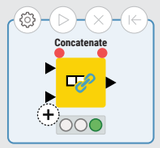
Cheat Sheet: Building a KNIME Workflow for Beginners

Open for Innovation
KNIME

Getting started with KNIME Analytics Platform

• Use the Getting Started Guide to take your first steps with visual workflows at: ***www.knime.com/getting-started-guide***
• Learn more about included nodes and explore working examples in the **KNIME Analytics Platform Version 5 Starter Perspective Collection** on **KNIME Community Hub**.



Node Action Bar: Interact directly with the node to, e.g., configure, execute, cancel or reset a node.
Configure: Open the configuration dialog.
Execute: Executes the node.
Cancel: Cancels the execution of the node.
Reset: Resets the node.
Node Labels: Double click *"Add comment"* below the node to add a comment/label.
Dynamic ports: Additional input ports can be added by clicking the plus on the left side of the node.

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.

Error: The node has encountered an error during execution.

READ

Reads CSV files. It has an auto-detect function to automatically guess the file structure. As for other reader nodes, you can add a "File system connection" input port to connect to different data sources.

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

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!

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

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.

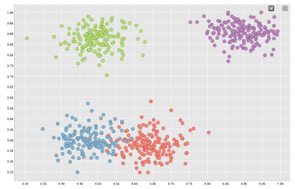
Reads data from a Google Sheet file. Authentication occurs on the Google site. Google credentials are not saved within the KNIME workflow.

In reader and writer nodes, the file path is expressed relatively to a key location of the current KNIME installation, like workflow, workflow data area, and mountpoint.

EXPLORE

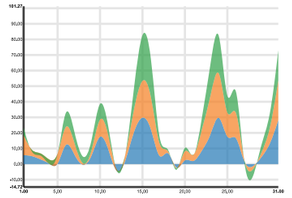
(All visualizations are interactive)

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.



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.

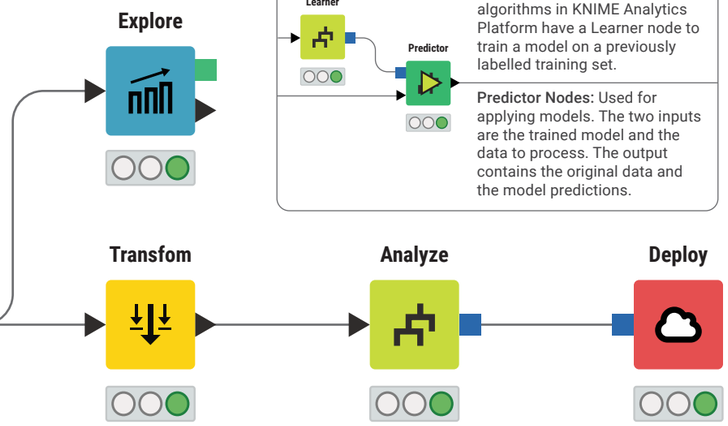
Provides an interactive view to summarize the statistics of the input data via statistical measures and histograms - for both numerical and nominal columns.

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.

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

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.

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.



Learner Nodes: Supervised algorithms in KNIME Analytics Platform have a Learner node to train a model on a previously labelled training set.

Predictor Nodes: Used for applying models. The two inputs are the trained model and the data to process. The output contains the original data and the model predictions.

DEPLOY

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.

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

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.

Writes the input data table into a CSV file or to a remote location denoted by an URL.

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.

Exports the input data table into a Tableau file or server for reporting.

ANALYZE

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.

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.

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.

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

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.

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).

TRANSFORM

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.

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.

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.

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

Merges two or more data tables vertically by piling up cells in columns with the same name. Cells in not overlapping columns are filled with missing values.

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.

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.

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.

Joins rows from two data tables based on common values in one or more key columns. The output - inner join, left outer join, right outer join, full outer join, or the respective antijoins - can be split into multiple output tables.

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

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.

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.

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.

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.

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.

Resources

• **E-Books**: KNIME Advanced Luck covers advanced features & more. Practicing Data Science is a collection of data science case studies from past projects. Both available at ***knime.com/knimepress***

• **KNIME Blog**: Engaging topics, challenges, industry news, & knowledge nuggets at ***knime.com/blog***

• **E-Learning Courses**: Take our free online self-paced courses to learn about the different steps in a data science project (with exercises & solutions to test your knowledge) at ***knime.com/knime-self-paced-courses***

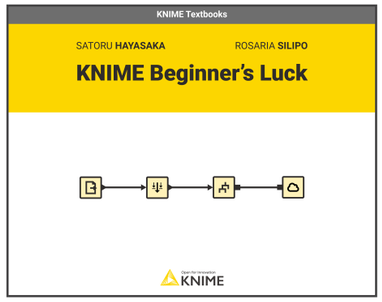
• **KNIME Community Hub**: Browse and share workflows, nodes, and components. Add ratings, or comments to other workflows at ***hub.knime.com***

