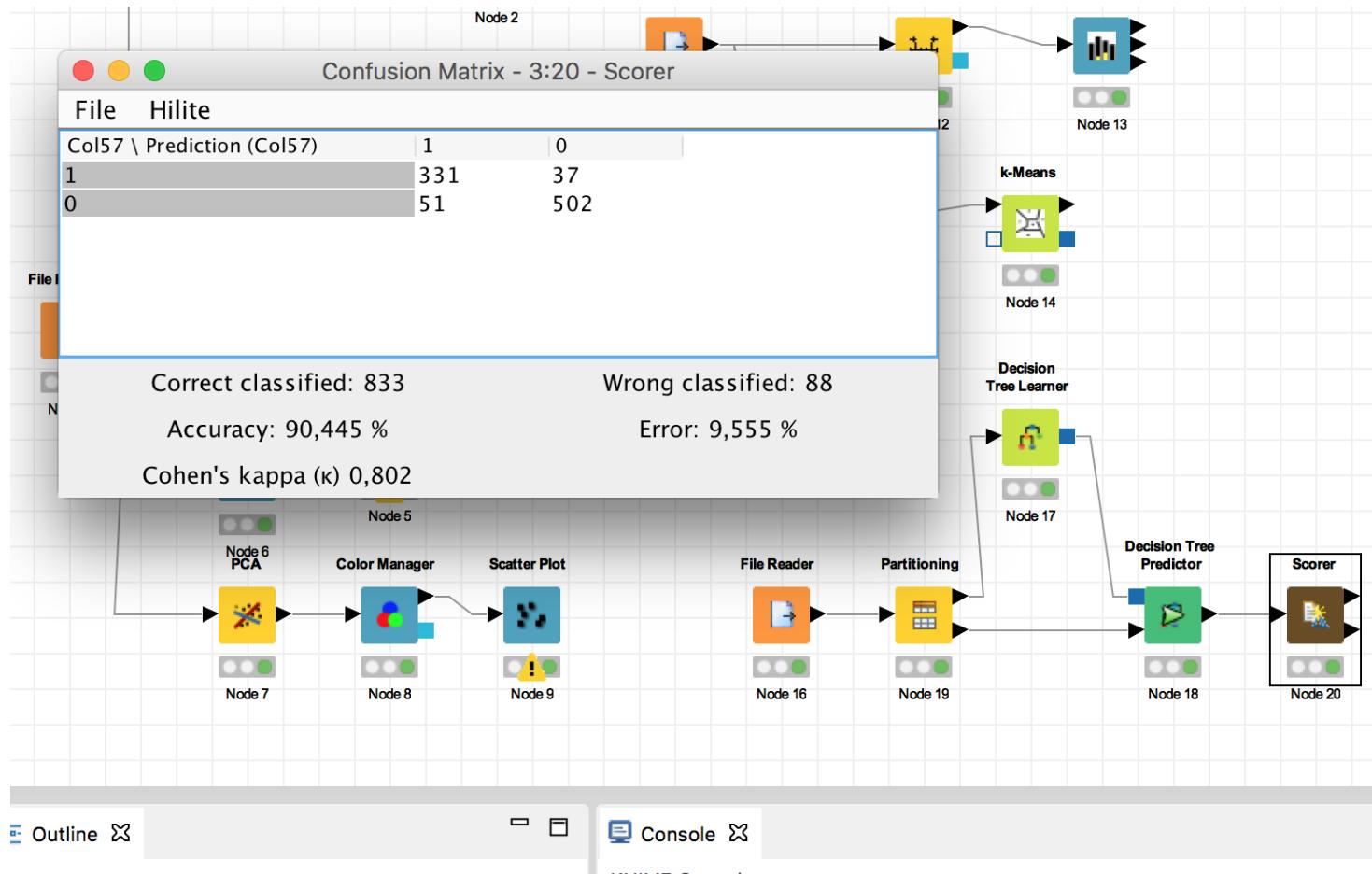
# Machine Learning (4)



# Machine Learning (5)



# Machine Learning (6)



shows the attribute possible underlying columns column and the matrix's matrix. Additionally, accuracy Positive Precision the overall

## Detailed View

### First column

The data

### Second column

The of th

# Machine Learning (7)

