Interpreting a Decision Tree in KNIME

This document describes how to interpret a decision tree classifier. We will use the tree created in the Classification Using Decision Tree in KNIME Hands-On.

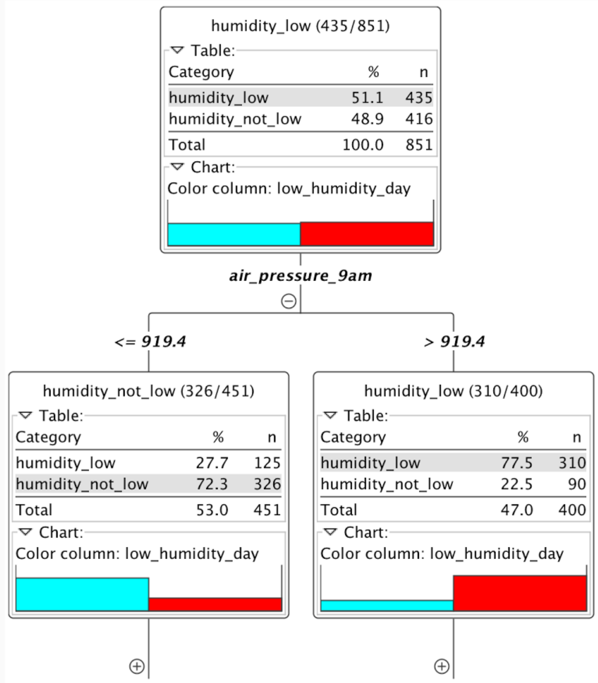
Classification Task

Recall that the task is to classify low-humidity days vs. days with normal or high relative humidity. The class label is based on the categorical variable **low_humidity_day**. This variable was created from the numeric variable relative_humidity_3pm. The class target low_humidity_day was created with the following categories:

- **humidity_low**: if relative_humidity_3pm < 25
- **humidity_not_low**: if relative_humidity_3pm >= 25

Decision Tree Model

First, let's take a look at just the first two levels of the tree. You can see the following image by right-clicking on the **Decision Tree Learner** node in the workflow and selecting "View: Decision Tree View":

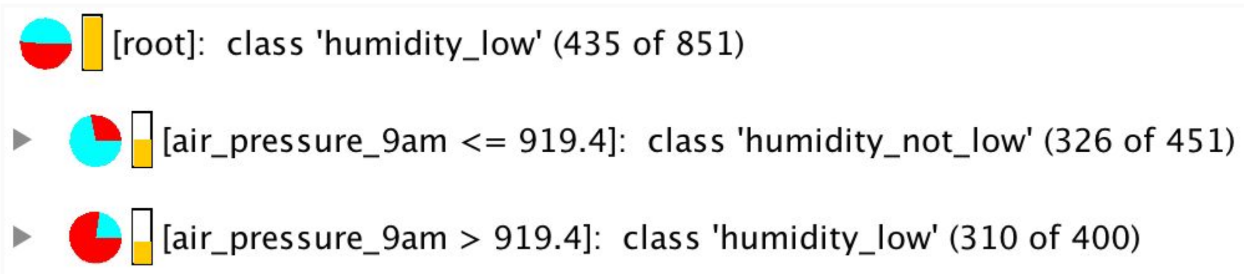


Looking at the root node (the top node), we see that there 851 samples in total. Of these, 435 or 51.1% of the samples are labeled as humidity_low; that is, the true label of these samples is humidity_low. Of the total number of samples, 416 or 48.9% are labeled as humidity_not_low. So at the root node, approximately half of the samples are humidity_low and half are humidity_not_low. This is indicated by the color bars at the bottom of the root node: blue is for humidity_not_low, and red is for humidity_low, and the height of each bar specifies the percentage of samples labeled with the respective category.

Split #1 on air_pressure_9am

The first split is on the variable **air_pressure_9am**. Samples with air_pressure_9am <= 919.4 are placed in the left child node where most of the samples are labeled as humidity_not_low. Samples with air_pressure > 919.4 are placed in the right child node where most of the samples are labeled as humidity_low. Note that the color red is associated with the humidity_low class. What this first split specifies is that high values of air_pressure are associated with humidity_low. This makes sense since high air pressure usually corresponds to sunny days, which have normal or high relative humidity.

