

# Data Mining, BI, Predictive Analysis, Visualization

with KNIME

# KNIME : create workflow

KNIME - /Users/nistor/knime-workspace

Quick Access

**KNIME Explorer**

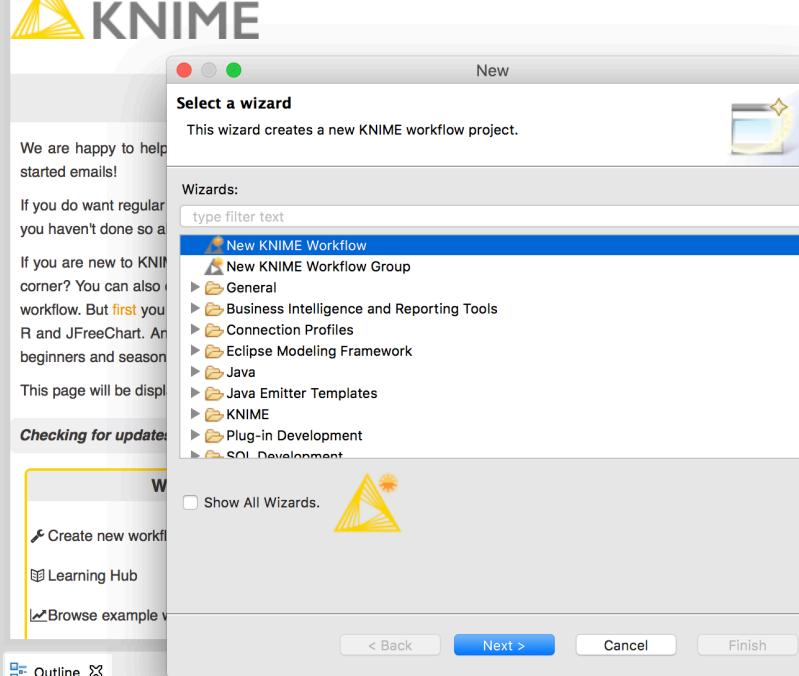
- 02\_SocialNetworkAnalysis
- 050018\_ChurnPrediction
- data
  - ChurnPredictionDeployment
  - ChurnPredictionTraining
- 050019\_TwitterAnalysis
- 002005\_PMMI\_Examples
- 002006\_PMMI\_Ensembles\_blog
- 050004\_lastfm\_Recommendations
- 050005\_Social\_Media\_Clustering
- Dim Reduction Techniques
- 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



## 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.

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 KntimeRemoteFileSystem
WARN KntimeRemoteFileSystem
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 Binary Node.
Column "Cluster" has no nominal values set: execute predecessor or add Binary Node.
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 Binary Node.
Column "Churn" has no nominal values set: execute predecessor or add Binary Node.
Column "Churn" has no nominal values set: execute predecessor or add Binary Node.
```

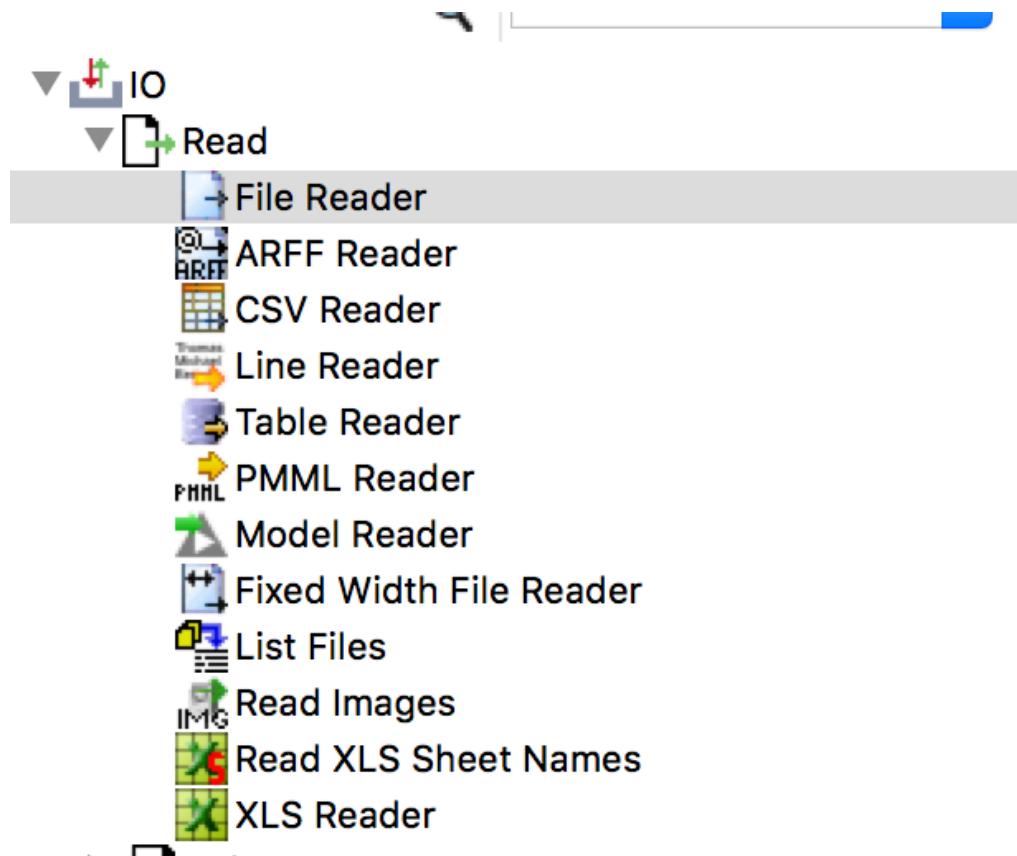
# Read data

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

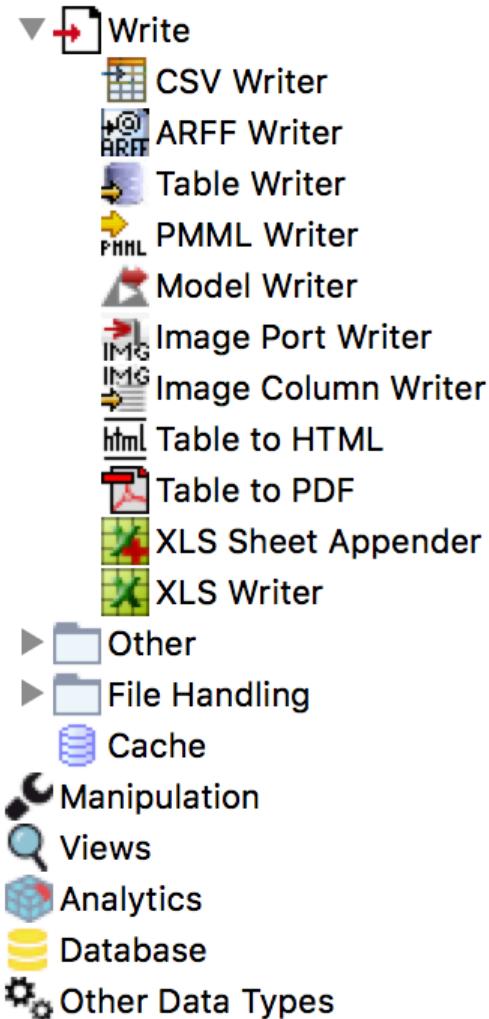
- KNIME Explorer**: Shows a project tree with nodes like "02\_SocialNetworkAnalysis", "050018\_ChurnPrediction", "050019\_TwitterAnalysis", and "002005\_PMML\_Examples".