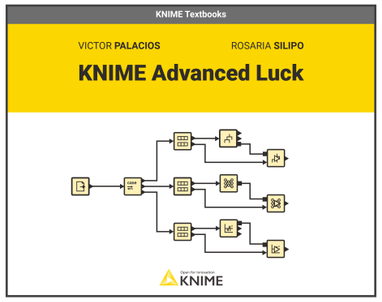
• **KNIME Forum**: Join our global community & engage in conversations at ***forum.knime.com***

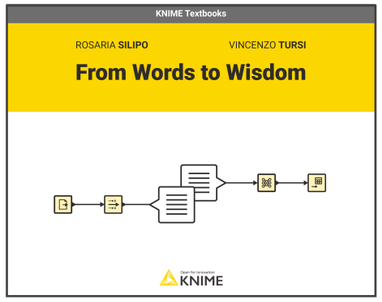
• **KNIME Business Hub** : For team-based collaboration, automation, management, & deployment check out KNIME Business Hub at ***knime.com/knime-business-hub***

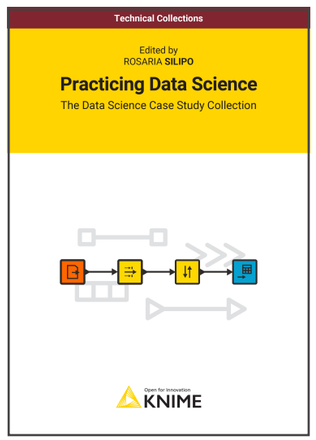
KNIME Press

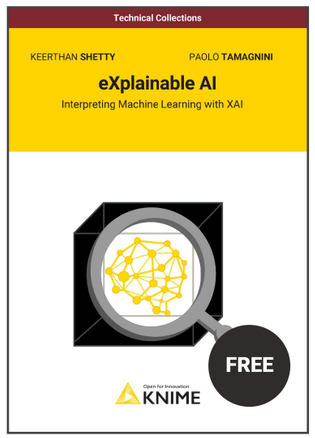
Extend your KNIME knowledge with our collection of books from KNIME Press. For beginner and advanced users, through to those interested in specialty topics such as topic detection, data blending, and classic solutions to common use cases using KNIME Analytics Platform - there's something for everyone. Available for download at www.knime.com/knimepress.

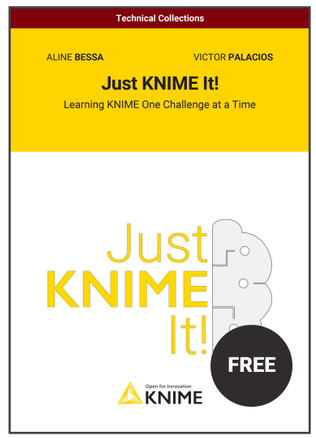


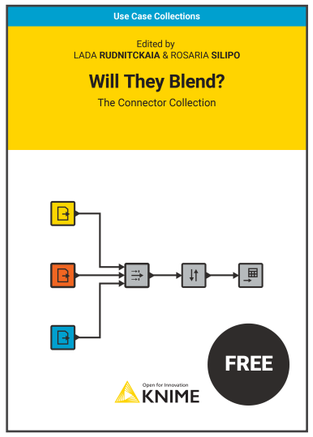


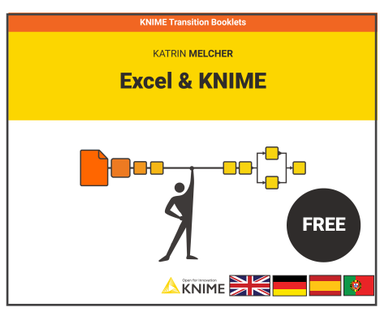


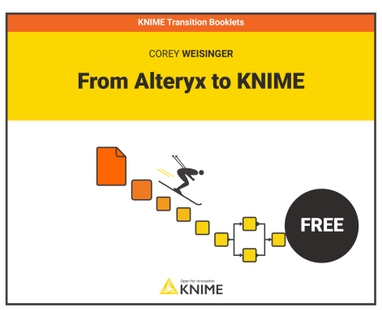


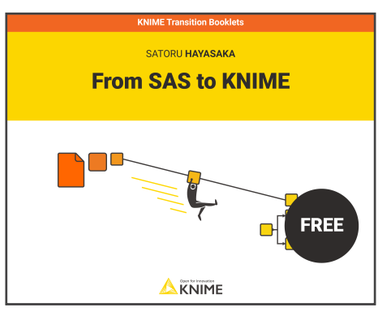


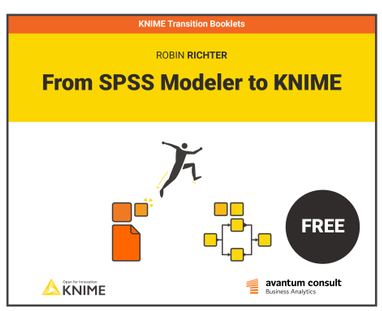












Need help?
Contact us!

© 2023 KNIME AG. All rights reserved. The KNIME® trademark and logo and OPEN FOR INNOVATION® trademark are used by KNIME AG under license from KNIME GmbH, and are registered in the United States. KNIME® is also registered in Germany.