To look at more levels in the decision tree, we need a more compact view. So we will now switch to the ‘simple’ view. The following image shows the same tree structure as the image of the decision tree above, and is generated by clicking on the Decision Tree Learner node and selecting “View: Decision Tree View (simple)”:

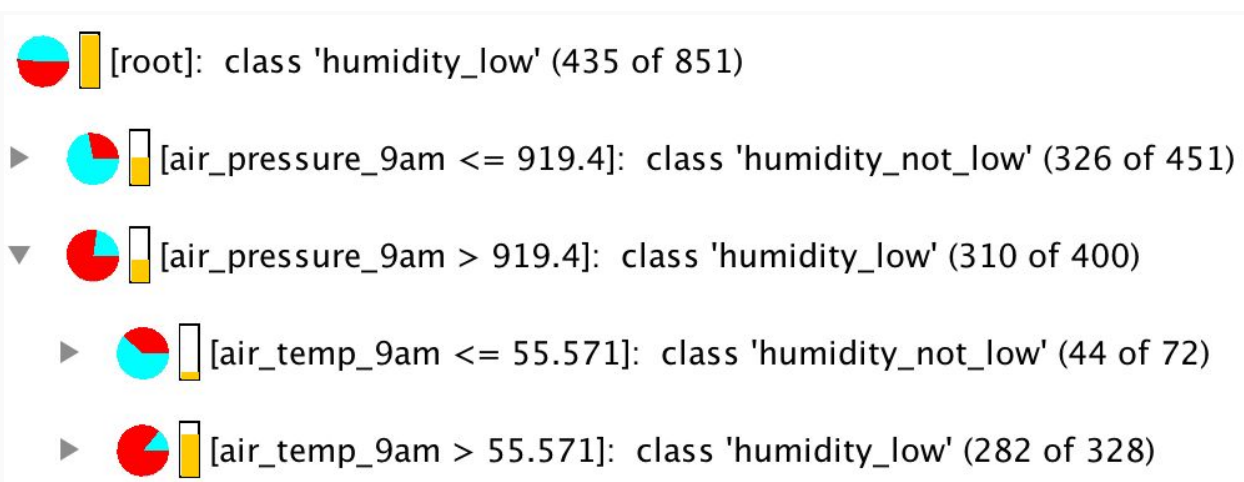


The root node is shown as the top line, followed by the children nodes resulting from the split on air_pressure_9am. Again, samples with air_pressure_9am <= 919.4 are placed in the left child node (shown right under the root node) where most of the samples are labeled as humidity_not_low. In other words, the true label for most samples in the left child node is humidity_not_low, which is indicated by the pie chart symbol being mostly blue, and the numbers in parentheses specifying that 326 out of 451 samples in that node are labeled humidity_not_low. Samples with air_pressure_9am > 919.4 are placed in the right child node where most of the samples are labeled as humidity_low.

Note that no prediction has been made yet since classification decisions are not made until a leaf node is reached.

