Cheat Sheet: Building a KNIME Workflow for Beginners

Follow me on LinkedIn for more:
Steve Nouri
https://www.linkedin.com/in/stevenouri/

Open for Innovation
KNIME

Getting started with KNIME Analytics Platform

- Read through the installation guide at knime.com/installation
- Check out the 7 things you should do after installing KNIME Analytics Platform at knime.com/blog/seven-things
- Take the E-Learning Course at knime.com/knime-introductory-course
- Browse the workflows on the public EXAMPLES Server available in the KNIME Explorer

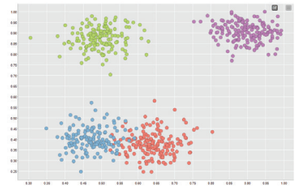
Understanding the traffic light system:

- Not configured: Node is not yet configured and cannot be executed with its current settings
- Configured: Node has been correctly configured and may be executed at any time
- Executed: Node has been successfully executed and results can be viewed and used in downstream nodes

EXPLORE

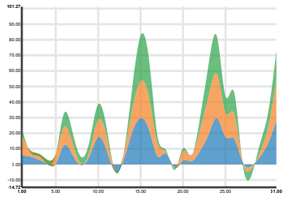
(All visualizations are interactive)

Scatter Plot: Represents input data rows as points in a two dimensional plot. Input dimensions (columns) on the x-y axis plot and graphical properties can be changed in the configuration window or interactively in the node view.



Sunburst Chart: Displays categorical columns through a hierarchy of rings. Each ring is sliced according to the nominal values in the corresponding column and to the selected hierarchy. This is a powerful chart for multivariate analysis.

Stacked Area Chart: Plots multiple numerical data columns on top of each other using the previous line as the base reference. The areas in between lines are colored for easier comparison. This chart is commonly used to visualize trending topics.



Line Plot

Line Plot: Plots numerical values in data columns (y-axis) against values in a reference column (x-axis). Data points are connected via colored lines. If the reference column on the x-axis contains sorted time values, the line plot graphically represents the evolution of a time series.

Color Manager

Color Manager: Assigns a color property to each input row based on the row's value in a selected column. This color property affects the graphical representation in the upcoming views.

Pie Chart

Pie Chart: Visualizes one aggregated metric for different data partitions with colored slices on a circle where the areas are proportional to the metric values. The partitions are defined by a categorical column.

Data Explorer

Data Explorer: Provides an interactive view to summarize the statistics of the input data via statistical measures and histograms - for both numerical and nominal columns.

Box Plot

Box Plot: Visualizes numeric columns using the quartile statistics. Watch out for the points at the end of the whiskers - they might mark outliers!

Bar Chart

Bar Chart: Visualizes one or more aggregated metrics for different data partitions with rectangular bars where the heights are proportional to the metric values. The partitions are defined by a categorical column.

ANALYZE

Decision Tree

Decision Tree: The Learner node trains a C4.5 or a CART decision tree. The configuration window includes options for pruning, early stopping, information measures, splitting values, and more. Both the Learner and the Predictor node provide an interactive view where the decision tree is displayed together with the input data propagation.

k-Means

k-Means: Implements the k-Means clustering algorithm. Number of clusters must be set prior to node execution. This node builds the clusters. The Cluster Assigner node finds the closest cluster and assigns it to the input data row. Being an unsupervised algorithm, this node pair doesn't follow the classic Learner - Predictor scheme.

Logistic Regression

Logistic Regression: The Learner node trains a logistic regression model to predict categorical target values. The configuration window includes options for solver, input feature choice, regularization functions to avoid overfitting, & more.

Scorer

Scorer: Calculates a number of performance measures such as accuracy, F1-score, or Cohen's Kappa, to quantify the quality of a classifier.

Numeric Scorer

Numeric Scorer: Calculates a number of numerical error measures, such as root mean squared error, mean absolute error, or R^2, to quantify the quality of a numerical predictor model.

ROC Curve

ROC Curve: Displays the Receiver Operating Characteristic (ROC) curve of a classifier working on a binary class problem. One of the two classes is arbitrarily chosen as the positive class and the ROC curve is built on the probabilities/scores produced for that class on the input data set.

Integrations to many open source data analytics tools are also available. Some use the KNIME node GUI (H2O, Weka, Keras, Spark MLlib). Others offer nodes with a development environment for scripting and debugging (R, Python, Java).

READ

File Reader

File Reader: Reads all text files, particularly character separated files, such as CSV files. The File Reader is the workhorse for reading text data.

Excel Reader (XLS)

Excel Reader (XLS): Reads content from sheets in Excel files (XLS, XLSX). Sheet and cells to be read can be defined in the configuration window.

Table Creator

Table Creator: Allows users to manually create a data table in its configuration window as a data sheet. Data cells can be copied and pasted in the sheet. Perfect for generating small data sets.

Model Reader

Model Reader: Reads machine learning models generated with any of the Learner nodes. Models are usually saved after training and reused in deployment.