- Node Repository**: Shows the "Read" category, which includes "File Reader", "ARFF Reader", "CSV Reader", "Line Reader", "Table Reader", "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".
- Node Description**: A panel titled "File Reader" describing its function: "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)."
- Console**: A log window showing KNIME Console output:

```
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  
WARN KnimeRemoteFileSystem  
WARN KnimeRemoteFileSystem  
WARN Color Manager     0:67  
WARN Color Manager     0:67  
WARN Color Manager     0:67  
WARN File Reader       3:1  
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 "Cluster" has no nominal values set: execute predecessor or add Binr  
Column "Churn" has no nominal values set: execute predecessor or add Binr  
Column "Churn" has no nominal values set: execute predecessor or add Binr  
Column "Churn" has no nominal values set: execute predecessor or add Binr  
No Settings available.
```

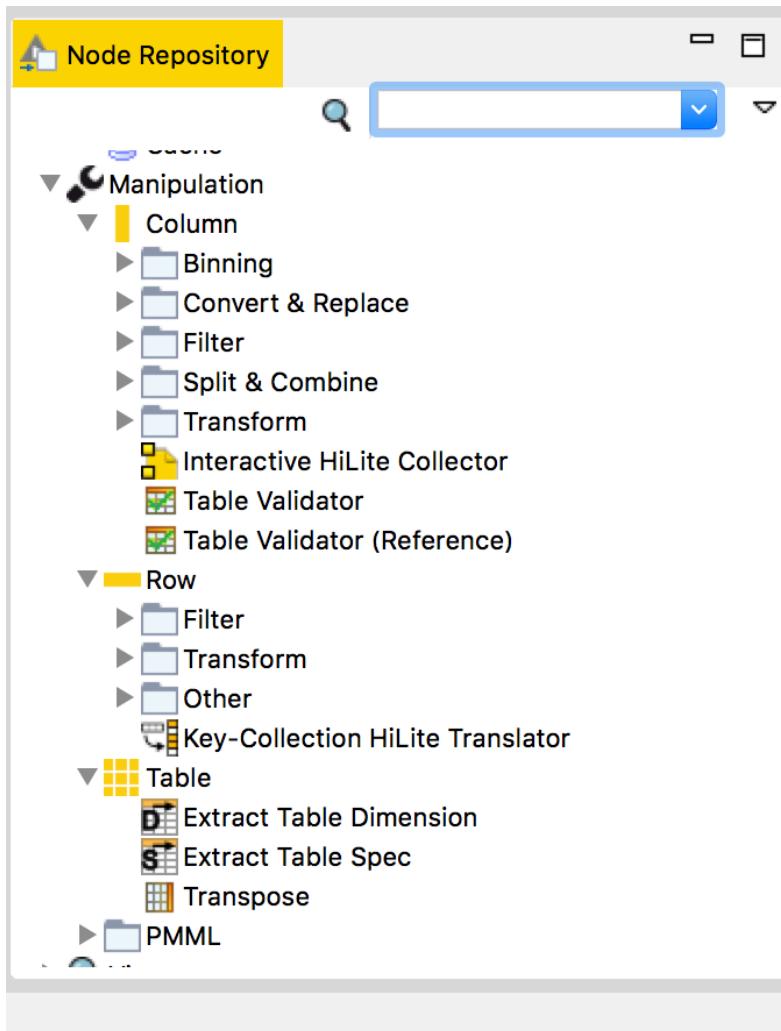
# Several types of importing



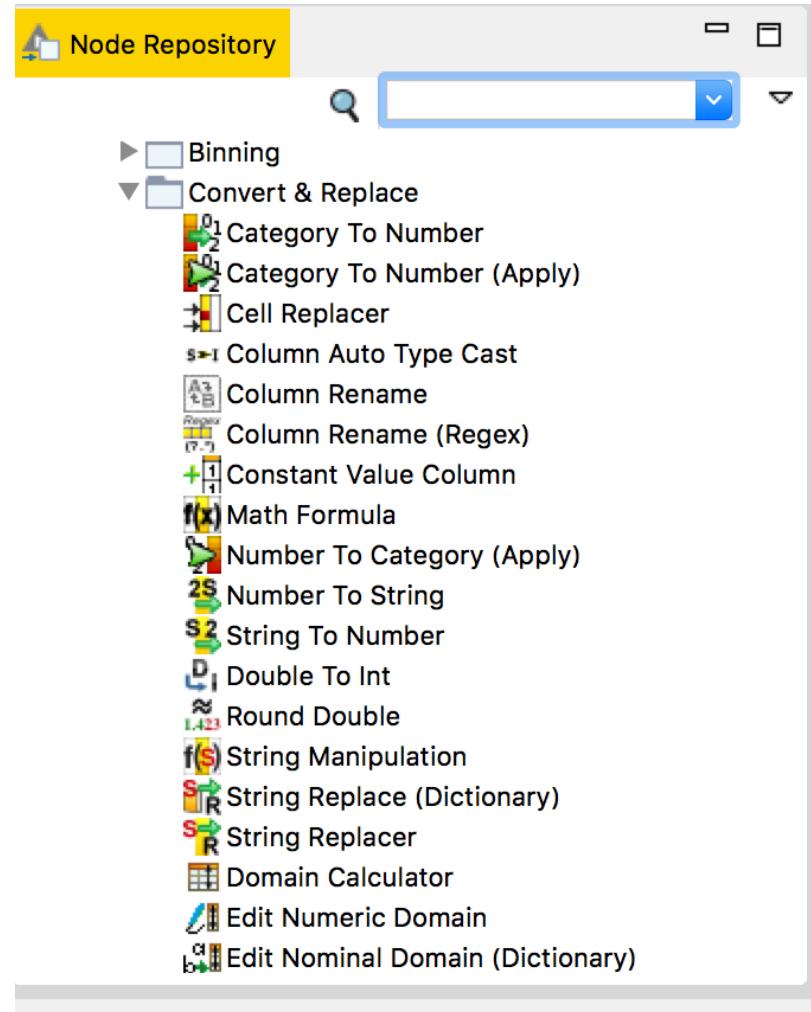
# Write data & results



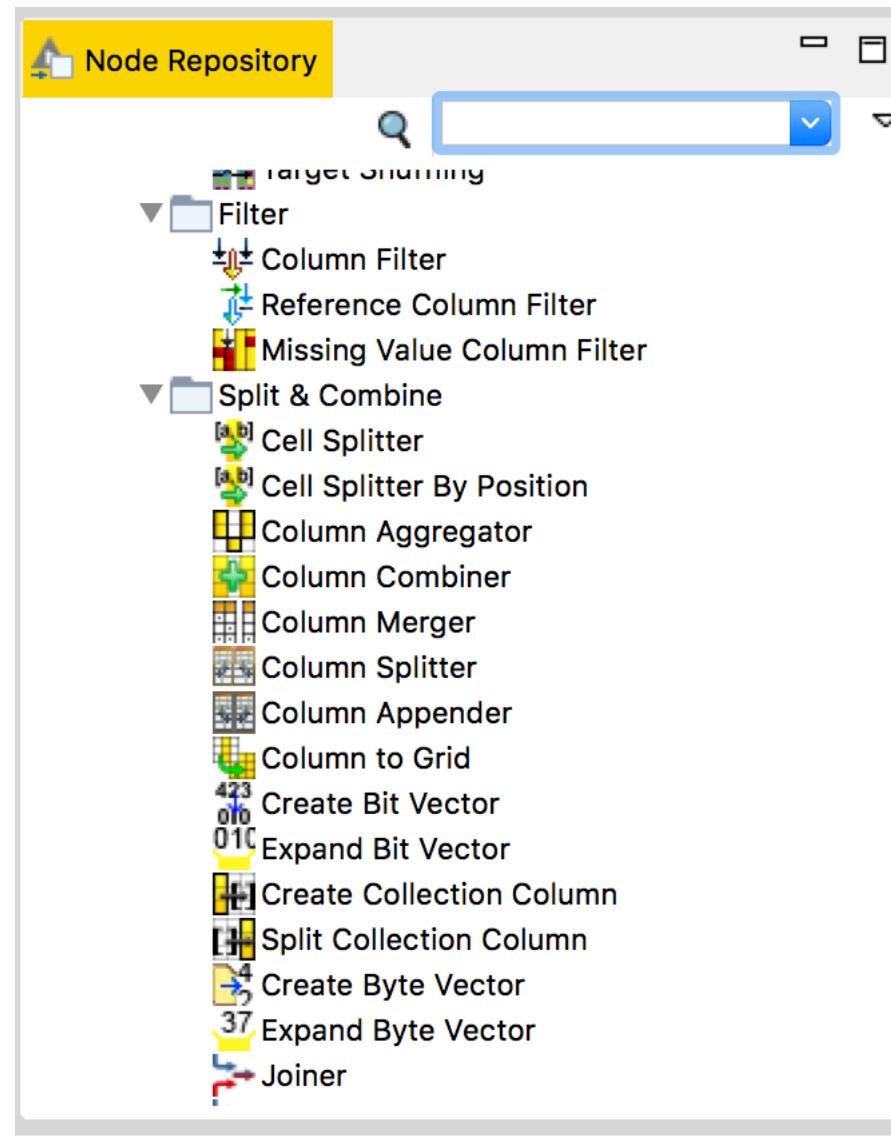
# Data manipulation



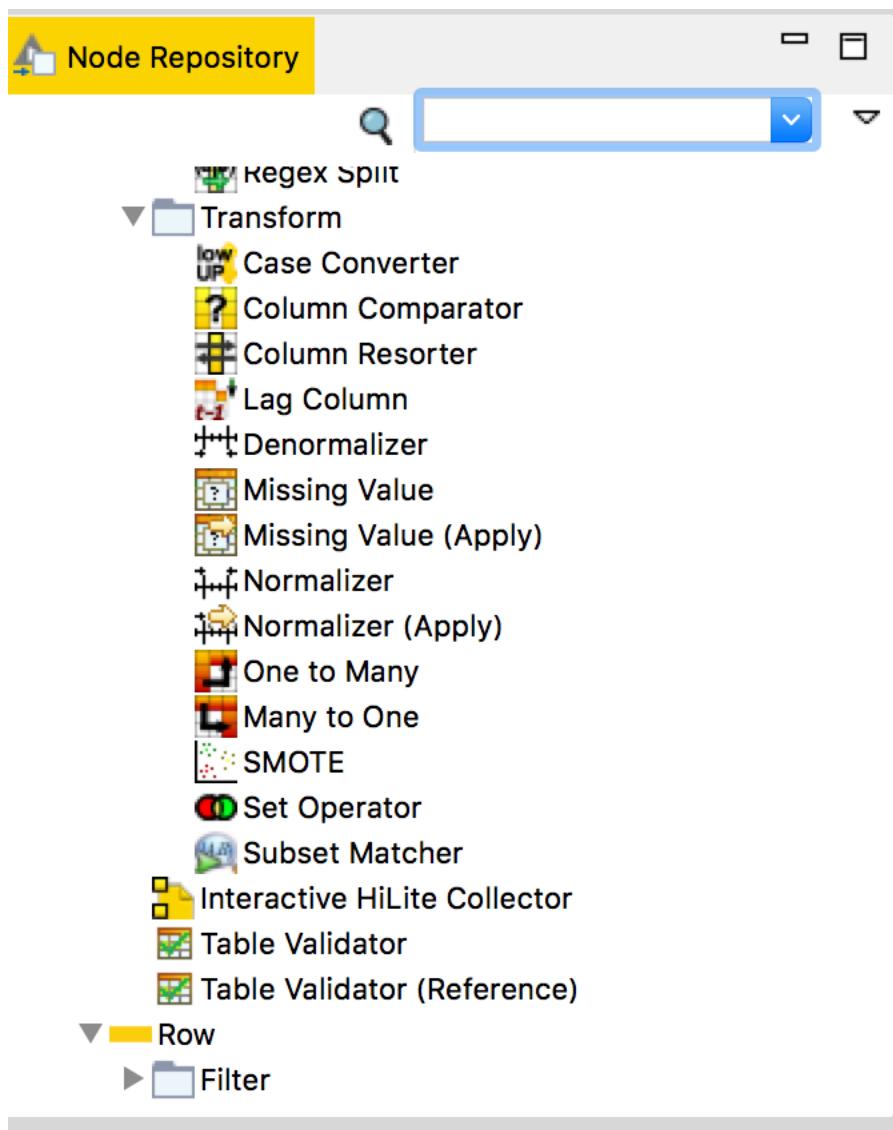
# Data : convert & replace



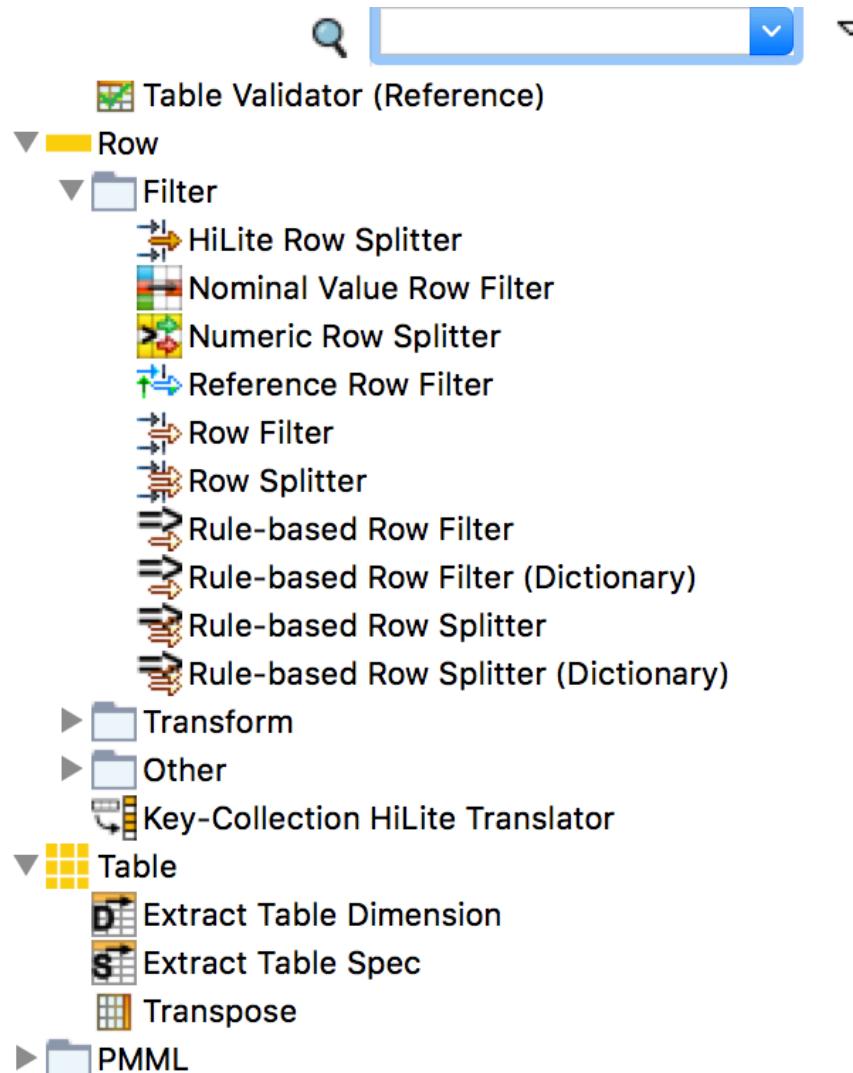
# Split & Combine data



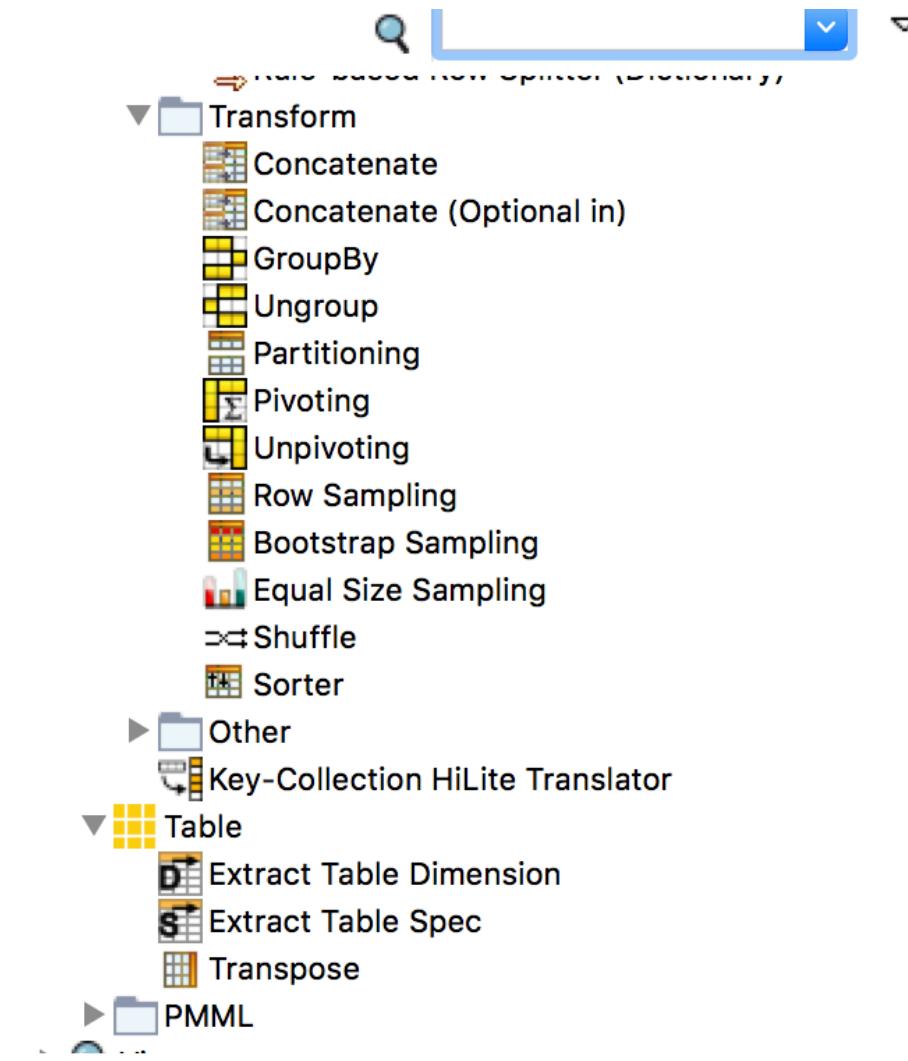
# Data transformation



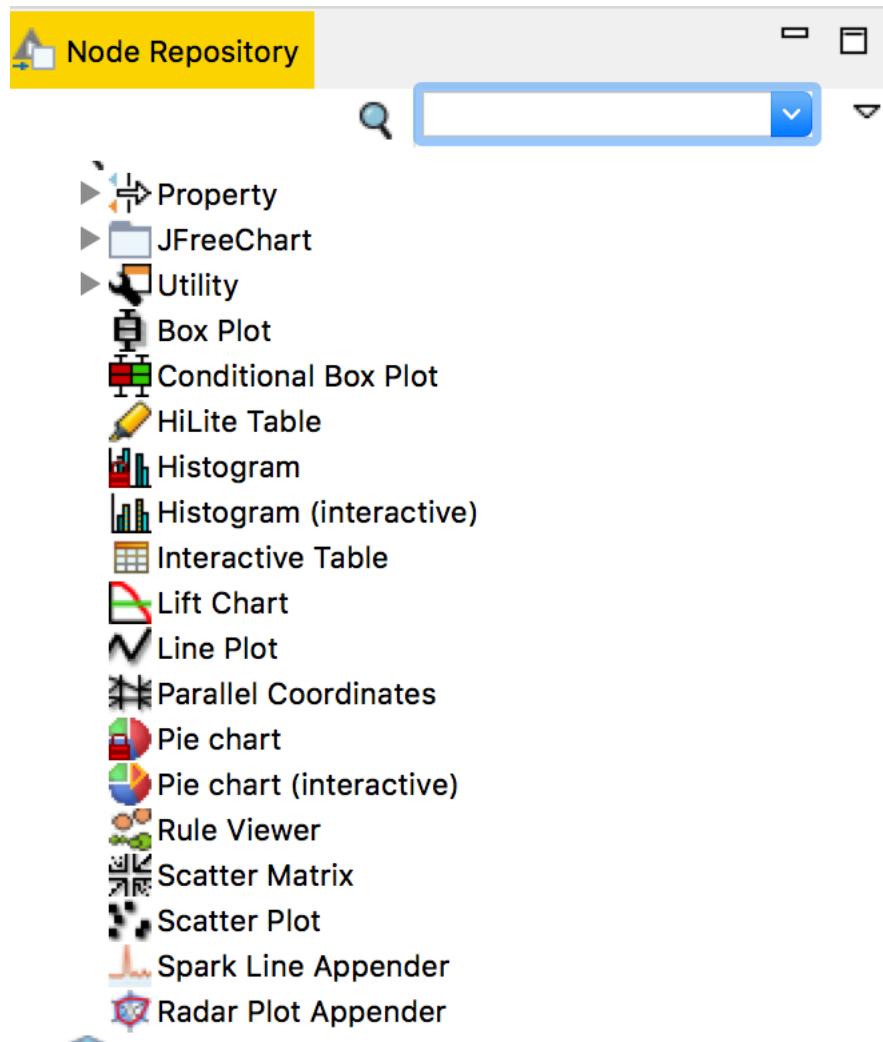