Split #2 on air_temp_9am

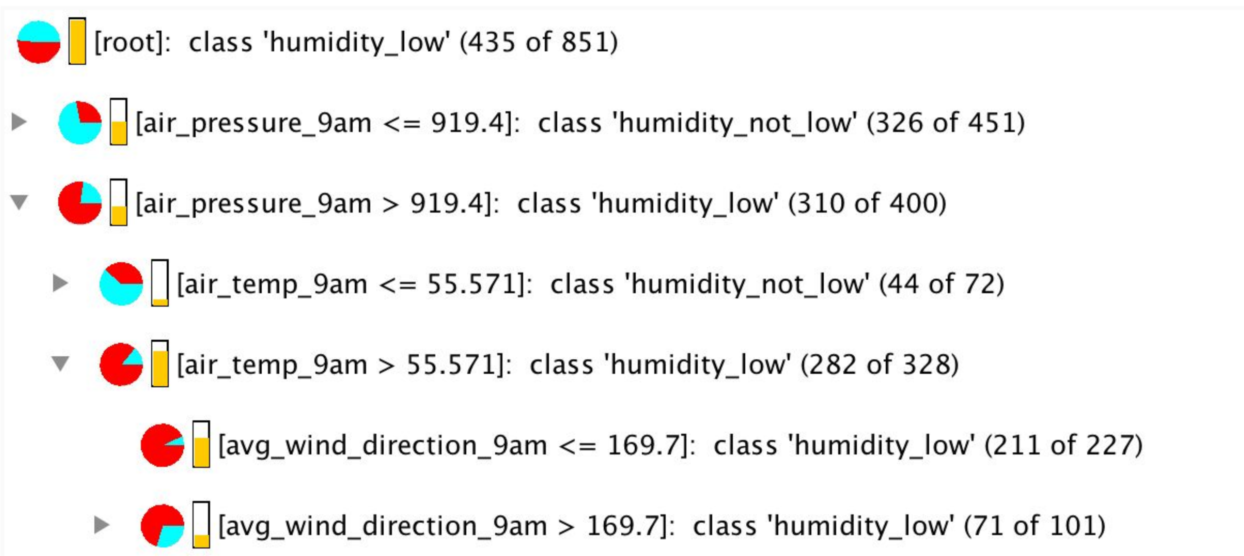
Let’s continue down the branch of the right child, where most of the samples have true label as humidity_low. If we expand that node, we get the following tree:



We see that this split is based on the variable **air_temp_9am**. If a sample has a value for air_temp_9am <= 55.571, then it placed in the left child node, where most of the samples are labeled as humidity_not_low. And if a sample has a value for air_temp_9am > 55.571, then it is placed in the right child node, where most of the samples are labeled as humidity_low. What this means is that low-humidity days are associated with warmer days. This makes sense since days with high humidity tend to be rainy days with cooler temperatures, while days with low humidity are sunny days with warmer temperatures.

Split #3 on avg_wind_direction_9am

Continuing with the humidity_low branch, we expand the right child node to get the following:



The third split is based on the variable **avg_wind_direction_9am**. Samples with avg_wind_direction_9am <= 169.7 are placed in the left child node where most of the samples are labeled as humidity_low. Samples with avg_wind_diection_9am > 169.7 are placed in the right child node. Notice that most of the samples in the right child node are also labeled humidity_low, but there is still additional processing needed with those samples since the right child node is not a leaf node.

