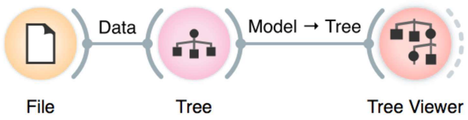
Lesson 6: Classification Trees

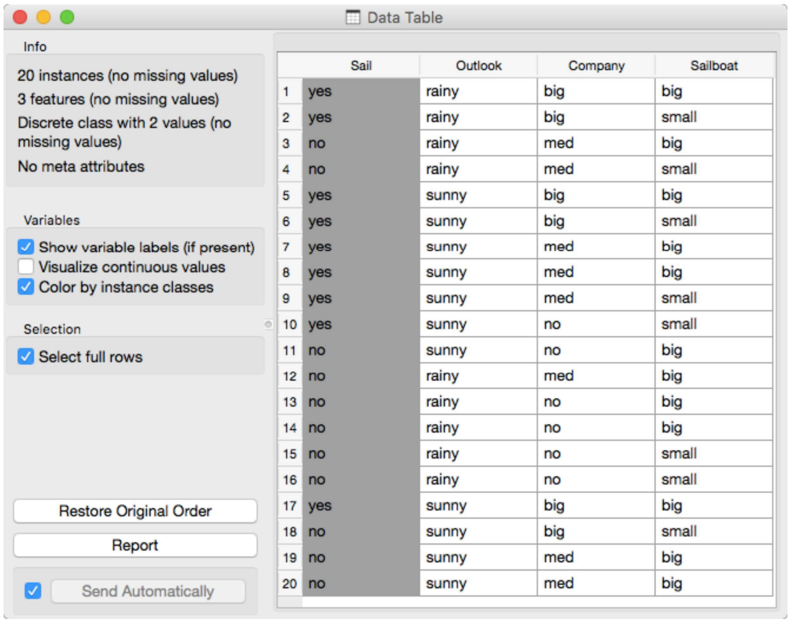
Classification tree is one of the oldest, but still popular, machine learning methods. We like it since the method is easy to explain and gives rise to random forests, one of the most accurate machine learning techniques (more on this later). So, what kind of model is a classification tree?

Let us load a data set from http://file.biolab.si/datasets/sailing.tab that records the conditions under which a friend skipper went sailing, build a tree and visualize it in the Tree Viewer.

The data set we will use is stored on a server. Copy the web address and paste it into URL entry box in the File widget. An alternative way to access this data is to use the Data Sets widget that is currently available in the Prototypes add-on.



Here's a warning: this sailing data is small. Therefore, any relations inferred from the classification tree on this page are unreliable. What should the size of the data set be to acquire stronger conclusions?



Info

20 instances (no missing values)
3 features (no missing values)
Discrete class with 2 values (no missing values)
No meta attributes

Variables

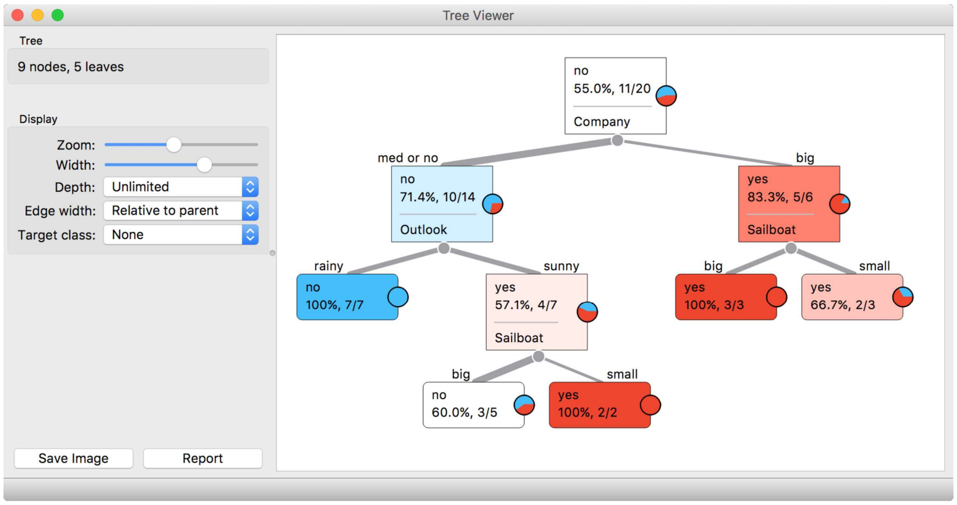
☑ Show variable labels (if present)
☐ Visualize continuous values
☑ Color by instance classes

