**Random Forest Algorithm Investigation**

**Random** - Two kinds of randomness  
During the construction process, two mechanisms of random selection are employed.

**Forest** - A forest  
It consists of multiple decision trees, which is why it is called a "forest."

**What are the advantages of Random Forest over Decision Trees?**Higher accuracy.

**How to build a Random Forest**

1. Randomly (1) select data from the original dataset (allowing repetition of the same data) to form a Bootstrap dataset. (Note: Approximately one-third of the original data will not enter the Bootstrap dataset. This data is referred to as Out-of-bag samples.)
2. Use the Bootstrap dataset to build decision trees, but in each step of the construction, only a randomly (2) selected subset of features can be used for building. For example, when building the first layer, features like Feature1 and Feature4 might be randomly selected, and for the second layer, Feature2 and Feature4 might be chosen.
3. Repeat step 2 to build multiple decision trees.

**How to use a Random Forest**Feed each row of the original dataset into every decision tree. Each decision tree will produce a classification result (different decision trees may produce different results). Finally, collect the classification results from all the decision trees, and the result with the highest number of votes becomes the final answer.

**How to verify the accuracy of a Random Forest**Pass the Out-of-bag samples mentioned earlier into each decision tree, and ultimately aggregate and compare the results to assess the accuracy.

* **Random** - 两个随机  
  构建过程中两次采用随机选取数据的机制。
* **Forest** - 一片森林  
  由多个Decision Tree组成，因而叫做森林（Forest）
* Random Forest相比Decision Tree有什么优势？  
  准确性更高。
* **如何构建一个Random Forest**
  1. 随机(1)从原始数据集中获取数据（可重复取用同一条数据）用来组建Bootstrap数据集。（注：大约有3分之1的原始数据不会进入Bootstrap数据集。这类数据被称作Out-of-bag samples）
  2. 用Bootstrap数据集构建Decision Tree，但是构建的每一步只能随机(2)选取部分feature来进行构建。例如：构建第一层时可能随机选取了Feature1和Feature4，构建第二层时可能会随机选取Feature2和Feature4。
  3. 重复步骤2，构建多个Decision Tree.
* **如何使用Random Forest**

把原始数据集的数据逐行分别放入每一个Decision Tree，每个Decision Tree都会得到一个分类结果（不同的Decision Tree所产生的结果可能不同），最终汇集所有Decision Tree的分类结果，票数高的那个结果为最终答案。

* **如何验证Random Forest**的准确性  
  将前面提到的Out-of-bag samples分别传入每个Decision Tree，最终汇总比对结果的准确性。