Classification Using Decision Tree in KNIME

Learning Objectives

By the end of this activity, you will be able to perform the following operations in KNIME:

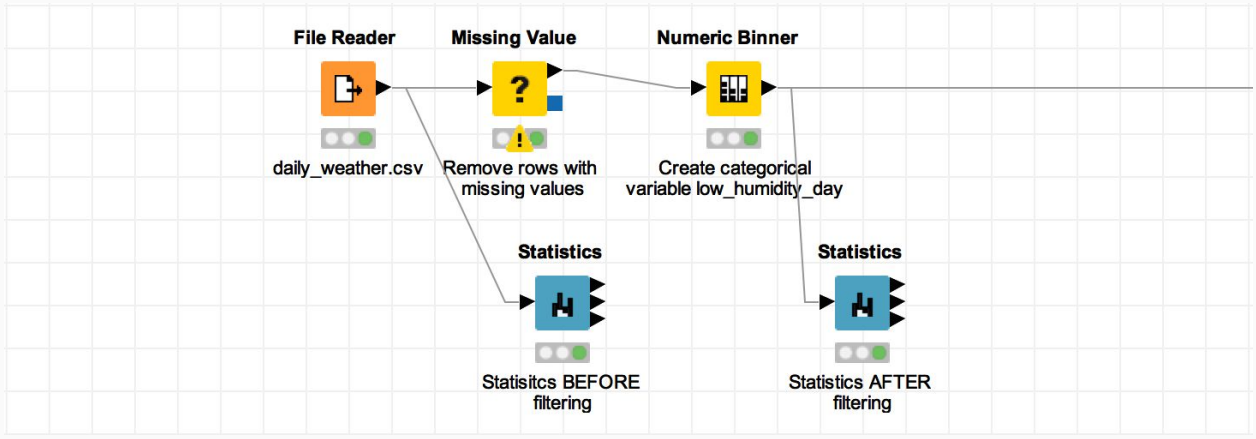
1. Create a categorical variable from a numeric variable
2. Examine the summary statistics of a dataset
3. Build a workflow for a classification task using a decision tree

Problem Description

Now that we have explored the data and looked at how to handle missing values, the next step is to build a classification model to predict days with low humidity. Recall that low humidity is one of the weather conditions that increase the dangers of wildfires, so it would be helpful to be able to predict low-humidity days. We will build a decision model to classify low-humidity days vs. non-low-humidity days based on weather conditions observed at 9am.

