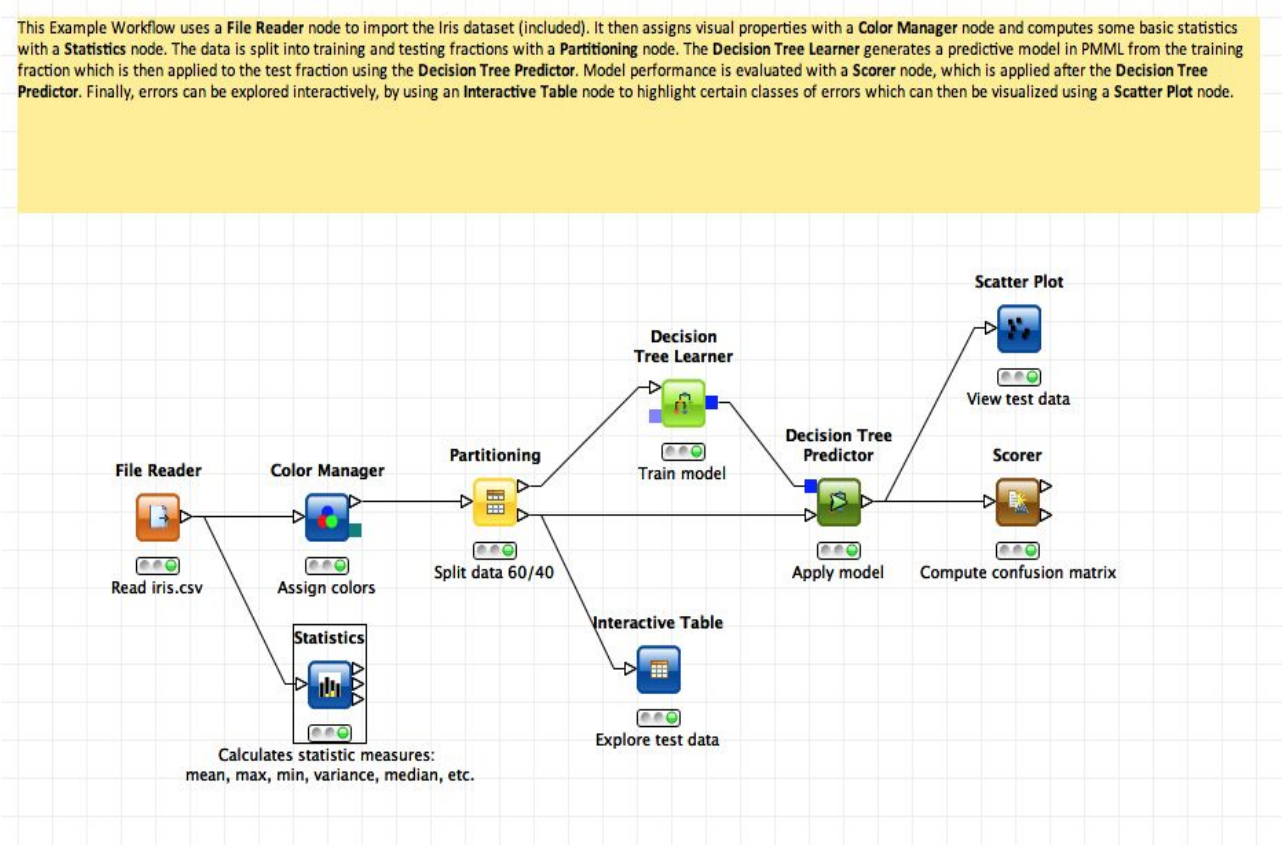
Building a Decision Tree in KNIME

This reading covers the following topics:

- Sample workflow for building a decision tree in KNIME
- Nodes commonly used for data preparation
- Node names and workflow annotation

A sample workflow to build a decision tree for classification can be found in KNIME. It is called Example Workflow under LOCAL in the KNIME Explorer. The workflow is shown below:



Example Workflow Nodes

The nodes in the above workflow are:

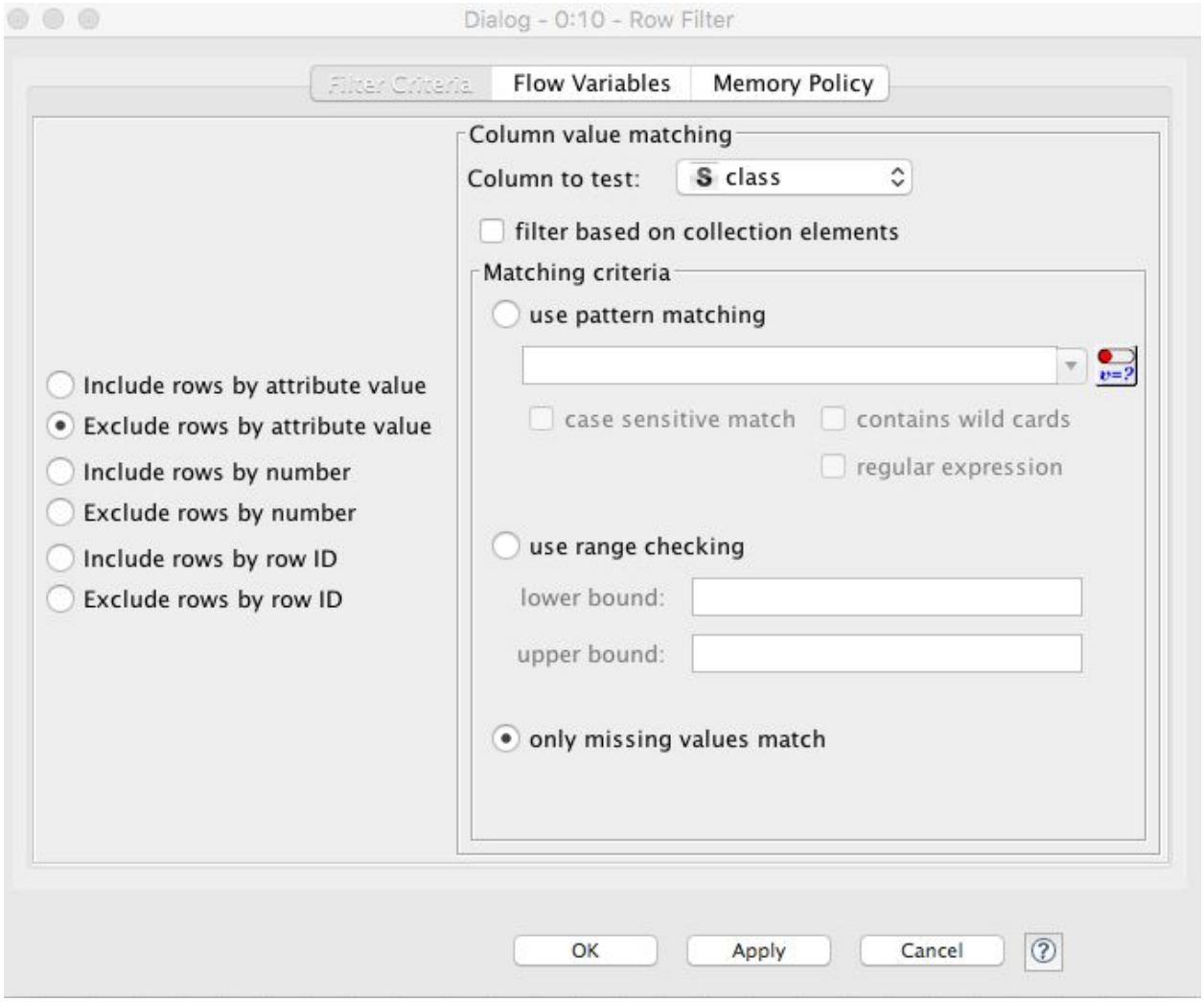
- File Reader
    - Reads data from an ASCII file or URL location.
    - After executing the node, right-click on it and select 'File Table' to view the data. The number of rows is displayed at the top.
- Statistics
    - Calculates statistical measures such as min, max, mean, standard deviation, etc.
    - You may want to increase 'Max no. of possible values per column (in output table)' to see all samples on the statistics view.

- Color Manager
    - Assigns colors to attributes to visualize results.
    - For example, you can assign different colors to different classes for the target attribute.
- Partitioning
    - Partitions the dataset into train and test data. This is important because you want to test your model on data that was not used to train (i.e., create) it.
    - Stratified sampling should be selected. This specifies that the distribution of classes will be (approximately) the same between the original dataset and the resulting data partitions.
    - Use random seed should also be selected. This ensures that you will get the same partitions every time you execute this node. This is important to get reproducible results. NOTE: This is not set by default, so you will need to set it when you use this node.
- Decision Tree Learner
    - Creates a decision tree model using the input data.
    - Pruning method should be set to MDL. Pruning reduces the tree size and usually leads to better generalization performance. NOTE: This is not set by default, so you will need to set it when you use this node.
- Decision Tree Predictor
    - Applies a trained decision tree model to input data, which should be test data (i.e., new data not used to create the model).
- Scorer
    - Compare two attributes and computes the confusion matrix. For classification, the predicted attribute and target attribute are compared.
    - Right-click and select View: Confusion Matrix to see the confusion matrix and overall accuracy.
    - Confusion matrix explanation:
- Scatter Plot
    - Creates scatter plot of two attributes.
    - Right-click and select ‘View: Scatter Plot’. Then in the node dialog, click on Column Selection to select attributes to display on plot.
- Interactive Table
    - Displays in a table view.
    - Rows can be selected and highlighted. Highlighted rows also show up in Scatter Plot. This can be used to analyze certain samples or types of errors.