# Transform data : by Row



# Transform data: by Column



# Data visualization



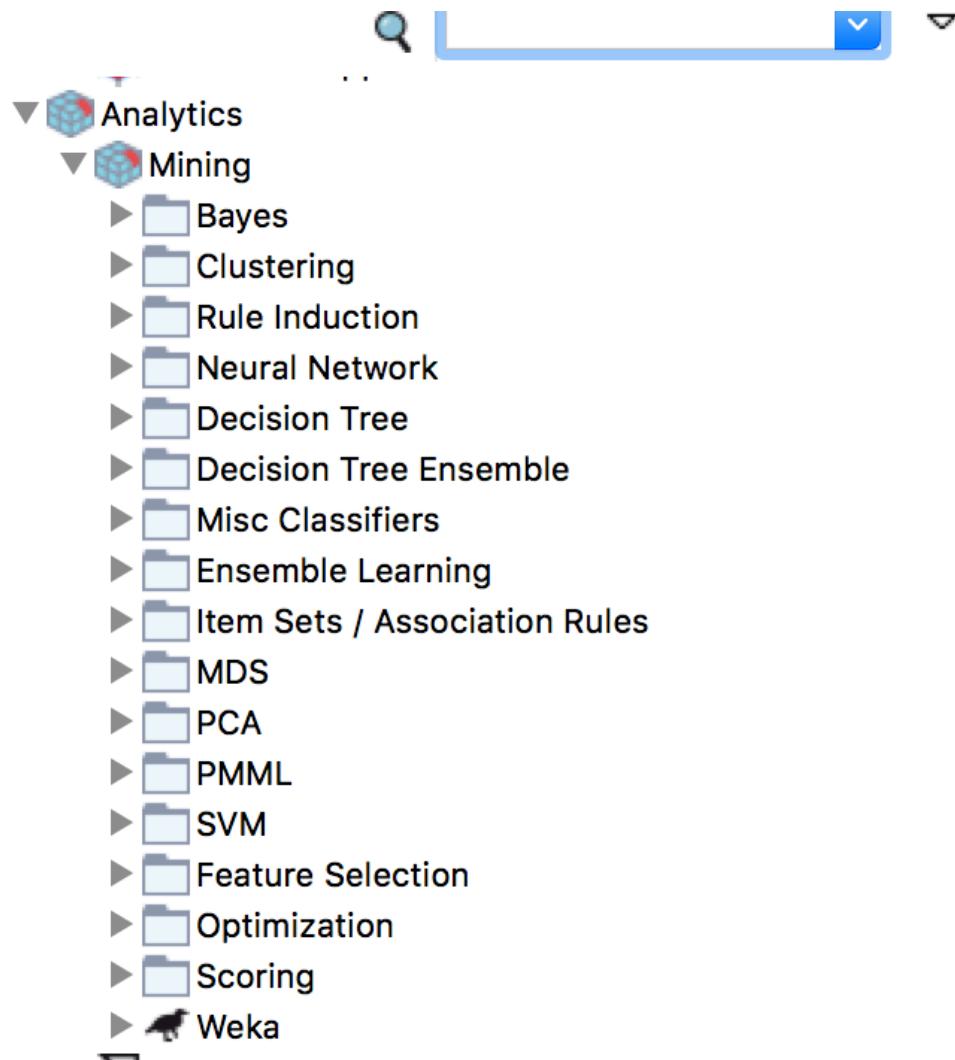
# Other functions



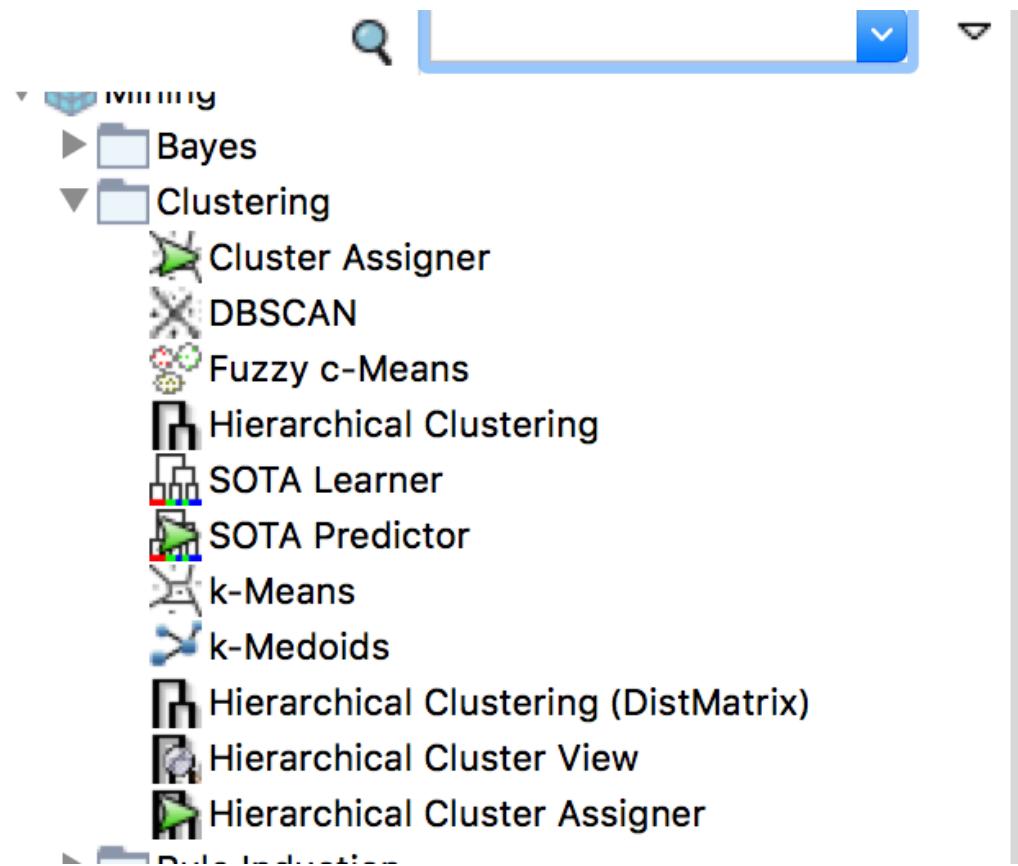
- ▶  Spark Line Appender
- ▶  Radar Plot Appender
- ▶  Analytics
  - ▶  Mining
  - ▶  Statistics
  - ▶  Distance Calculation
  - ▶  Database



# Knime: Analytics & Data Mining



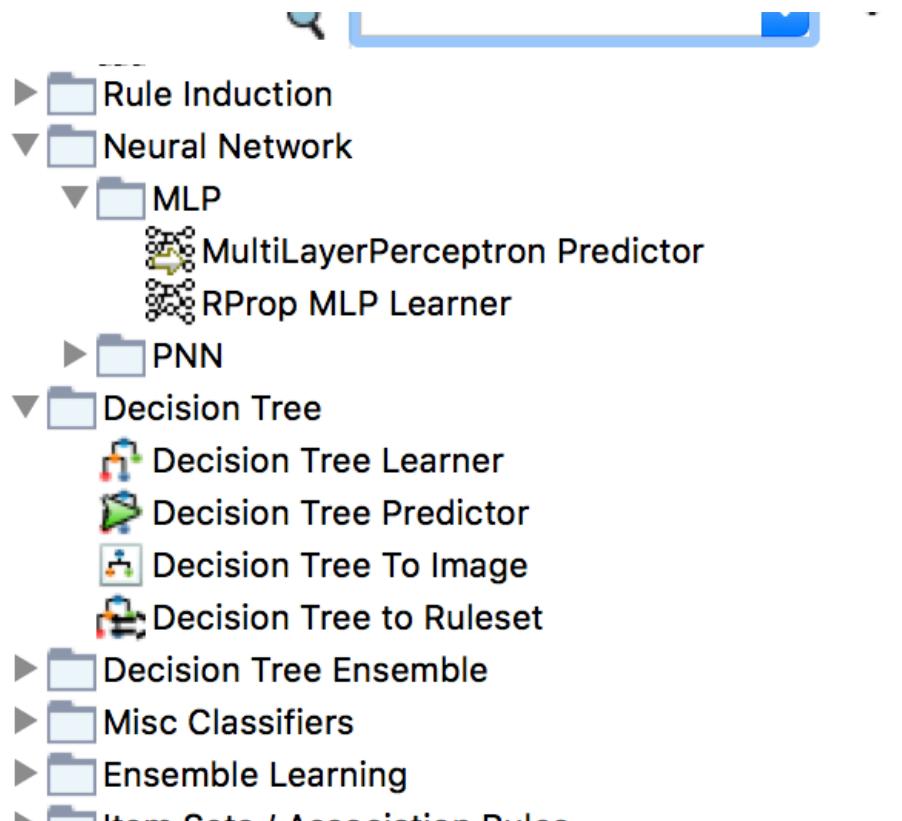
# Clustering data



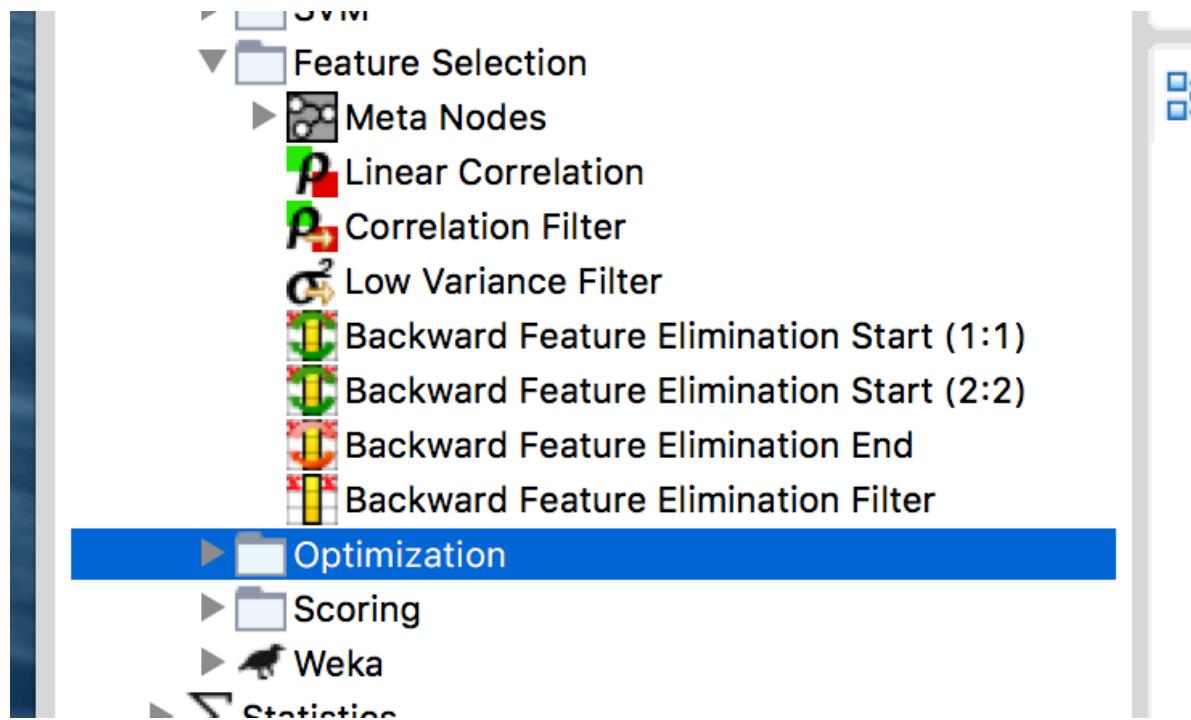
# Data projections methods

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

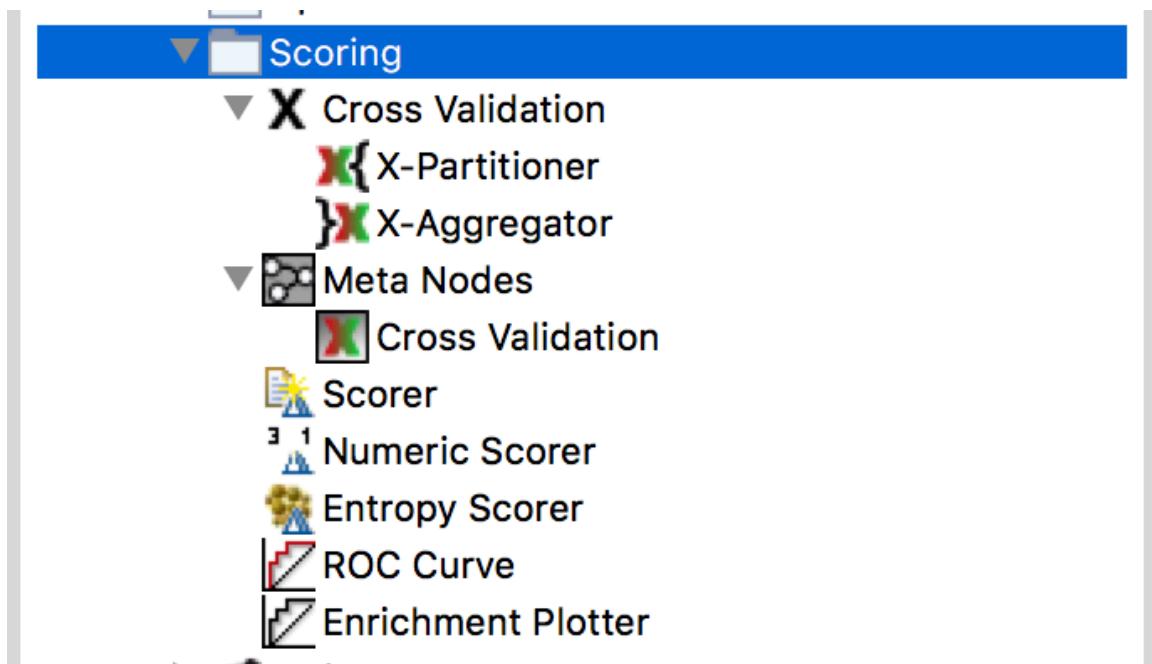
# Predictions models



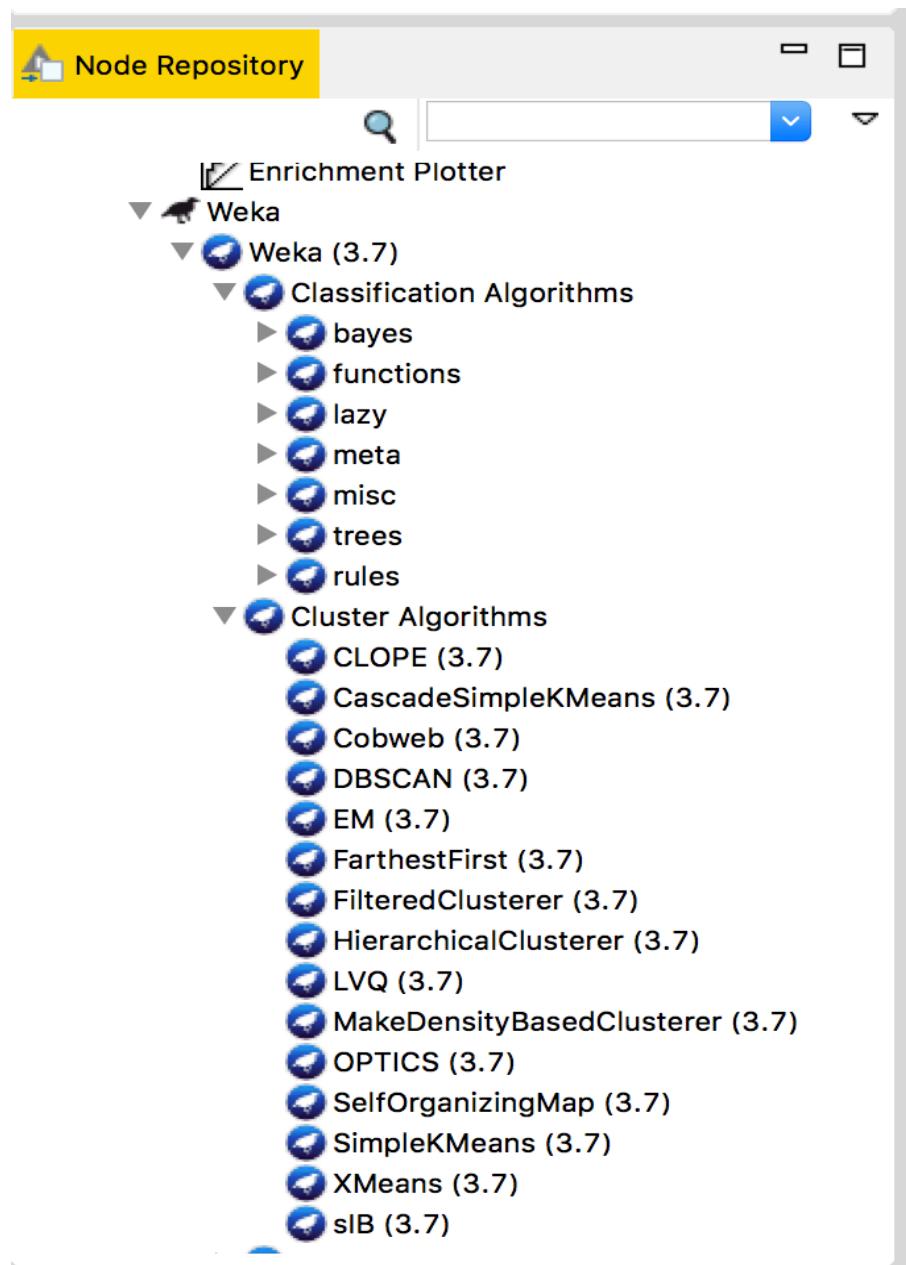
# Features selections approaches



# 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

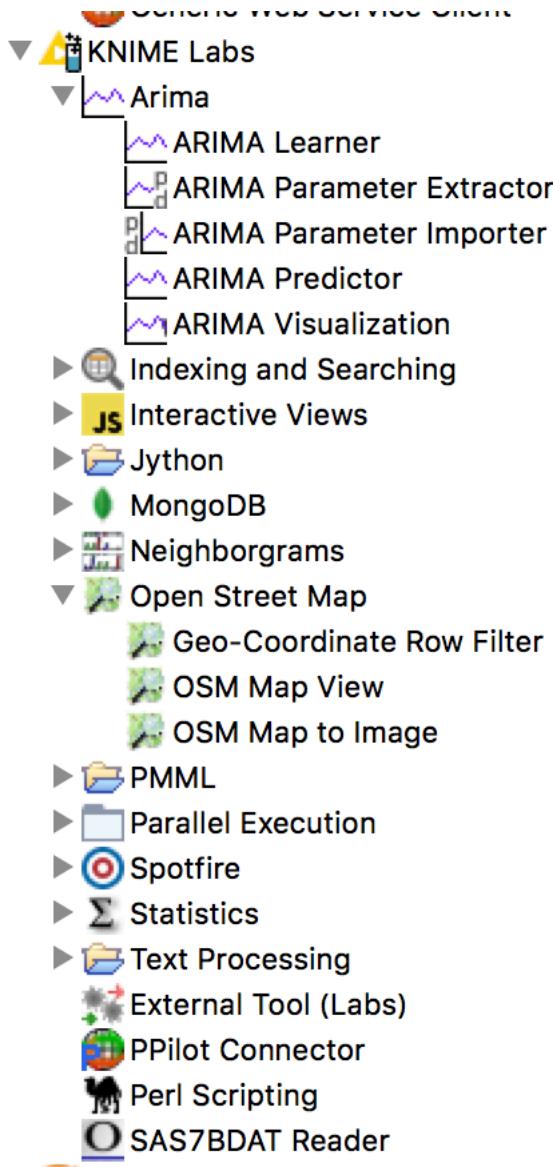
# Statistics

## Distance computing

## Distance matrix

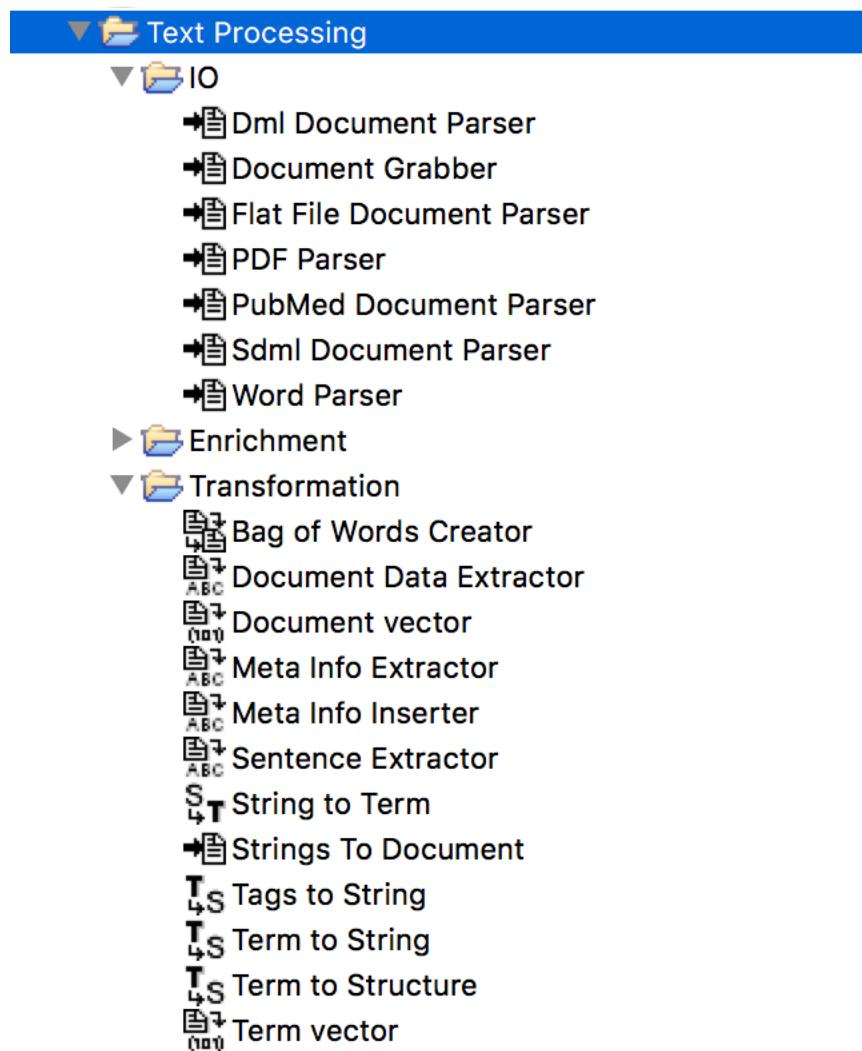
# Scripting

- ▼  Scripting
  - ▼  Java
    -  Java Snippet
    -  Java Snippet (simple)
    -  Java Snippet Row Filter
    -  Java Snippet Row Splitter
  - ▼  Python
    -  Python Edit Variable
    -  Python Source
    -  Python Script
    -  Python Script (2:1)
    -  Python View
    -  Python Object Reader
    -  Python Object Writer
    -  Python Learner
    -  Python Predictor
    -  Python Script (DB)
  - ▼  R
    -  Meta Nodes
    -  IO
      -  R Source (Table)
      -  R Source (Workspace)
      -  R Snippet
      -  R View (Table)
      -  R View (Workspace)
      -  R to Table
      -  Table to R