Steps

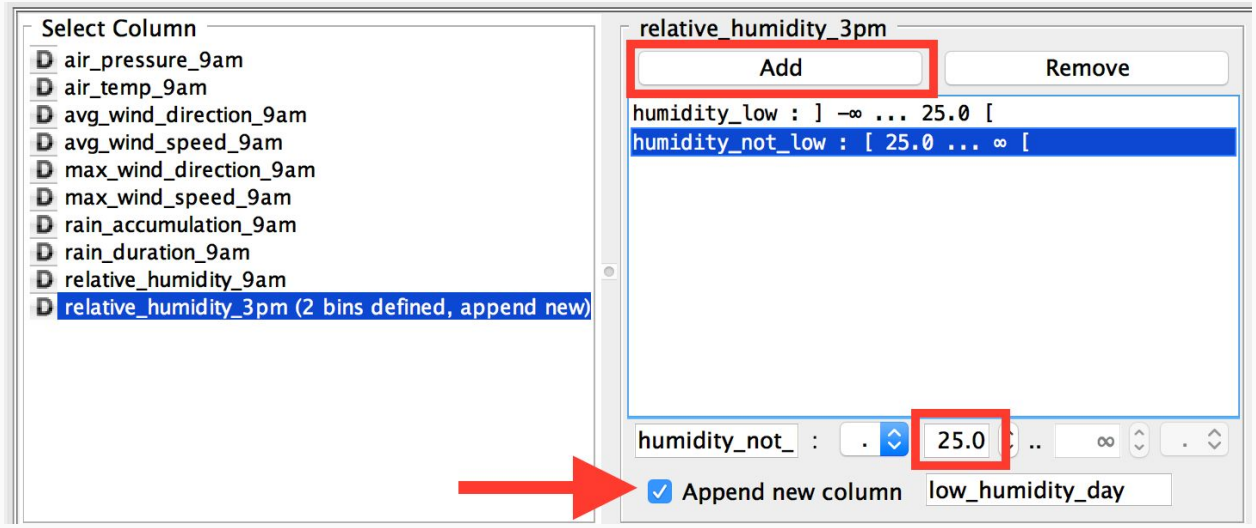
Prepare the data



Let's build a workflow to build a decision tree model to classify low-humidity days vs. days with normal or high humidity. The model will be used to predict low-humidity days

1. Start a new workflow in your local workspace.
2. Use a **File Reader** node to import the daily_weather.csv dataset. Use the configuration dialog to specify the location of the daily_weather.csv file.
3. Connect a **Missing Value** node to the File Reader node. This will handle the missing values that the dataset contains so the data can be analyzed properly. In the configure dialog, in the **Default** tab, choose **Remove Row*** to remove all rows with missing values.
4. As with the Data Exploration Hands-On, use the **Numeric Binner** node to create a new categorical variable with the condition "*if relative_humidity_3pm < 25% then humidity_low is true, else humidity_not_low is true*".

- Locate the **Numeric Binner** node, which is in the Manipulation>Column>Binning category. Drag it to the Workflow Editor, and connect it to the **Missing Value** node.
- Open the Configure Dialog for the Numeric Binner node. Select **relative_humidity_9am**, and **Add** 2 bins. Make one bin called "humidity_low" with the range **-∞ to 25** excluding 25, and another called "humidity_not_low" with the range**25 to** ∞. The endpoint brackets specify that humidity_low excludes 25.0, while humidity_not_low includes 25.0. This is necessary to capture the condition "if relative_humidity_3pm < 25% then low_humidity_day=1, else low_humidity_day=0". Click the checkbox to "**Append new column**" and name it **low_humidity_day**.