Common Data Preparation Nodes

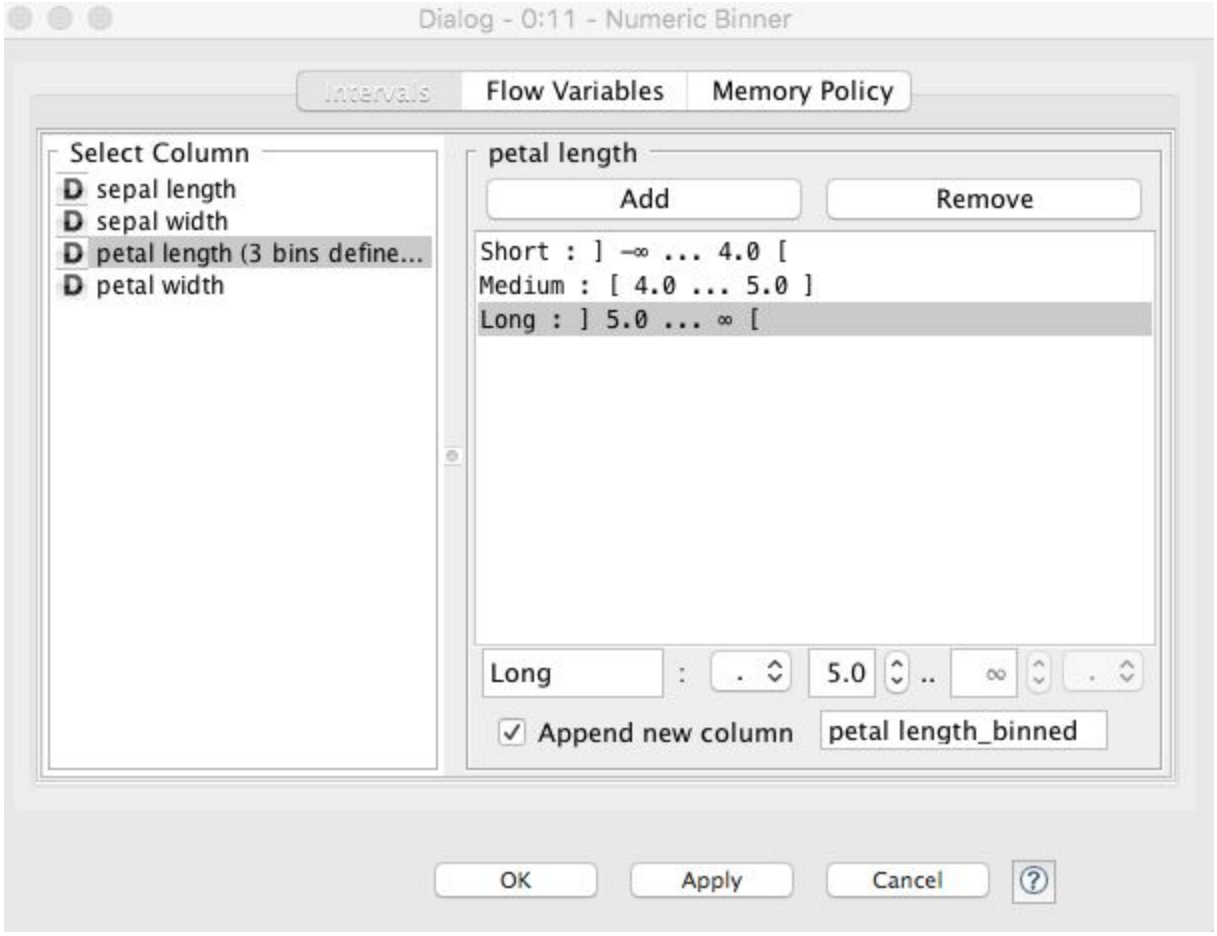
KNIME also provides several nodes for data manipulation. Some nodes commonly used for data preparation are the following:

- Row Filter
    - Removes rows matching specified criteria from the input data.
    - To filter rows with missing values (e.g., samples without a value for 'class' in the Iris dataset), configure the node as follows:

