Learning Objectives

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

1. Create and interpret a confusion matrix for a decision tree
2. Determine the accuracy rate of a decision tree model
3. Use highlighting to analyze classification errors

Problem Description

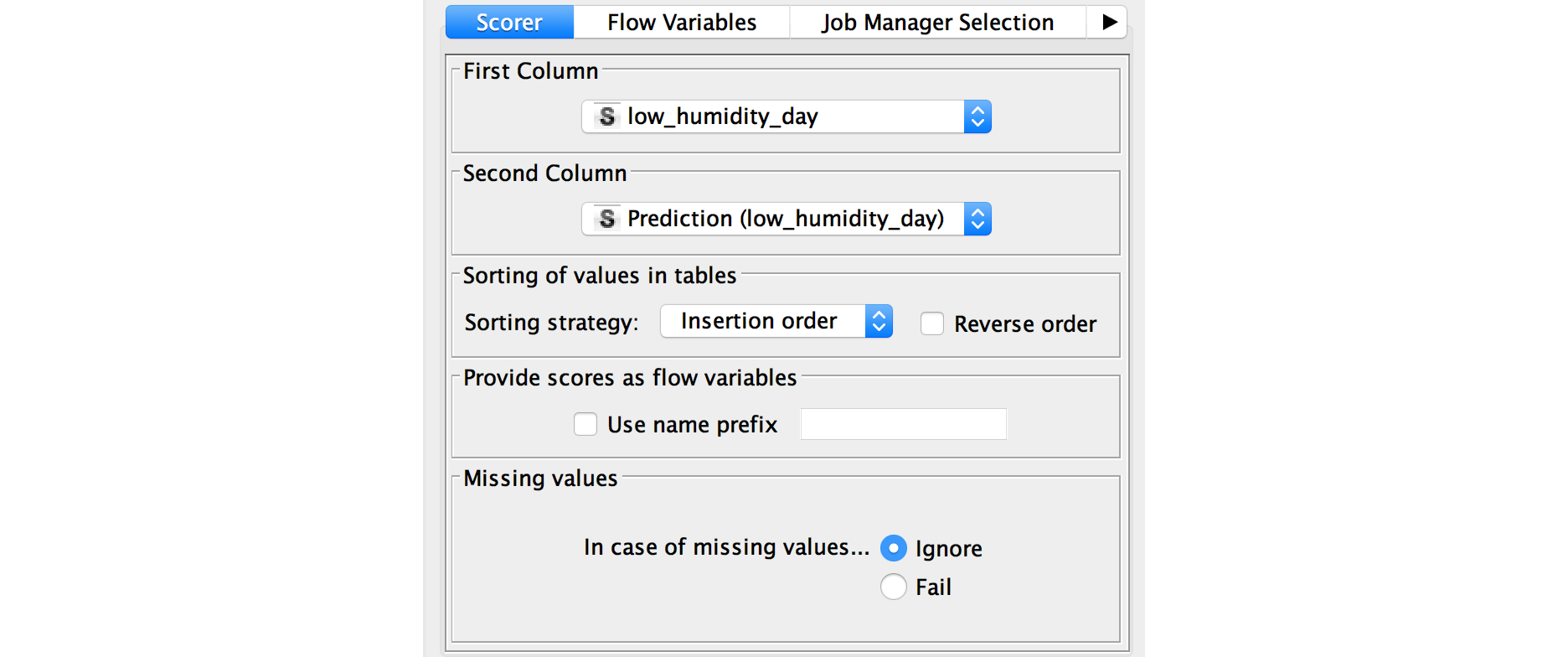
With the decision tree classifier built, we now need to evaluate its performance.

