KNIME Analytics Platform: advanced functionalities and basic Machine Learning

Data manipulation and aggregation processing steps are often used to transform the input data before machine learning methods are applied. Data mining and machine learning algorithms are then used to generate predictive and/or descriptive data models. Descriptive models help to explore and understand historical data and to extract interesting patterns from them. Predictive models are applied to new data to make predictions about a specific target attribute.

Generating data models can be as simple as a single node, or it may require a more complex workflow that combines several nodes.

Moreover, the estimation of the accuracy of a data model is often useful (or required) before the data model is deployed in a production environment. Model evaluation requires a complex workflow that must follow appropriate methodologies.

Advanced KNIME functionalities

#	Type	Functionality	Task
1	Workflow control and parametrization	Flow Variables	Flow Variables allow you to pass information between nodes using programming variables represented by red dots on nodes and red edges between nodes.
2	Workflow control	Loops	Iterate a workflow: loops require a start loop node and an end loop node to delimit the nodes included in the iteration block.
3	Workflow control	Control structures	Conditional workflow execution: If-switch, Case-switch, Try-Catch
4	Workflow Abstraction and modularization	Metanodes	They allow to build new reusable nodes that encapsulate an arbitrarily complex workflow inside them. Flow variables have global scope.
5	Workflow Abstraction, modularization and parametrisation	Component (Wrapped Metanodes)	Upgraded version of Metanode: it can use Quickforms for parametrisation, has flow variables with local scope, can filter imported/exported variables, can be used with the KNIME Web portal (server), can rename I/O ports, etc.
6	Reusability	Shared Component (Workflow Templates)	Exported Metanodes that are stored in the workspace (file system) for later reuse and sharing.
7	User input/output and parametrisation	Abstractions (Quickform nodes)	Nodes for component configuration, component input/output and visualization widgets. These nodes explicitly model the parameters of the component configuration, which data are passed in (or out) of the component, and which visualisations are used to compose the component’s view.

Useful KNIME nodes for advanced analytics

#	Type	Node	Task
1	Data source	Data Generator	Create an artificial data set
2	Data source	Table Creator	Interactively create a data table
3	Data source	List Files	List the files in a directory
4	Data Manipulation	Row ID	Reset the row ID of the rows of a data table
5	Attribute transformation	Normalizer	normalizes the values of the numeric columns (min-max, z-score or scaling)
6	Attribute transformation	PCA	Principal Component Analysis (PCA)
7	Flow Variable declaration	Table Row to Variable	create a flow variable from a table row
8	Quickform	String Input	Interactively create a flow variable
9	Quickform	Double Input	Interactively create a flow variable
10	Quickform	Date&Time Input	Interactively create a flow variable
11	Quickform	Value Selection	Interactively create a flow variable from the values in a given column
12	Clustering	k-Means	Partitional Clustering based on the k-Means algorithm: it uses the Euclidean distance, thus can only be applied to numerical attributes.
13	Clustering	k-Medoids	Partitional Clustering based on the k-Medoids algorithm for arbitrary distance functions.
14	Clustering	Hierarchical Clustering	Hierarchical Clustering based on a top-down (divisive) approach
15	Regression	Linear regression Learner and Regression Predictor	Multivariate linear regression
16	Regression	Simple regression Tree Learner and Predictor	Multivariate regression based on a single regression tree according to the algorithm CART (Classification and Regression Trees)
17	Classification	Random Forest learner and predictor	Bag of Decision Trees: random subspace method (aka attribute bagging)
18	Classification	Tree Ensemble learner and predictor	Random Forest variant providing both attribute and record bagging: random subspace method and bootstrap aggregating
19	Classification	Gradient Boosted Trees learner and predictor	An ensemble of decision trees based on boosting
20	Classification	Decision Tree learner and predictor	Decision Trees

21	Performance estimation	Scorer	Typically used for computing Classification accuracy
22	Performance estimation	Numerical Scorer	Typically used for computing performance statistics of Regression (incl. the coefficient of determination R^2)
23	Performance estimation	Entropy Scorer	Typically used for computing Clustering validity indeces
24	Data manipulation for performance estimation	Partitioning	Splits the input data table into two disjoint tables for the hold-out method for the estimation of the accuracy of a predictive model.
25	Data manipulation for performance estimation	X-Partitioner and X-Aggregator	The cross-validation method for the estimation of the accuracy of a predictive model.
26	Optimization	Parameter Optimization Loop Start and Loop End	Exhaustive (brute force) and heuristic methods to find optimal values for parameters. (Note: it requires the KNIME Optimization Extension)