Control Parameters

The following parameters can be used to control or limit how a decision tree is built:

- The minimum number of observations that must exist in a node in order for a split to be attempted. This parameter helps us control the complexity of the decision tree (by keeping it smaller than it otherwise would be). For example, if "minimum split" parameter is set to 100, then if a node has fewer than 100 observations in it, the algorithm will create no further splits from that node.

- The minimum number of observations in any terminal node, usually called "minimum bucket size." This means that if a split results in a node that has fewer than this number of observations, then the split will not be made. Usually only one of this parameter and the minimum split parameter will need to be specified as they work in a very similar way to control decision tree complexity.

- A complexity parameter (CP), which indicates the minimum amount of impurity reduction required for a split to be made. Any split that does not decrease the overall lack of fit (this is referring to the measure of impurity, error, or goodness of fit) by a factor of CP is not attempted. For instance, with anova splitting (using R-squared impurity for regression), this means that the overall R-squared must increase by CP at each step. The main role of this parameter is to save computing time by pruning off splits that are obviously not worthwhile. Note, though, that it is possible that highly valuable splits may occur *after* an invaluable split, which is why it is often preferred to do retroactive pruning.

- The maximum depth of any node of the final tree, with the root node counted as depth 0. Again, this parameter controls the complexity of the tree. A similar notion is setting the maximum number of nodes.