Steps

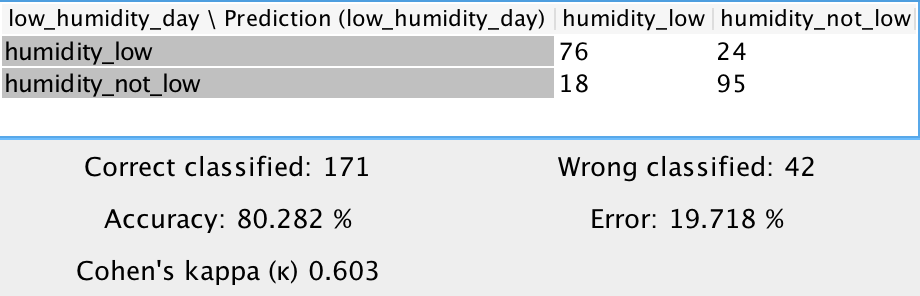
Generate a Confusion Matrix and Determine Accuracy Rate

A confusion matrix shows the type of errors and correct classifications that a classifier makes. It can be generated using a **Scorer** node.

1. Open the Decision Tree Workflow that you created from the Classification Hands-On reading.
2. Connect a **Scorer** node to the existing **Decision Tree Predictor**.
3. The Scorer Configure Dialog should look like this by default. Click OK.



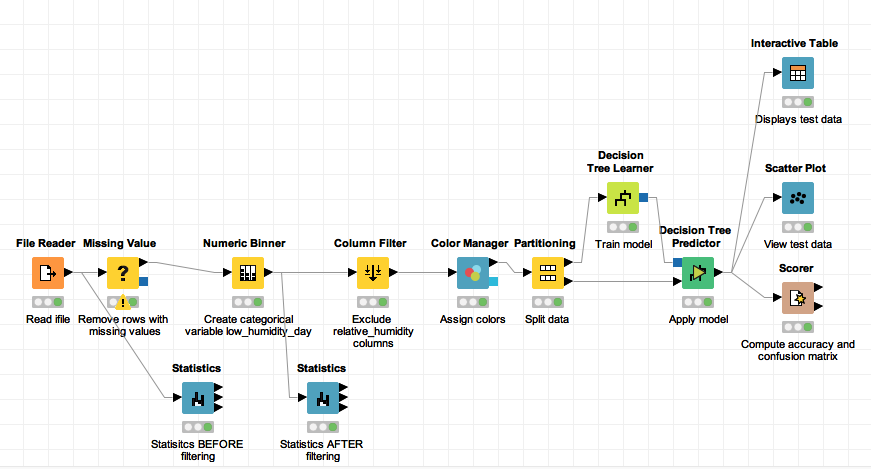
Execute and view the **Scorer** node. It shows the confusion matrix, along with the accuracy of the prediction. Here you should see an accuracy rate of 80.282% if you followed all the hands-on instructions.



From the confusion matrix, we see the following:

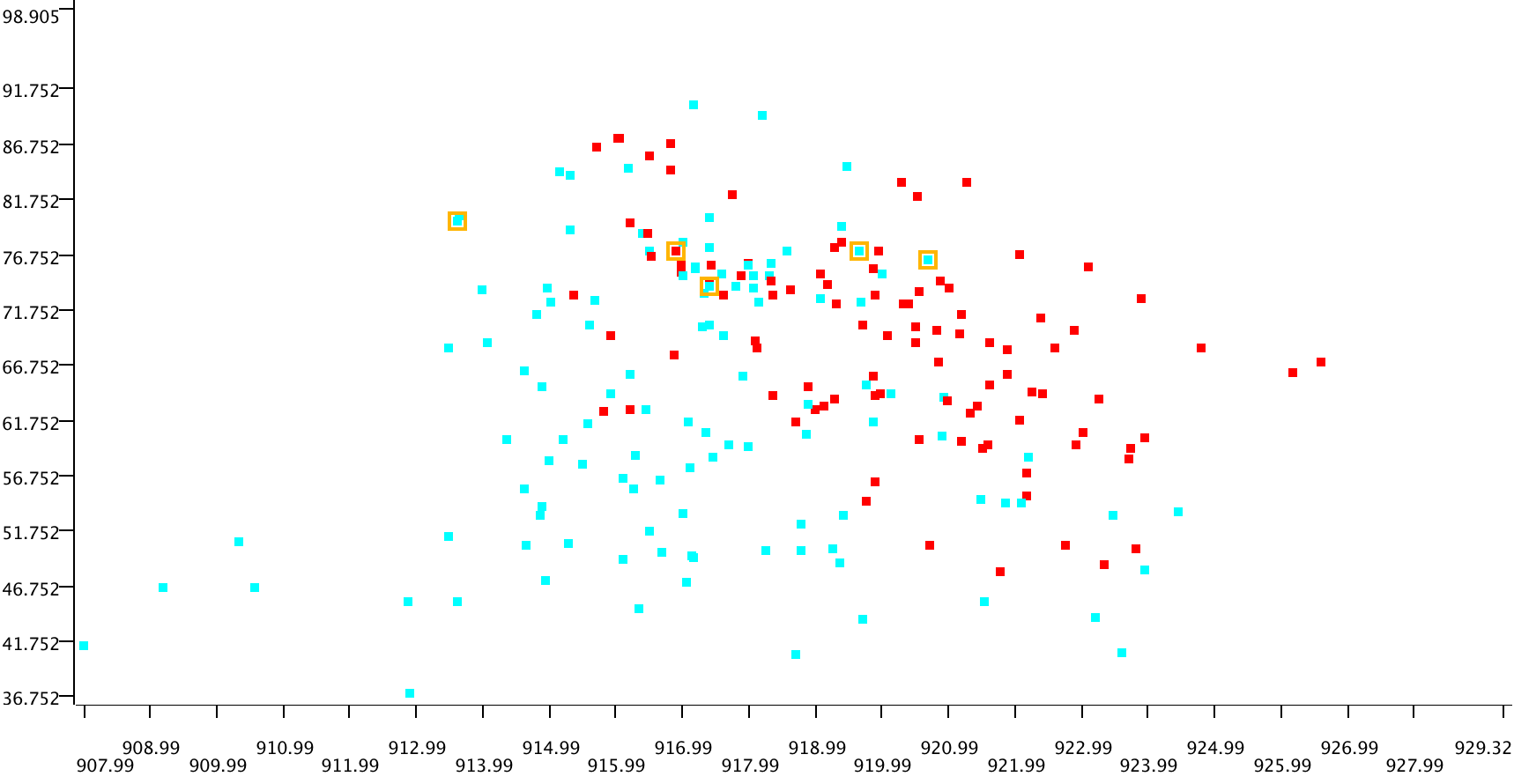
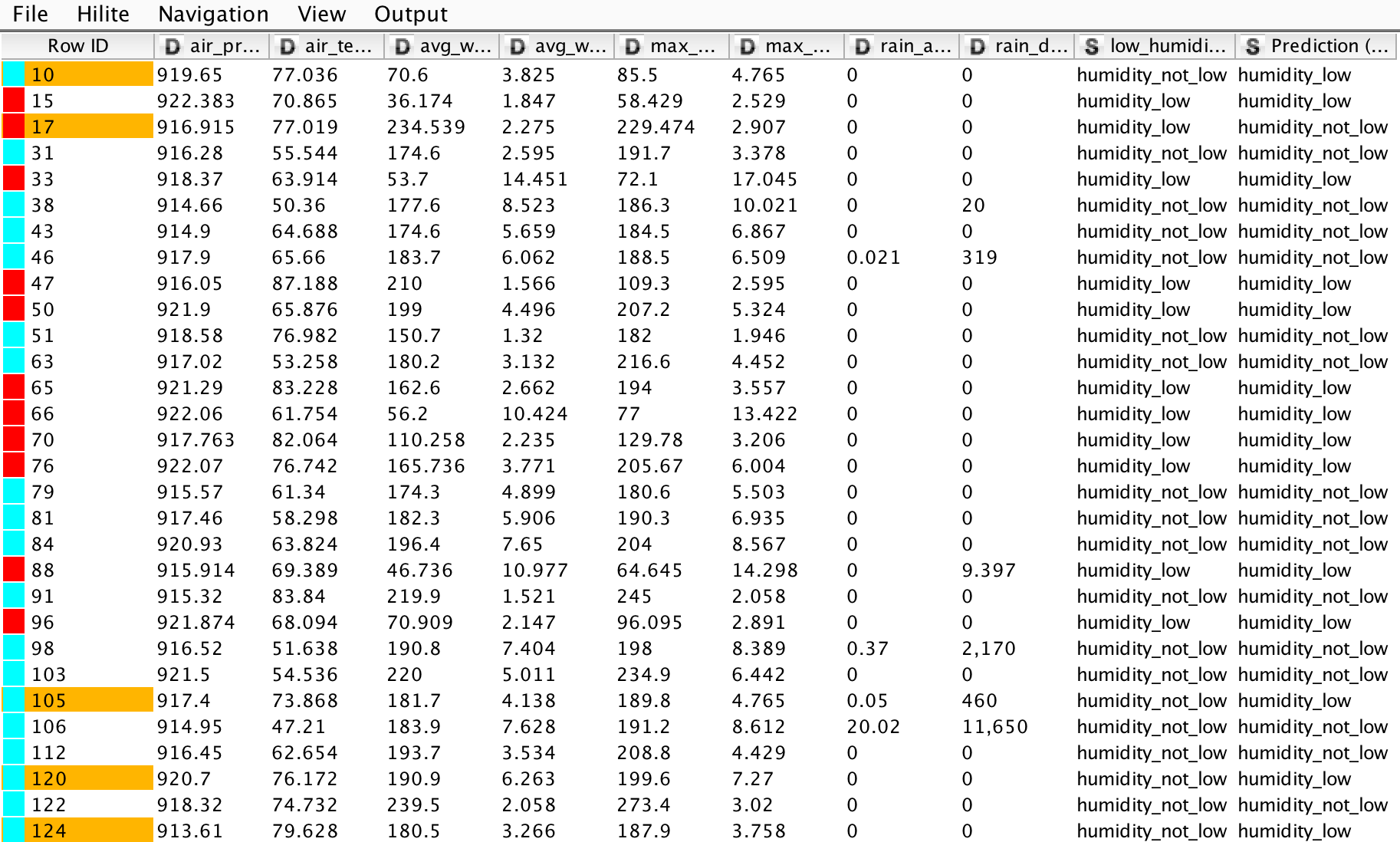
* There are 213 samples in the test data set (the sum of all the values in the confusion matrix)
* 76 humidity\_low samples with were correctly classified
* 95 humidity\_not\_low samples were correctly classified
* The accuracy rate is (76 + 95) / 213 = 171 / 213 = 80.282%
* 24 humidity\_low samples were incorrectly classified as humidity\_not\_low
* 18 humidity\_not\_low samples were incorrectly classified as humidity\_low
* The error rate is (24 + 18) / 213 = 42 / 213 = 19.718%