# Knime Labs

# Text Processing (1)



# Text Processing (2)

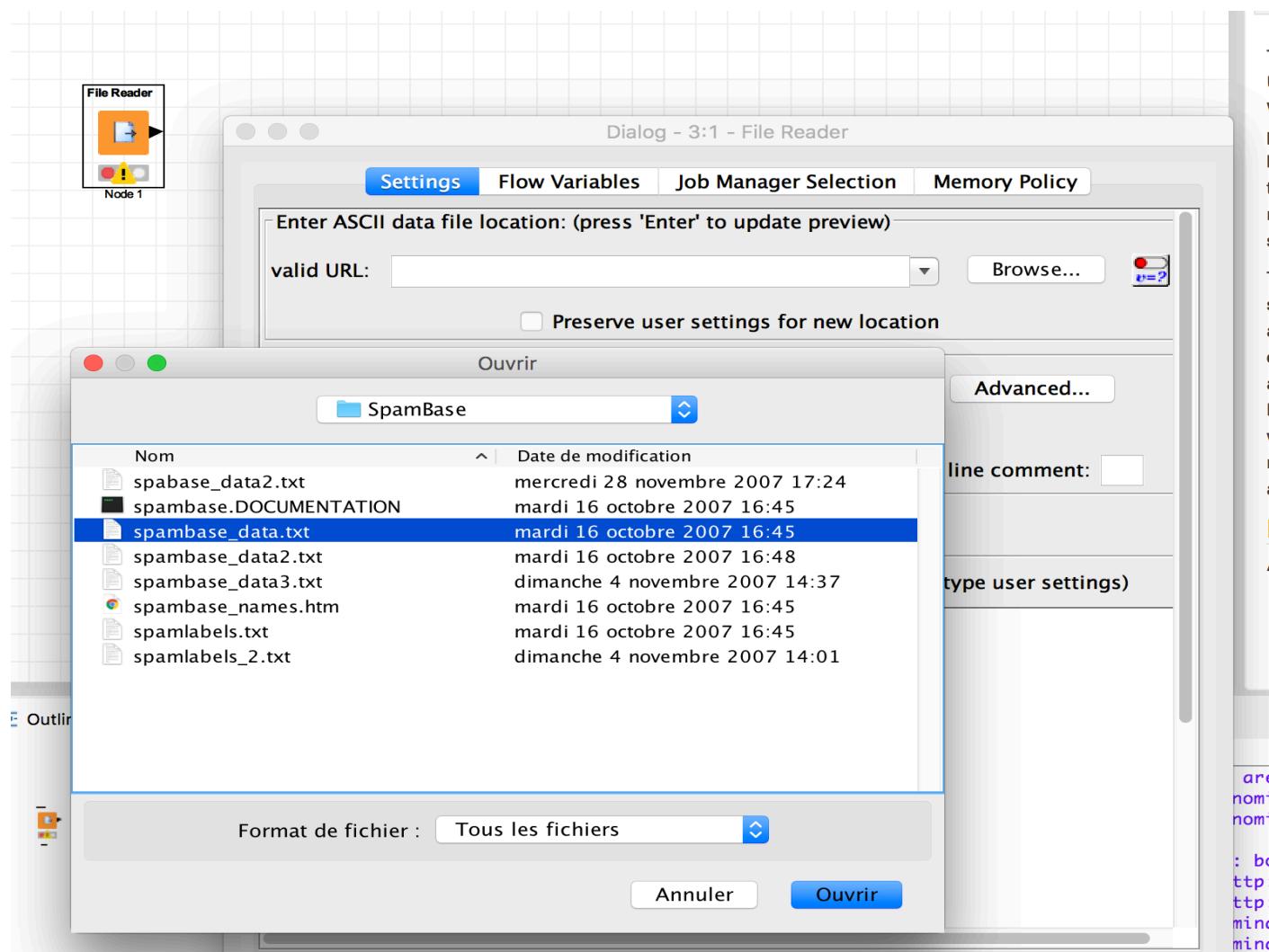
The screenshot shows a software interface with a sidebar containing a tree view of text processing components. The root node is 'Text Processing', which is expanded to show its sub-categories: 'IO', 'Enrichment', 'Transformation', 'Preprocessing', 'Frequencies', 'Mining', and 'Misc'. Each category contains several specific components, each represented by a small icon followed by a descriptive name.

- Text Processing
  - 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

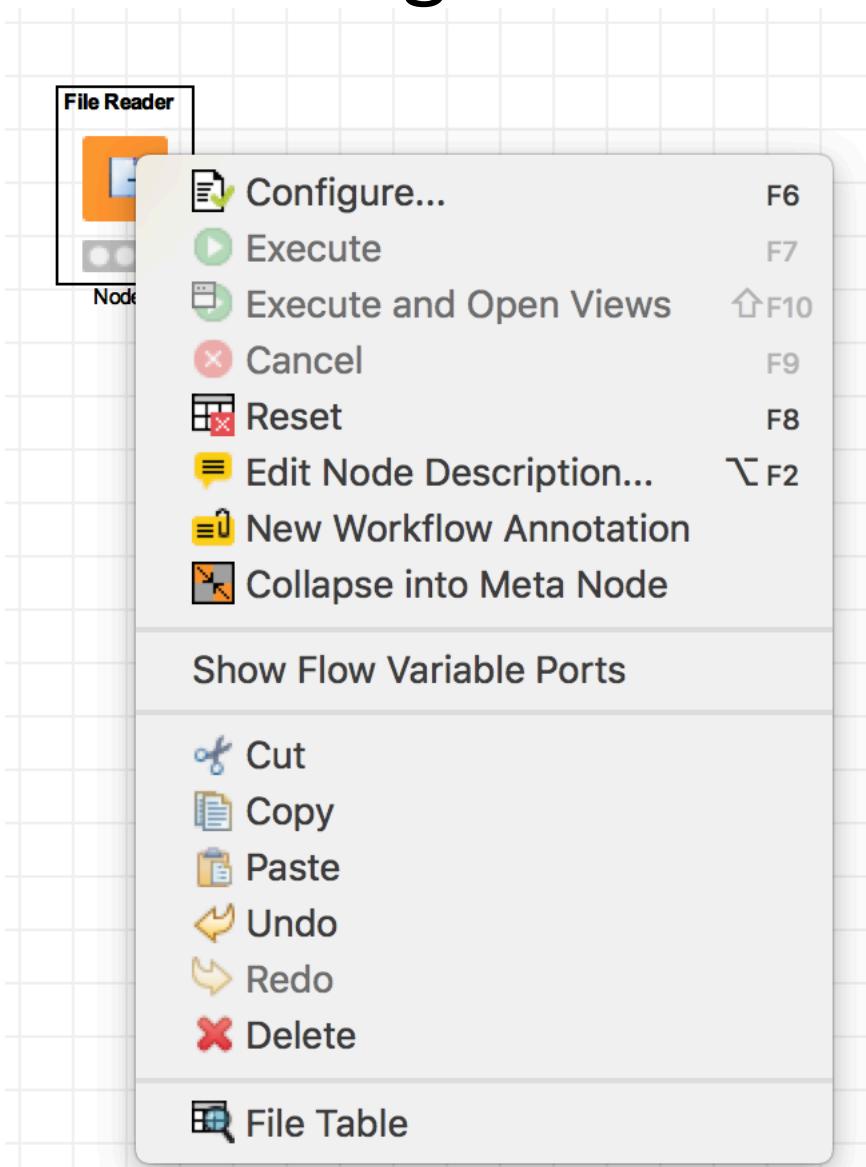
# Other utilities

-  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

The screenshot shows the KNIME interface with a 'File Reader' node selected. The node icon is orange with a white file icon and a green arrow. Below it, the text 'Node 1' is visible. To the right of the node, a tooltip provides information about the node's function: "This node can be used to read files from a URL location. It can handle both local and remote files. When you open the node configuration dialog, you can provide a filename, i.e., a URL, or a file path by analyzing the corresponding input port." The main window displays a table titled 'File Table - 3:1 - File Reader'. The table has a header row with columns: Row ID, Col0, Col1, Col2, Col3, Col4, Col5, Col6, Col7, and Col8. The data rows are labeled 'Row0' through 'Row27'. The first few rows of data are as follows:

Row ID	Col0	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8
Row0	0	0.64	0.64	0	0.32	0	0	0	0
Row1	0.21	0.28	0.5	0	0.14	0.28	0.21	0.07	0
Row2	0.06	0	0.71	0	1.23	0.19	0.19	0.12	0
Row3	0	0	0	0	0.63	0	0.31	0.63	0
Row4	0	0	0	0	0.63	0	0.31	0.63	0
Row5	0	0	0	0	1.85	0	0	1.85	0
Row6	0	0	0	0	1.92	0	0	0	0
Row7	0	0	0	0	1.88	0	0	1.88	0
Row8	0.15	0	0.46	0	0.61	0	0.3	0	0
Row9	0.06	0.12	0.77	0	0.19	0.32	0.38	0	0
Row10	0	0	0	0	0	0	0.96	0	0
Row11	0	0	0.25	0	0.38	0.25	0.25	0	0
Row12	0	0.69	0.34	0	0.34	0	0	0	0
Row13	0	0	0	0	0.9	0	0.9	0	0
Row14	0	0	1.42	0	0.71	0.35	0	0.35	0
Row15	0	0.42	0.42	0	1.27	0	0.42	0	0
Row16	0	0	0	0	0.94	0	0	0	0
Row17	0	0	0	0	0	0	0	0	0
Row18	0	0	0.55	0	1.11	0	0.18	0	0
Row19	0	0.63	0	0	1.59	0.31	0	0	0
Row20	0	0	0	0	0	0	0	0	0
Row21	0.05	0.07	0.1	0	0.76	0.05	0.15	0.02	0
Row22	0	0	0	0	2.94	0	0	0	0
Row23	0	0	0	0	1.16	0	0	0	0
Row24	0	0	0	0	0	0	0	0	0
Row25	0.05	0.07	0.1	0	0.76	0.05	0.15	0.02	0
Row26	0	0	0	0	0	0	0	0	0
Row27	0	0	0	0	0	0	1.66	0	0

# Statistics view

Analysis  
Prediction

nDeployment  
nTraining

nodes

File

statistics

ing  
e t-test  
groups t-te

DVA

ession Learne  
egression 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

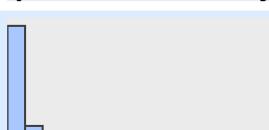
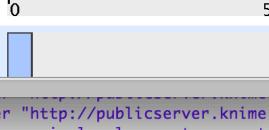
File Reader → Statistics

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 counts of unique values together with their occurrences. The output is a table with one row per column.

