# Tree-Based Models and Variable Importance

## Variable Importance in Decision Trees and Random Forests

Run CHUNK 16 to see how to build a decision tree in R.

You don't need to worry about the detail of this algorithm; it will be covered in Module 6. For the purposes of feature selection, we care more about the fact that the algorithm actually "chooses" the most useful variables at each split in the tree. Thus, by observing the final constructed decision tree (or random forest etc.) we can calculate a measure of "variable importance" based on the number of times a variable is selected for use in the model and the resulting improvement it results in, weighted by the number of data points in the subset of data at that point in the tree. This variable importance can be used to rank variables and the top n selected for use in further modeling in a similar way to filter-based feature selection however this method allows us to acknowledge non-linear variable effects.

You can see an example of what the variable importance output from a tree based model looks like by running CHUNK 17. Further details of this method will be covered in Module 6.