Use Highlighting and Scatter Plot to Analyze Classification Errors



A good way to enhance analysis of incorrect predictions is to visualize them. This can be accomplished using a feature called **hiliting**, and viewing the data in a **Scatter Plot** node.

1. Connect an **Interactive Table** node to the **Decision Tree Predictor**.
2. Execute and view this **Interactive Table** to see the input values for each sample (row), along with the ACTUAL/TRUE low\_humidity\_day value and the PREDICTED low\_humidity\_day value. The red and blue squares next to the Row ID color-codes the actual/true label (low or not). You can use this table to analyze samples whose true value differs from the predicted value (incorrect prediction).
3. Connect a **Scatter Plot** node to the **Decision Tree Predictor**.
4. Execute and view the **Scatter Plot** node, and place the window side-by-side with the **Interactive Table** window.
5. Go through the the table looking for rows with predictions that are different from the true value.
6. When you find such a row, click anywhere on that row. At the top of the window click **Hilite > Hilite Selected**. This will make that row yellow in the table and in the Scatter Plot. It may be easiest to use the up and down arrow keys to navigate the rows of the table. In this example, we are just going to highlight the first 5 misclassifications.
7. Do this for any row with a misclassification. This allows you to pinpoint the misclassified samples and analyze them further. Analyzing the misclassified samples can bring insight into how to improve model performance. For example, if many samples with avg\_temperature\_9am between 60 and 70 degrees are misclassified, this suggests that more samples with these values for avg\_temperature\_9am are needed to train the 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.