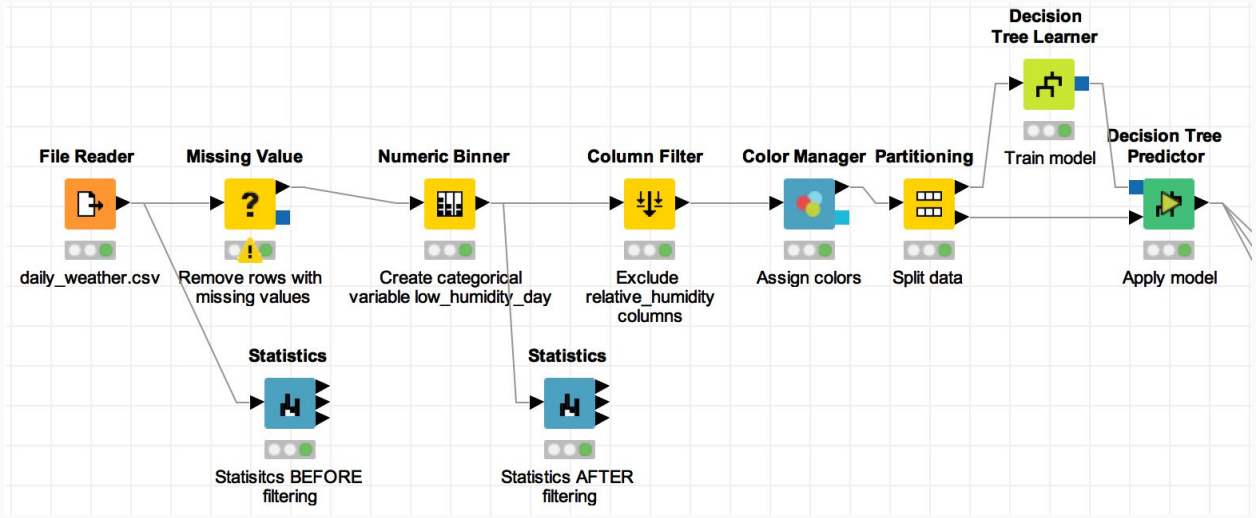


Examine Summary Statistics

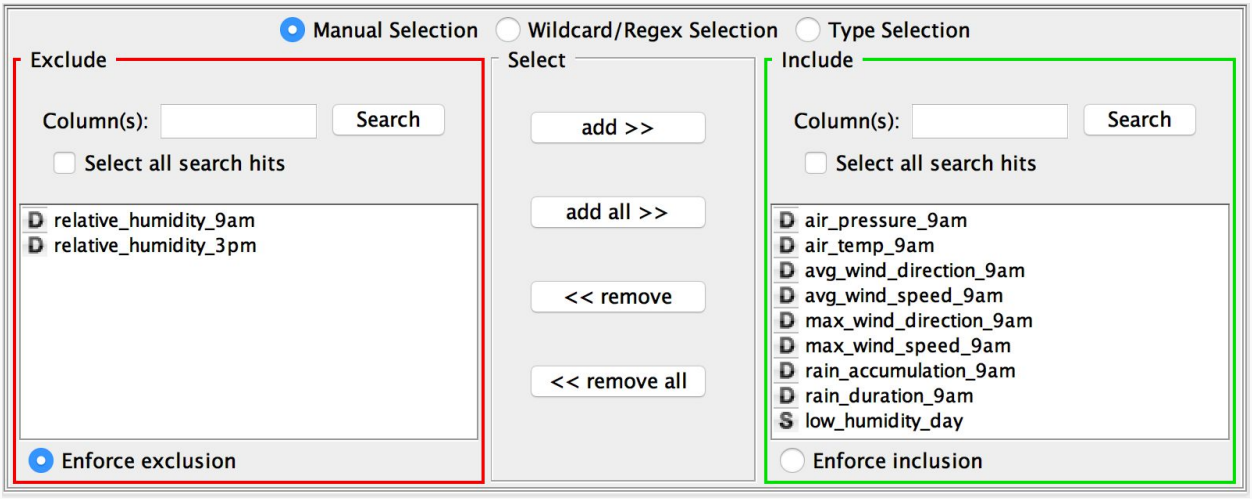
Before we build the workflow any further, let's use some **Statistics** nodes to check a few things about our processed data.

1. Connect a **Statistics** node to the output of the **File Reader** node. In the Configure Dialog of the Statistics node, change **Max no. of possible values per column (in output table)** to **1,500**, and **add all >>** columns to the **Include**side. Rename this node to "Statistics BEFORE filtering".
2. Connect a **Statistics** node to the output of the **Numeric Binner** node. In the Configure Dialog of this Statistics node, change **Max no. of possible values per column (in output table)** to **1,500**, and **add all >>** columns to the **Include**side. Rename this node to "Statistics AFTER filtering".
3. Execute and view both Statistics nodes, and you should see the resulting histograms have the same features in both. This is to ensure that the way we handled the missing values did not skew our data. Now we can check the following:
4. There are missing values for many of the variables in the "Statistics BEFORE filtering" node, but zero missing values in the "Statistics AFTER filtering" node.
5. The distributions of each variable in both "Statistics BEFORE filtering" and "Statistics AFTER filtering" should be about the same. You can spot check a couple of variables by looking at histograms, min, max, mean, and standard deviation.
6. In the "Statistics AFTER filtering" node, look at the **Nominal** tab to see the distribution of low_humidity_day. This shows that the samples are equally distributed between low-humidity days and days with normal or high humidity.

