# Introduction

Step 1 is to create a bootstrap dataset that is the same size as the original (randomly sample from the dataset, repeats are allowed)

Step 2 create a decision tree from the bootstrap dataset in step 1. Critically we only use a random subset of all the features at each step to build the next part of the tree.

In prediction we use the ensemble of trees and they each vote for a classification. This method is called bagging. Bagging is bootraping data and using aggregate to decide is called bagging.

Out of bag dataset is the data that wasn’t used to create each tree, each different tree has a different out of bag. This gives us training metric of the different tree’s performance. A test set that was reserved and was not in any trees can be used to see the test metrics of the model.

To know how many random variables to consider at each step we test different parameters, then we measure the accuracy on the out of bag data (train data not selected in this boostraping step) whatever parameters provide the best results we use. Commonly we use the square of the number of variables.

Table

Description automatically generated

A picture containing timeline

Description automatically generated

# Missing values and clustering

We can make an initial guess of missing data using the median value. We can then refine the guesses by looking for samples similar. We can do this by building a random forest. We then go through every tree in the random forest and build a similarity matrix. Two entries are similar if they finish in the same leaf node in a tree in the random forest. We then divide the proximity values by the total number of trees. We can use the proximity values to weight the frequency of occurrences. This then becomes the revised guess.

For a regression based column we would calculate the weighted average of all the values. This is basically the sum of row.value \* weight[row]

We repeat these steps 6 or 7 times or until the values we assign stop changing. This means we have found good values to use here.

We could then do a heatmap or an MDS plot to show sample correlation. This would be a useful technique for determining data correlation.

1 – proximity matrix is the distance matrix.

Table

Description automatically generated

Graphical user interface, application

Description automatically generated

Application

Description automatically generated with low confidence