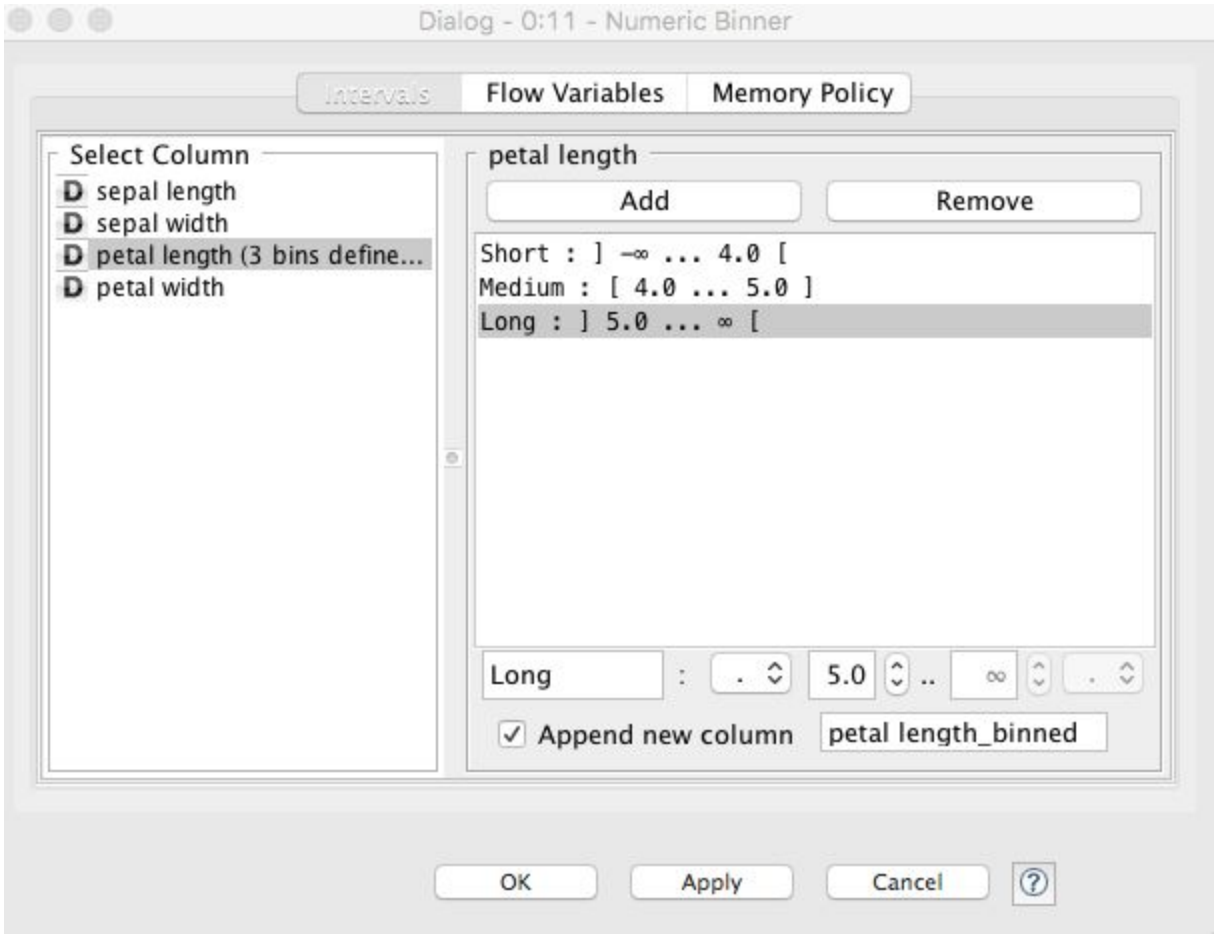


    - Right-click and select 'Filtered' to see the resulting data with the specified rows removed. The number of rows is displayed at the top.

- Column Filter
    - Removes columns (i.e., attributes) from the input data.
    - To remove attributes ‘sepal length’ and ‘sepal width’ for the Iris data, configure this node as follows:



- Numeric Binner
    - Converts a numeric attribute to a categorical attribute by binning the values.
    - Bins are defined for an attribute by specifying the beginning and end points of the interval for each bin. In specifying the interval endpoints, [means inclusive, and] means exclusive.
    - Check ‘Append new column’ to add the new binned attribute while still keeping the original attribute. This is recommended so that the two attributes can be compared to check that the intervals were defined correctly. If ‘Append new column’ is not checked, the original attribute is replaced by the new binned attribute.
    - To create a new categorical attribute ‘petal_length_binned’ with three bins from the numerical attribute ‘petal_length’ for the Iris dataset, configure this node as follows:



- After executing the node, right-click and select ‘Binned Data’ to see the data with the newly created categorical feature.

Node Name and Workflow Annotation

The text under a node is referred to as the node name. The default name is something similar to “Node 1”. Be sure to modify this field to accurately reflect the purpose of the node (e.g., “Read Iris Data”).

Similarly, the yellow textbox at the top of a workflow is the workflow annotation. This should provide a brief description to explain the purpose of the workflow.