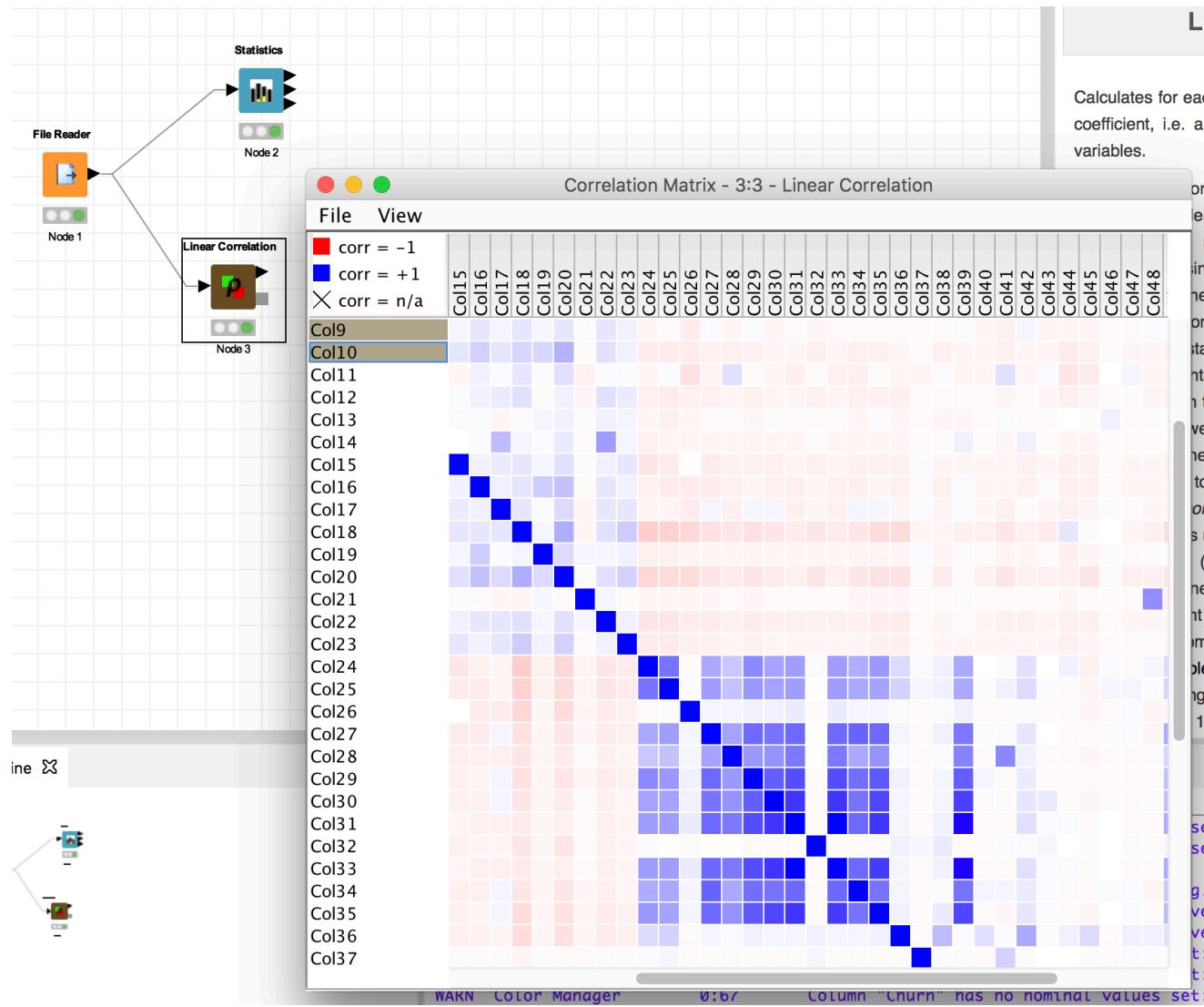
Statistics View - 3:2 - Statistics

Numeric Nominal Top/bottom

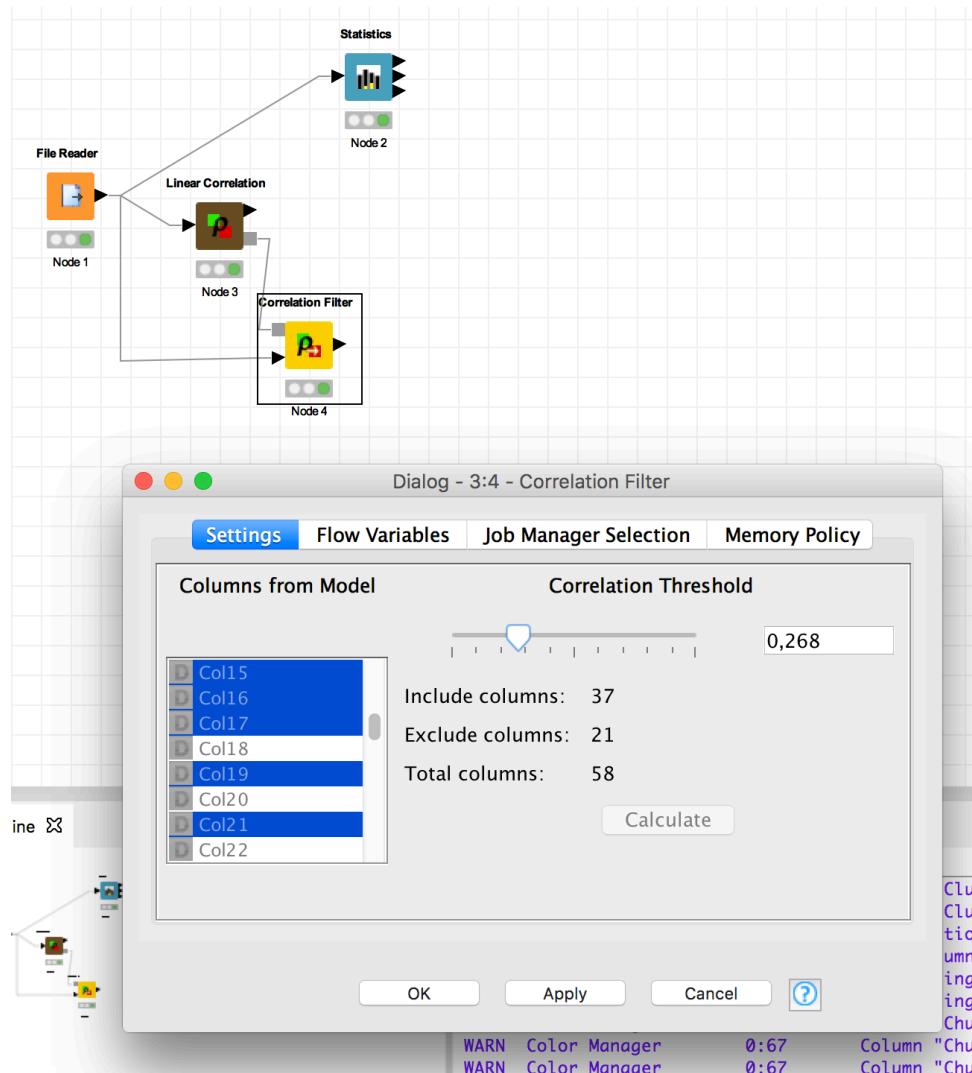
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
Col0	0.0	0.1046	?	4.54	0.3054	5.6756	49.3051	0	0	0	
Col1	0.0	0.213	?	14.28	1.2906	10.0868	105.6475	0	0	0	
Col2	0.0	0.2807	?	5.1	0.5041	3.0092	13.3087	0	0	0	
Col3	0.0	0.0654	?	42.81	1.3952	26.2277	726.4515	0	0	0	

WARN KnimeRemoteFileSystem  
WARN Cell Manager 0.67  
Connecting to server "http://publicserver.knime.org:80/tomcat" "Churn" has been loaded

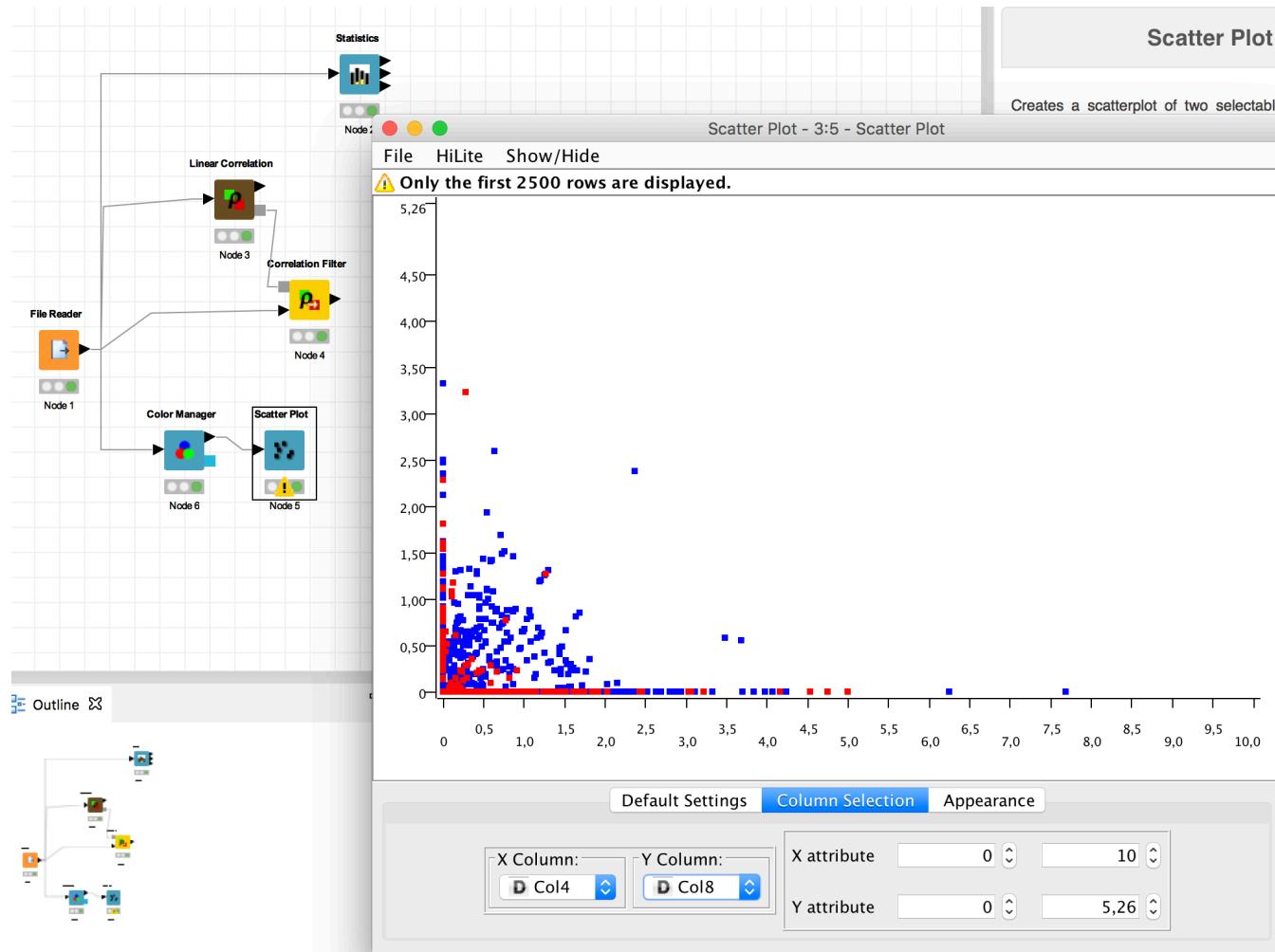
# Correlation view



# Correlation Filter



# Visualization: Scatter Plot



# Principal Component Analysis (PCA)

The screenshot shows a data mining workflow and a detailed configuration dialog for a PCA node.

**Flow Diagram:**

- A "File Reader" node (Node 1) feeds into a "Color Manager" node.
- The output of the "Color Manager" node goes to a "PCA" node (Node 6).
- The output of Node 6 goes to another "Color Manager" node (Node 8), which then feeds into a "Scatter Plot" node (Node 9).
- The "Color Manager" node (Node 1) also outputs to a "Scatter Plot" node (Node 5).
- A "Linear Correlation" node (Node 3) receives input from the "File Reader" node and outputs to a "Correlation Filter" node (Node 4).
- The "Correlation Filter" node outputs to a "Statistics" node (Node 2).

**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 Tab:**

- Fail if missing values are encountered (skipped per default)

**Target dimensions:**

Dimensions to reduce to

Minimum information fraction to preserve (%)

Replace original data columns

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

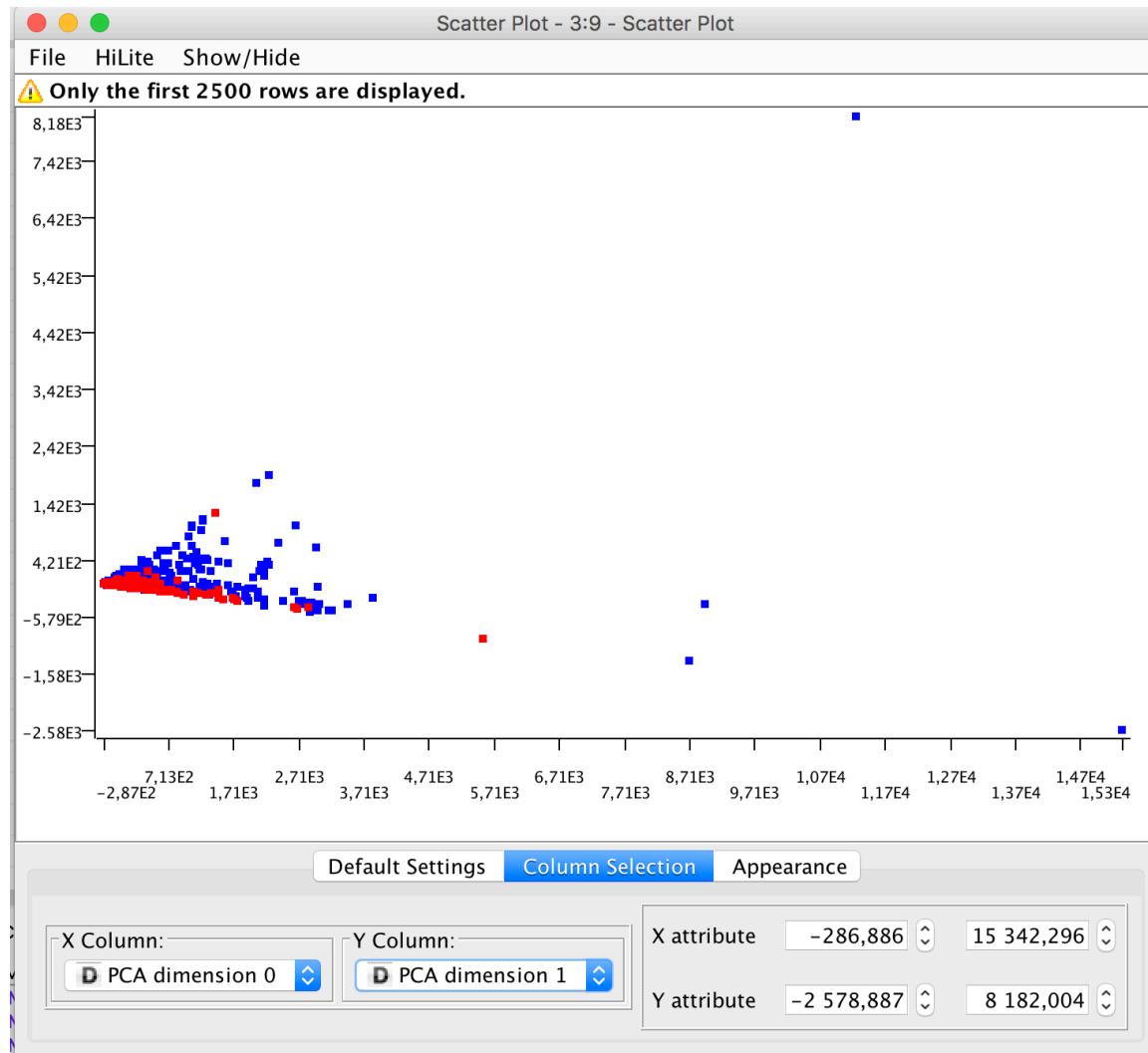
**Select:** add >>

**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 ?

# Project the data using PCA



# Data normalization

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**

The screenshot shows a KNIME workflow with a 'File Reader' node (Node 11) connected to a 'Normalizer' node (Node 12). The 'Normalizer' node has a warning icon. Below the workflow is the 'Dialog - 3:12 - Normalizer' window.