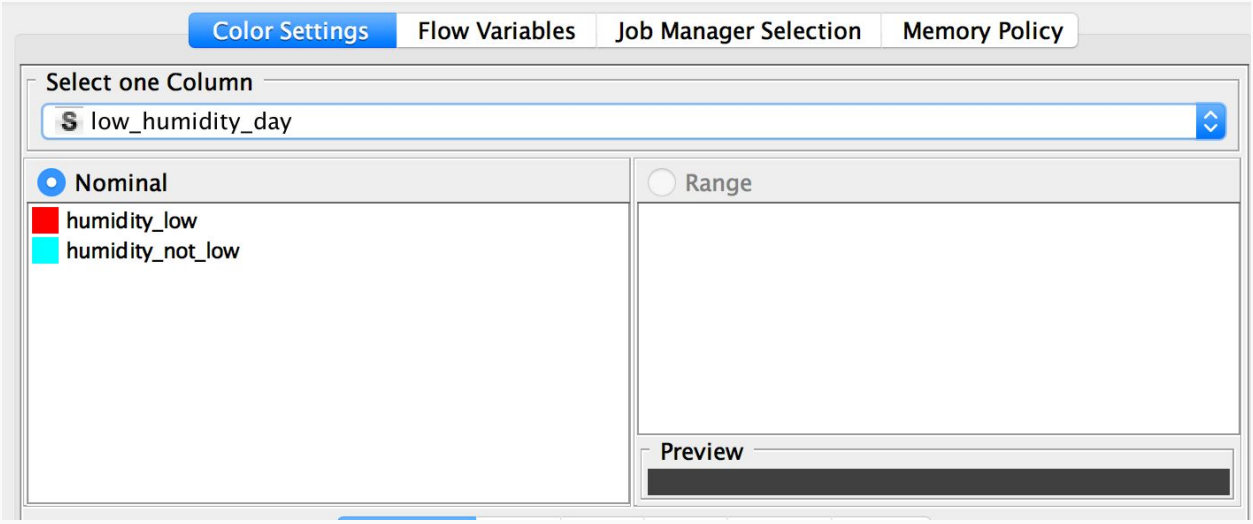
Build a Decision Tree Workflow



1. Connect a **Column Filter** node to the **Numeric Binner** node. In the Configure Dialog of the Column Filter node, exclude only the **relative_humidity_9am and relative_humidity_3pm** columns.

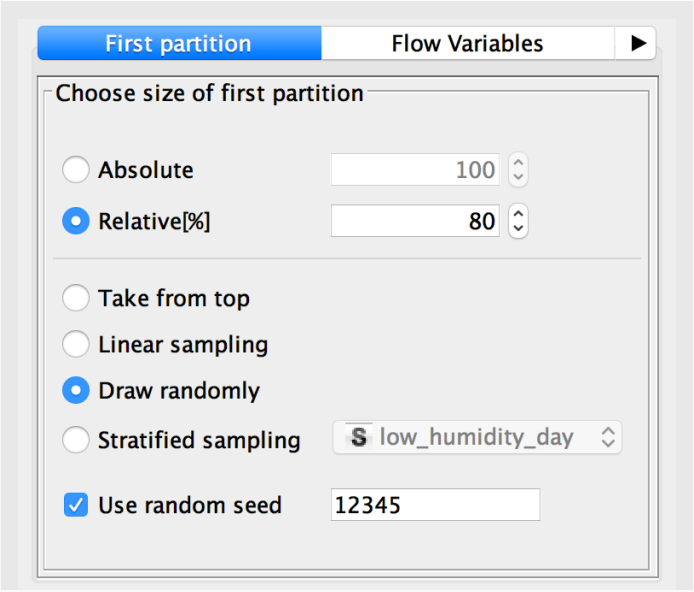


2. Connect a **Color Manager** node to the **Column Filter** node. This will color-code our categorical low_humidity_day variable so it is easier to visualize later on in the workflow. In the Configure Dialog of the Color Manager node, check that for low_humidity_day, the humidity_low is colored red and the humidity_not_low is colored blue.



3. Connect a **Partitioning** node to the **Color Manager** node. The Partitioning node is needed to split the data into training and testing portions. Training data is used to build the decision tree, and test data is used to evaluate the classifier on new data. In the Configure Dialog of the Partitioning node, choose **Relative[%] 80**, **Draw Randomly**, and **Use random seed 12345**. This will randomly select 80% of the data to the 1st output (the training data), and the rest to the 2nd output (the test data).

