**Regression tree -> need to revise**

**Classification tree -> need to revise**

**Random forest**

* Step 1: Create a bootstrapped dataset BD
* Step 2: Create a decision tree using the BD, but use only random subset of variables (columns) of each step
* Go back to step 1 and repeat

**How to use it**

* We have a new record -> run with first tree, second tree, … -> all the tree
* Receive the option that most trees make
* Bootstrapping the data plus using the aggregate to make a decision is called Bagging

**When using bootstrap to create the data set -> typically 1/3 of the original data does not end up in the BD**

-> this is called the “Out-Of-Bag Dataset”

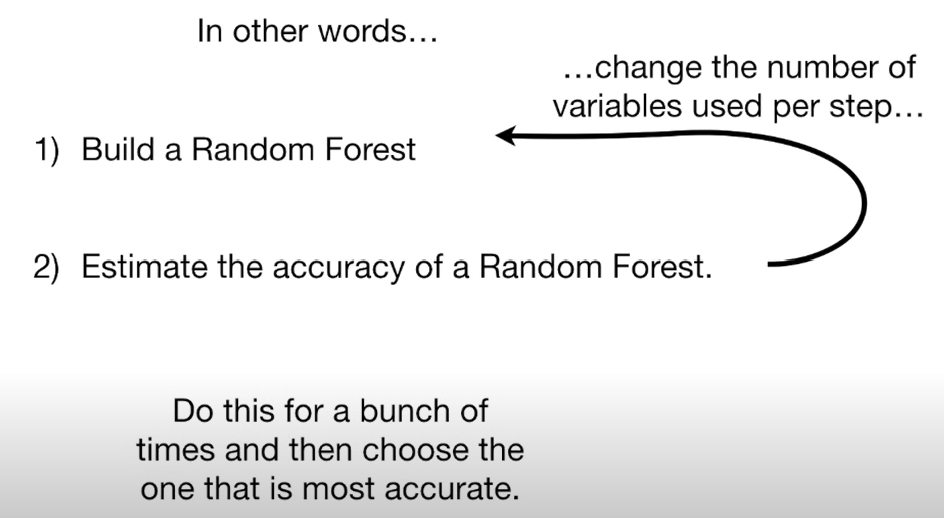
-> we run Out-Of-Bag sample through all of the other trees

-> the label with the most votes wins, it is the label that we assign this Out-Of-Bag sample

-> do the same thing for all of other Out-Of-Bag samples

-> we can measure how accurate our random forest is by the proportions of OOB samples that were correctly classified

-> the proportion of OOB samples that were incorrectly classified is the OOB Error



**Missing data and clustering**

Random forests consider 2 types of missing data

* Missing data in the original dataset used to create the random forest
  + First guess the missing value based on other samples with same label
    - Quantitative -> median
    - Qualitative -> most common
  + Build a random forest, run the data through the trees, recalculate the proximities and recalculate the missing values (do this 6-7 times until the missing values converge)
  + 1 – proximity matrix = distance matrix -> we can draw a heatmap, MDS plot
* Missing data in a new sample that you want to categorize
  + Create 2 copies of the data, ex one has heart disease, and one doesn’t have heart disease
  + Using above iterative method to make a good guess about the missing values