Classification Rules

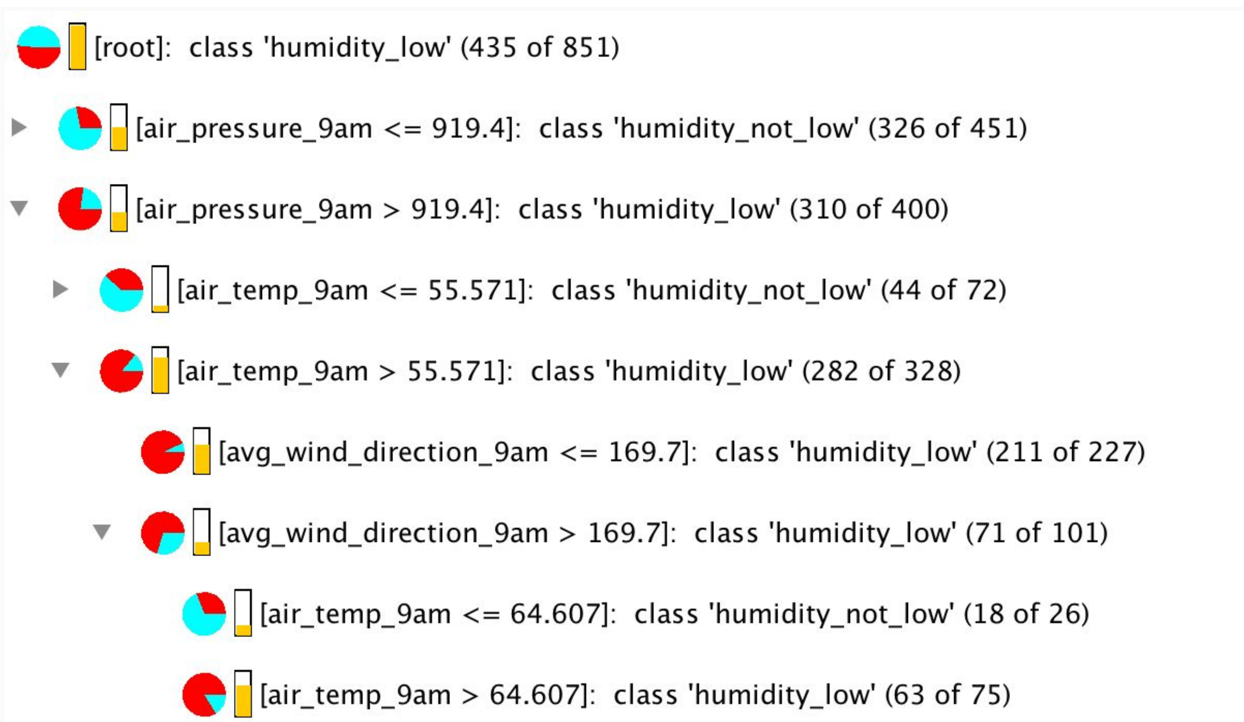
With the left child node, we have now reached a leaf node! Traversing from the root node to this leaf node, we can now see how a sample is classified as humidity_low:

1. If air_pressure_9am > 919.4 and
2. If air_temp_9am > 55.571 and
3. If avg_wind_direction_9am <= 169.7
4. Then sample is classified as humidity_low

This translates to days with high air pressure and warmer temperatures, with wind direction from the east are likely to be days with low relative humidity.

We have discussed above that low humidity is more likely to occur on sunny days with high air pressure and warmer temperatures. Now let's consider wind direction. Values for wind direction start at 0 degree for due North, and increases clockwise. So wind direction <= 169.7 means that the wind is from an eastern direction. For San Diego, this means warmer, drier air from the inland areas as opposed to cooler air with more moisture from the ocean. So this relationship between winds from the east and days with low humidity makes sense.

Expanding the right child node with avg_wind_direction_9am > 169.7, we get:



For the left leaf node, we see the following rules:

1. If air_pressure_9am > 919.4 and
2. If air_temp_9am > 55.571 and
3. If avg_wind_direction_9am > 169.7 and

4. If air_temp_9am <= 64.607
5. Then sample is classified as humidity_not_low

This translates to the following: Days with high air pressure, winds from the west, and temperatures between 56 and 65 degrees Fahrenheit are likely to be days with normal or high relative humidity.

For the right leaf node, we get:

1. If air_pressure_9am > 919.4 and
2. If air_temp_9am > 55.571 and
3. If avg_wind_direction_9am > 169.7 and
4. If air_temp_9am > 64.607
5. Then sample is classified as humidity_low

This translates to: Days with high air pressure, winds from the west, and temperatures greater than 65 degrees Fahrenheit are likely to be days with low humidity.

This branch is now complete. There are three leaf nodes, so there are three ways to assign a prediction of either humidity_low or humidity_not_low to each sample that is sent down this branch of the tree.

Other branches of the tree can be interpreted in a similar way. As with any other real dataset, some cases may require complex rules to form a classification decision.

Feature Importance

Aside from interpretability, another advantage of decision tree is that the resulting model tells you which features are important in the classification task. If you expand the tree to show all leaf nodes, you will see all the variables that the tree uses to perform the classification task.

For our daily weather dataset, note that out of the seven original input variables, only four variables (air_pressure_9am, air_temp_9am, avg_wind_direction_9am, max_wind_direction_9am) are used in the construction of the tree. These four variables are deemed important variables for this classification task, while the other variables do not contribute to the classification decisions made by the model.