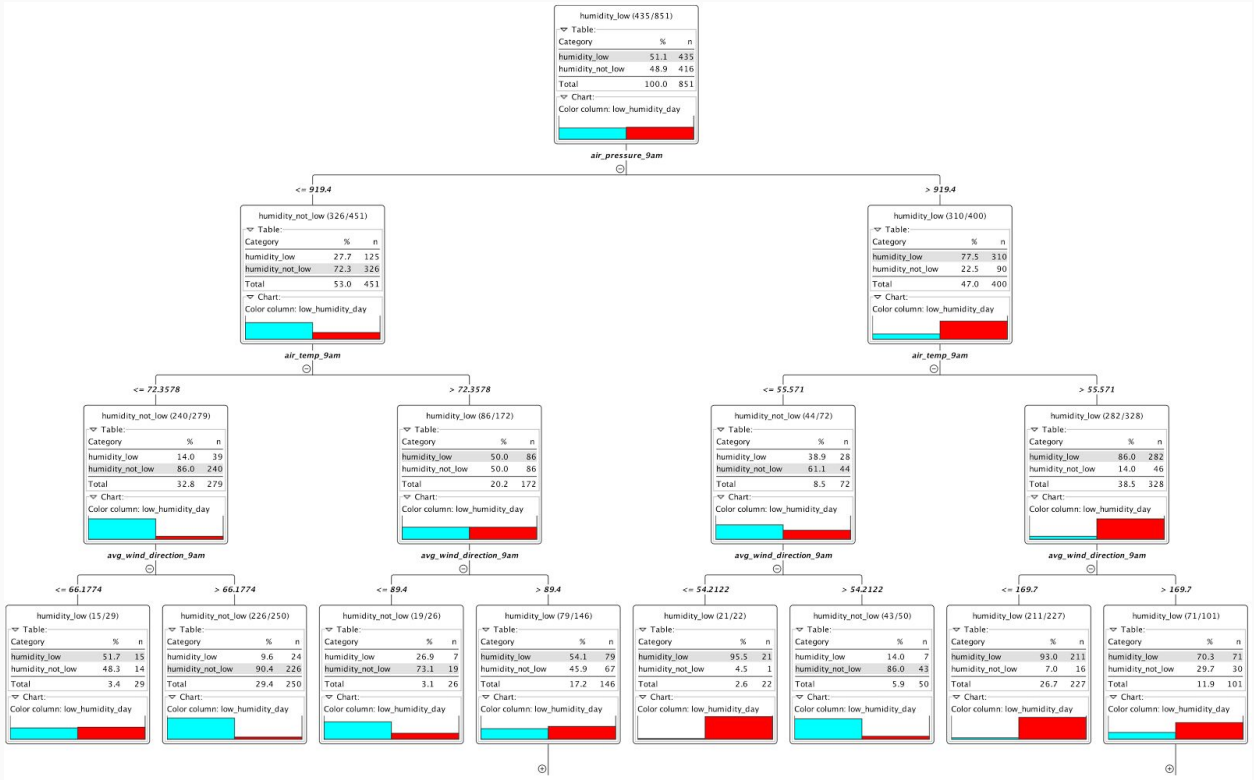
The random seed is set so that everyone can get the same training and test data sets to train and test the decision tree model for this exercise.



4. Connect a **Decision Tree Learner** node to the 1st output of the **Partitioning** node. This node will generate the decision tree classifier using the training data. In the Configure Dialog, change the **Min number of records per node** to **20**. This is a stopping criterion for the tree induction algorithm. It specifies that a node with this number of samples can no longer be split. The default value for this is 2, which is very small, and may result in overfitting.

5. Connect a **Decision Tree Predictor** to the 2nd output of the **Partitioning** node and to the output from the **Decision Tree Learner** node. This node will apply the model to the test data.

6. Execute the workflow. Right-click on the **Decision Tree Predictor** node and select 'View: Decision Tree View' to see the generated decision tree. You can zoom out and click the little '+' buttons to expand the nodes. The next Reading, **Interpreting a Decision Tree in KNIME**, describes how to interpret the resulting decision tree model.



Save Your Workflow

Save your workflow using <control>-s on Windows or <command>-s on Mac, or selecting File>Save or File>Save As.