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

Manual Selection  Wildcard/Regex Selection

**Exclude**     
**Select**

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

**Settings**

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

Min:   
Max:

# Data normalization

The screenshot shows a KNIME workflow interface. At the top, there is a diagram of the workflow:

```
graph LR; Node11[File Reader] --> Node12[Normalizer]; Node12 --> Node13[Statistics]
```

Below the diagram, a tooltip provides a detailed description of the Statistics node:

This node calculates statistical information for numeric columns. It computes minimum, maximum, mean, standard deviation, median, overall sum, number of missing values across all numeric columns, and unique values together with their occurrence.

The main window title is "Statistics View - 3:13 - Statistics". The menu bar shows "File". The toolbar includes buttons for "Numeric" (selected), "Nominal", and "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

File Reader → Normalizer → Statistics

k-Means

This node outputs the number of clusters (no means performs a crisp vector to exactly one c when the cluster assignn. The clustering algorithm the selected attributes. T node (if required, you "Normalizer" as a prepro If the optional PMML is preprocessing operations:

Dialog - 3:14 - k-Means

K-Means Properties Flow Variables Job Manager Selection Memory Policy

number of clusters: 3 ↕

max. number of iterations: 99 ↕

Search hits

Select

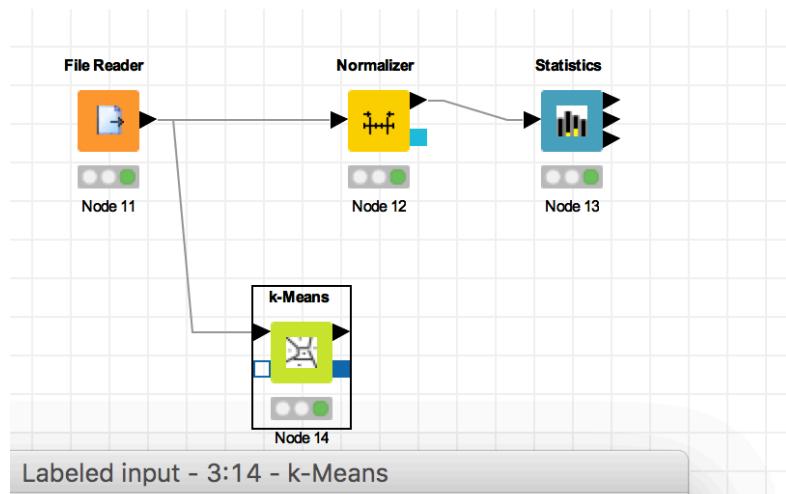
Include

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

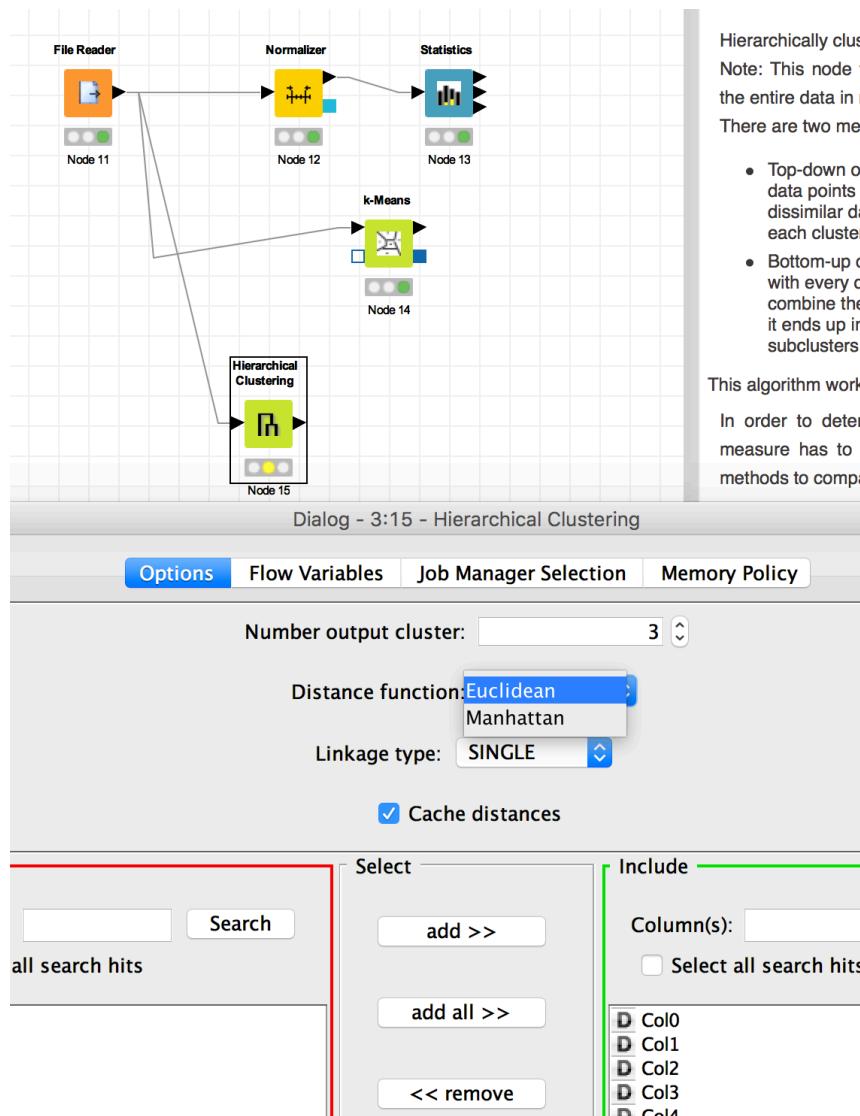
	Col55	Col56	Col57	Cluster
4	61	278	1	cluster_0
	101	1028	1	cluster_1
	485	2259	1	cluster_1
	40	191	1	cluster_0
	40	191	1	cluster_0
	15	54	1	cluster_0
	4	112	1	cluster_0
	11	49	1	cluster_0
	445	1257	1	cluster_1
	43	749	1	cluster_1
	6	21	1	cluster_0
	11	184	1	cluster_0
	61	261	1	cluster_0
	7	25	1	cluster_0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

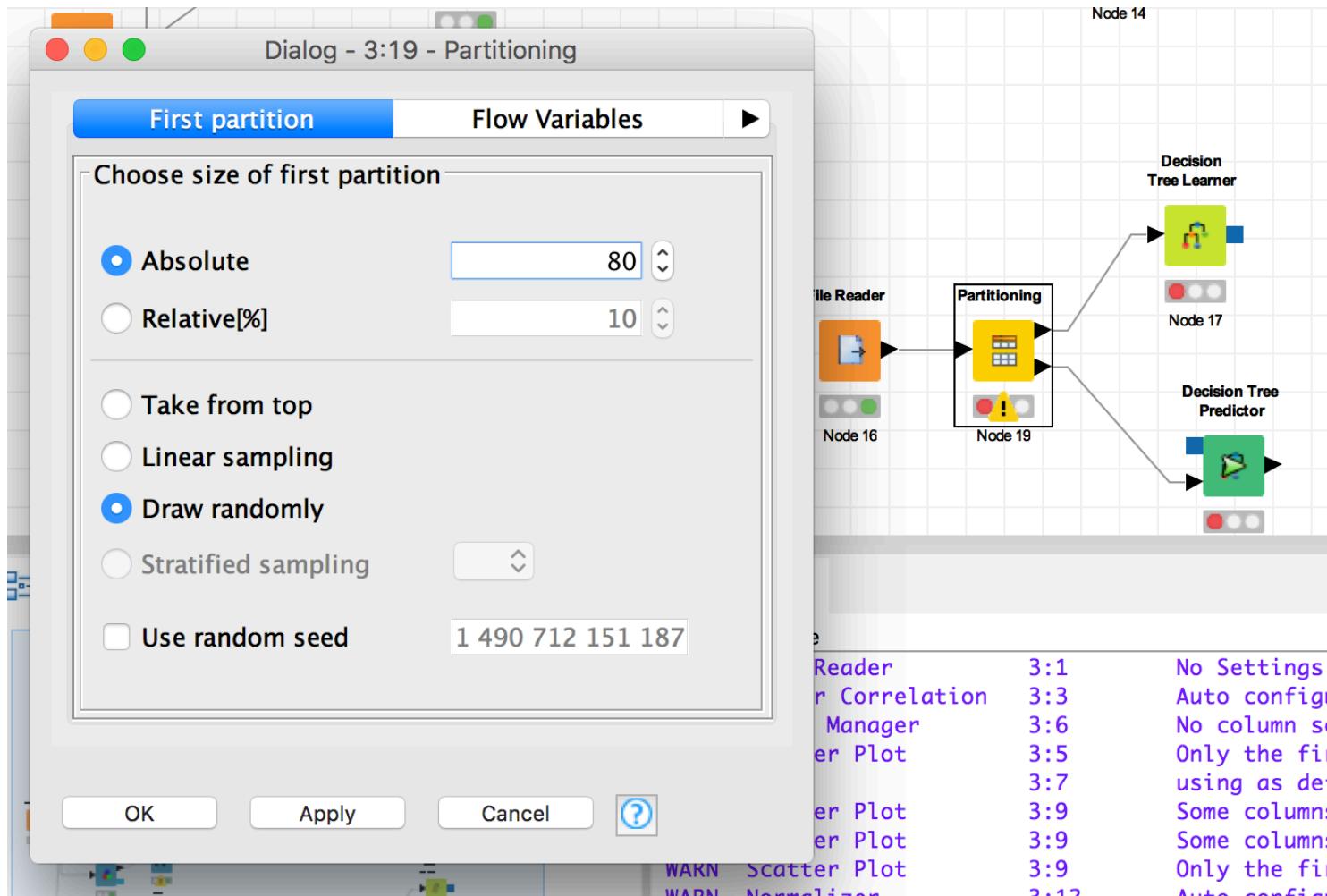
Note: This node v  
the entire data in n  
There are two met

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

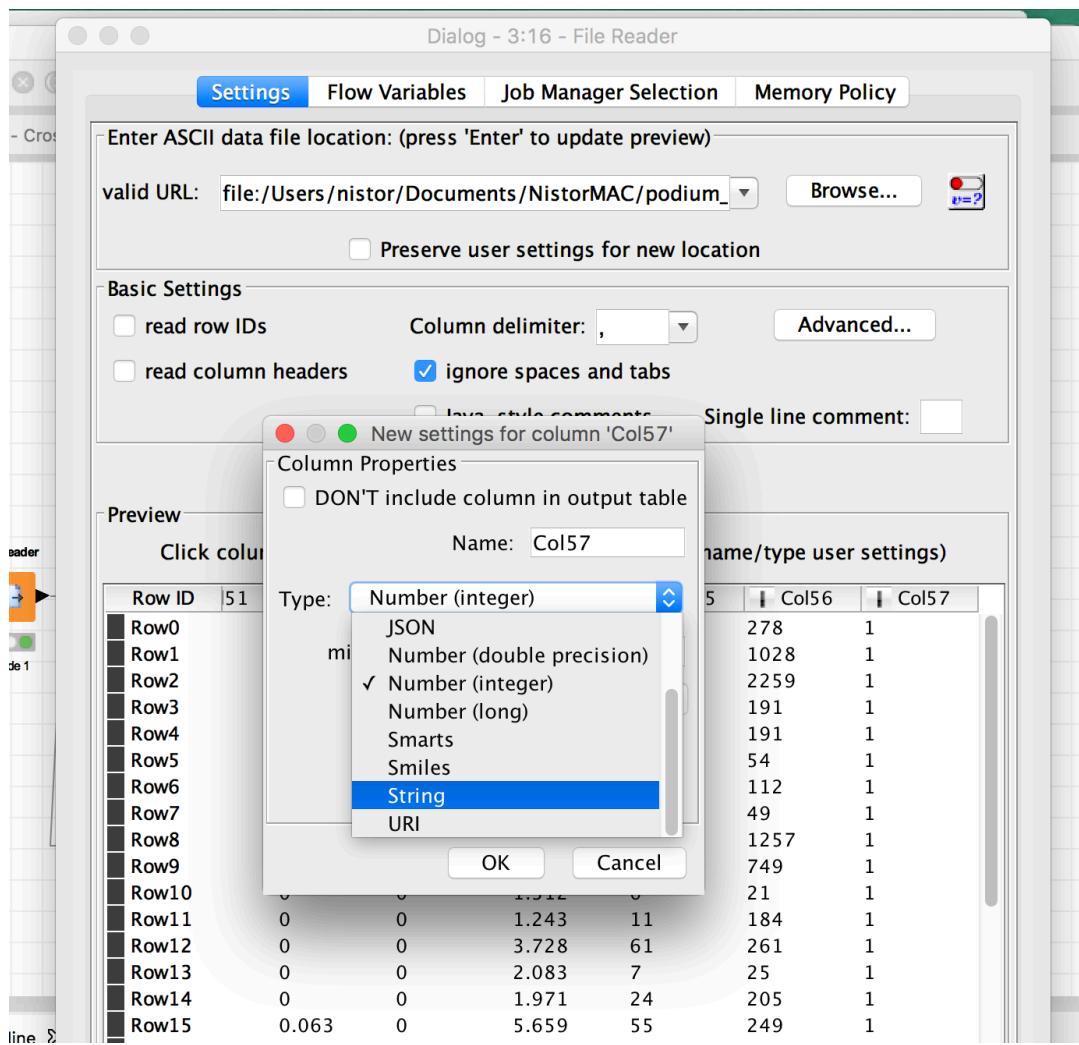
This algorithm work

In order to deter  
measure has to b  
methods to compa

# Machine Learning (1)



# Machine Learning (2)



# 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

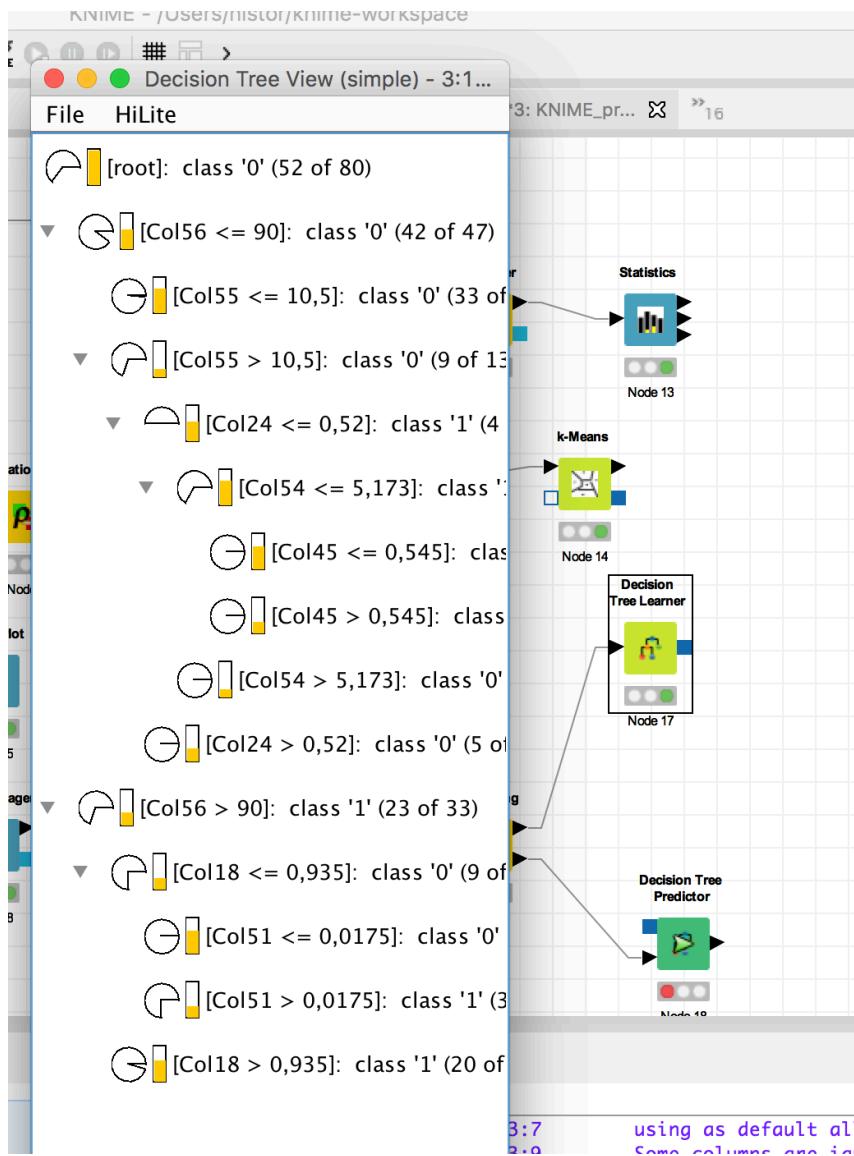
Decision Tree Learner  
Node 17

Decision Tree Predictor

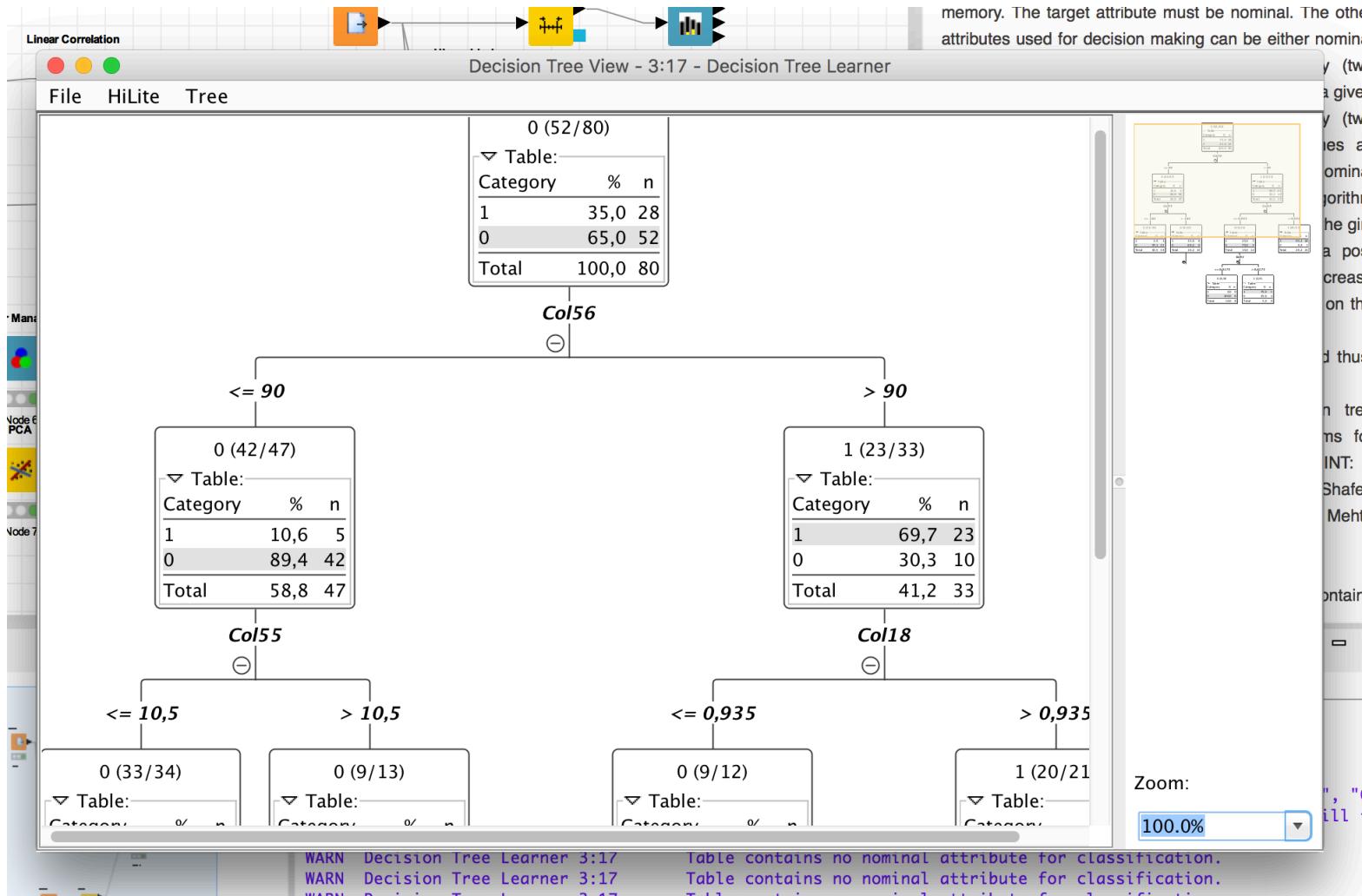
using as default  
Some columns are  
Some columns are  
Only the first 2  
Auto-configure:  
Potential deadlo  
No sampling meth  
Table contains  
Table contains  
Table contains

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

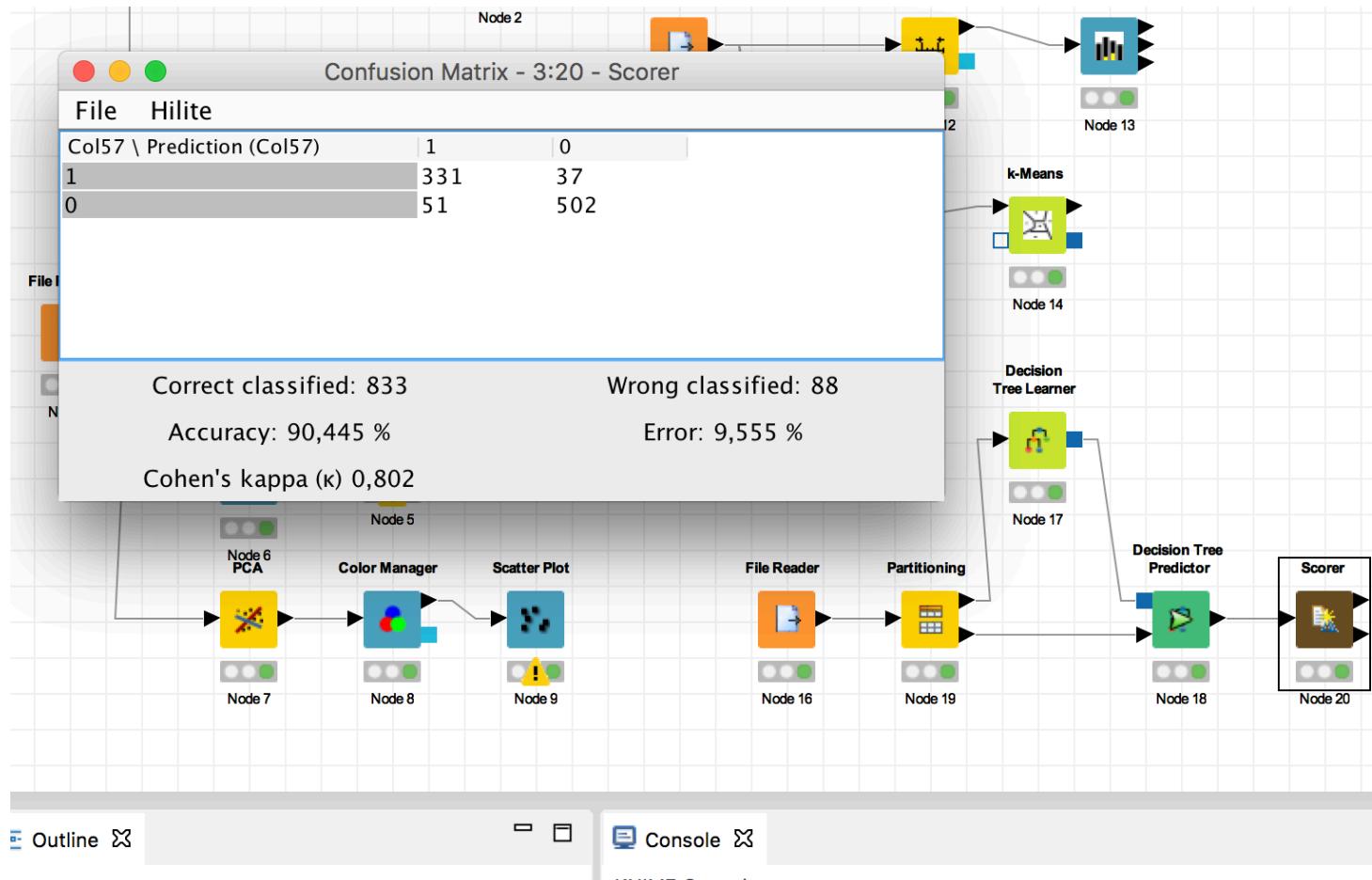
# Machine Learning (4)



# 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)