Selection

☑ Select full rows

☑ Send Automatically

	Sail	Outlook	Company	Sailboat
1	yes	rainy	big	big
2	yes	rainy	big	small
3	no	rainy	med	big
4	no	rainy	med	small
5	yes	sunny	big	big
6	yes	sunny	big	small
7	yes	sunny	med	big
8	yes	sunny	med	big
9	yes	sunny	med	small
10	yes	sunny	no	small
11	no	sunny	no	big
12	no	rainy	med	big
13	no	rainy	no	big
14	no	rainy	no	big
15	no	rainy	no	small
16	no	rainy	no	small
17	yes	sunny	big	big
18	no	sunny	big	small
19	no	sunny	med	big
20	no	sunny	med	big



Tree

9 nodes, 5 leaves

Display

Zoom:
Width:
Depth: Unlimited
Edge width: Relative to parent
Target class: None

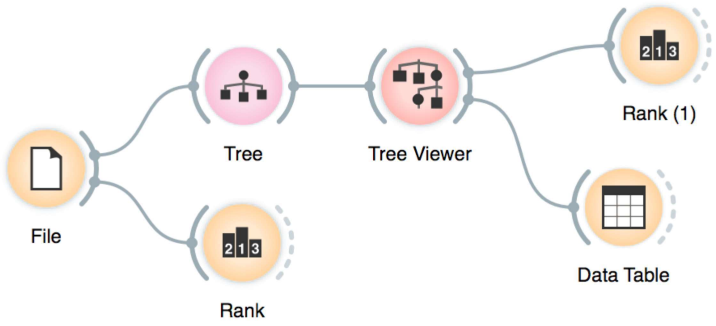
We read the tree from top to bottom. It looks like this skipper is a social person; as soon as there's company, the probability of her sailing increases. When joined by a smaller group of individuals, there is no sailing if there is rain. (Thunderstorms? Too dangerous?) When she has a smaller company, but the boat at her disposal is big, there is no sailing either.

Classification trees were hugely popular in the early years of machine learning, when they were first independently proposed by the engineer Ross Quinlan (C4.5) and a group of statisticians (CART), including the father of random forests Leo Brieman.

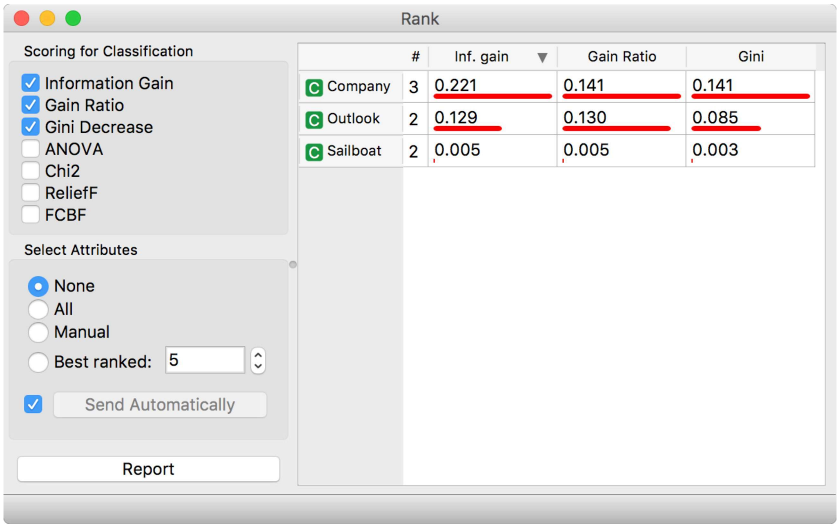
The Rank widget could be used on its own. Say, to figure out which genes are best predictors of the phenotype in some gene

Trees place the most useful feature at the root. What would be the most useful feature? It is the feature that splits the data into two purest possible subsets. These are then split further, again by the most informative features. This process of breaking up the data subsets to smaller ones repeats until we reach subsets where all data belongs to the same class. These subsets are represented by leaf nodes in strong blue or red. The process of data splitting can also terminate when it runs out of data instances or out of useful features (the two leaf nodes in white).

We still have not been very explicit about what we mean by "the most useful" feature. There are many ways to measure this. We can compute some such scores in Orange using the Rank widget, which estimates the quality of data features and ranks them according to how much information they carry. We can compute the scores from the whole data set or from data corresponding to some node of the classification tree in the Tree Viewer.



In this class, we will not dive into definitions. If you are interested, there's a good explanation of information gain on stackoverflow.com.


Rank

Scoring for Classification

- ☑ Information Gain
- ☑ Gain Ratio
- ☑ Gini Decrease
- □ ANOVA
- □ Chi2
- □ ReliefF
- □ FCBF

	#	Inf. gain ▼	Gain Ratio	Gini
C Company	3	0.221	0.141	0.141
C Outlook	2	0.129	0.130	0.085
C Sailboat	2	0.005	0.005	0.003

Select Attributes

- ● None
- ○ All
- ○ Manual
- ○ Best ranked: 5

☑ Send Automatically

Report