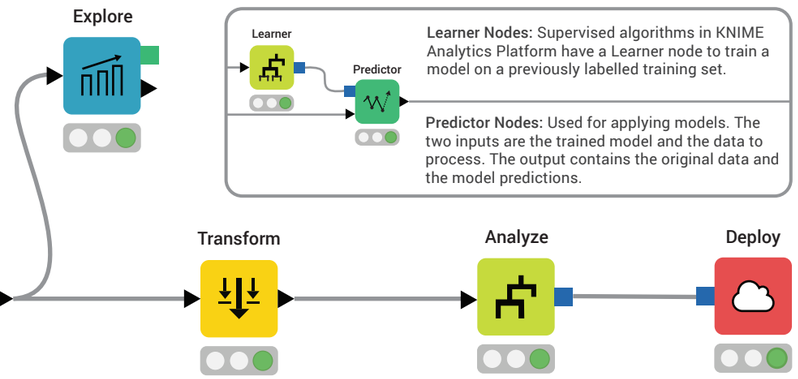
Table Reader

Table Reader: Reads data from a .table file. .table files are organized using a KNIME proprietary format, including the full file structure and are optimized for space and speed - providing maximum performance with minimum configuration!

Google Sheets Reader

Google Sheets Reader: Reads data from a Google Sheet file. Authentication occurs on the Google site. Google credentials are not saved within the KNIME workflow.

knime:// protocol: References a file path relative to some key location of the current KNIME installation like *knime://knime.workflow/../<filename> or knime://<knime.-server.mountpoint>/<path>/<filename>*



TRANSFORM

GroupBy

GroupBy: Groups the rows of a table by the unique values in selected columns and calculates aggregation and statistical measures for the defined groups. Despite its simple name, it offers powerful functionality and has many unsuspected usages. For example - row deduplication.

Pivoting

Pivoting: Extends the aggregation functionality of the GroupBy node by creating an output data table with columns and rows for the unique values in selected input columns. Note: the unique values of the grouping column become rows and the unique values of the pivoting column become columns.

Rule Engine

Rule Engine: Applies a set of rules to each row of the input data table. All Rule Engine operators are also available in the Column Expressions node.

Partitioning

Partitioning: Splits data into two subsets according to a sampling strategy. This node is generally used to produce a training and a test set to train and evaluate a machine learning model.

Row Filter

Row Filter: Filters rows in or out from the input data table according to a filtering rule. The filtering rule can match a value in a selected column or numbers in a numerical range.

Math Formula

Math Formula: Implements a number of math operations across multiple input columns, from simple sum and average, to logarithms and exponentials. All Math Formula operators are also available in the Column Expressions node.

String to Date&Time

String to Date&Time: Converts values in a String column into Date&Time values. The Date&Time format contained in the String values can be manually defined or auto guessed.

Cell Splitter

Cell Splitter: Splits values in a selected column into two or more substrings, as defined by a delimiter match. Delimiter is a set character, such as a comma, space, or any other character or character sequence.

Column Filter

Column Filter: Filters columns in or out from the input data table according to a filtering rule. Columns to be retained can be manually picked or selected according to their type, or of a regex expression matching their name.

Column Rename

Column Rename: Assigns new names and types to selected columns, as configured in the dialog.

Joiner

Joiner: Joins rows from two data tables based on common values in one or more key columns. The most common join types are possible: inner join, left outer join, right outer join, and full outer join.

Sorter

Sorter: Sorts the table in ascending or descending order based on the values of a chosen column. In addition, it is possible to sort based on multiple columns.

Concatenate

Concatenate: Merges vertically two data tables, by piling up cells in columns with the same name. Cells in uncommon columns are filled with missing values. The Concatenate (Optional in) node merges vertically up to four data tables.

Missing Value

Missing Value: Defines a strategy to deal with missing values in the input data table - either globally on all columns, or individually for each single column.

String Manipulation

String Manipulation: Performs operations on String values in columns, such as combining two or more Strings together, extracting one or more substrings, trimming blank spaces, and so on. All operators are also available in the Column Expressions node.

DEPLOY

Data to Report

Data to Report: Marks the data table to be exported to BIRT - a partially open source reporting tool integrated within KNIME. When switching from KNIME to BIRT, the marked data sets are imported into BIRT. The Image To Report node marks the input images to be exported to BIRT.

Excel Writer (XLS)

Excel Writer (XLS): Writes the input data table to a sheet in an Excel file (XLS or XLSX).

Table Writer

Table Writer: Writes the input data table to a file using the .table KNIME proprietary format. This format includes the full file structure and is optimized for space and speed. Including the table structure in the file is a great advantage - especially when exchanging data files among users.

CSV Writer

CSV Writer: Writes the input data table to a CSV file.

Google Sheets Writer

Google Sheets Writer: Writes the input data table into a Google Sheet file. Authentication occurs on the Google site. Google credentials are not saved within the KNIME workflow.

Send to Tableau Server

Connectors to Tableau: Export input data table into a Tableau file or server for reporting.

Resources

- **KNIME Forum:** Join our global community and engage in conversations at forum.knime.com
- **KNIME Books:** More tips, ideas, and lessons from knime.com/knimepress
- **KNIME Events:** Take a course, attend a workshop, or join a meetup at knime.com/events
- **KNIME Blog:** Engaging topics, challenges, industry news, and knowledge nuggets at knime.com/blog
- **Workflow Hub:** Browse our example workflows and/or share your own workflows. Show appreciation for others by adding ratings, or comments at workflows.knime.com
- **More Guides:** Still using SAS or Excel? Transition to KNIME Analytics Platform with these handy guides at knime.com/knimepress
- **KNIME Server:** For team-based collaboration, automation, management, and deployment check out KNIME Server at knime.com/server