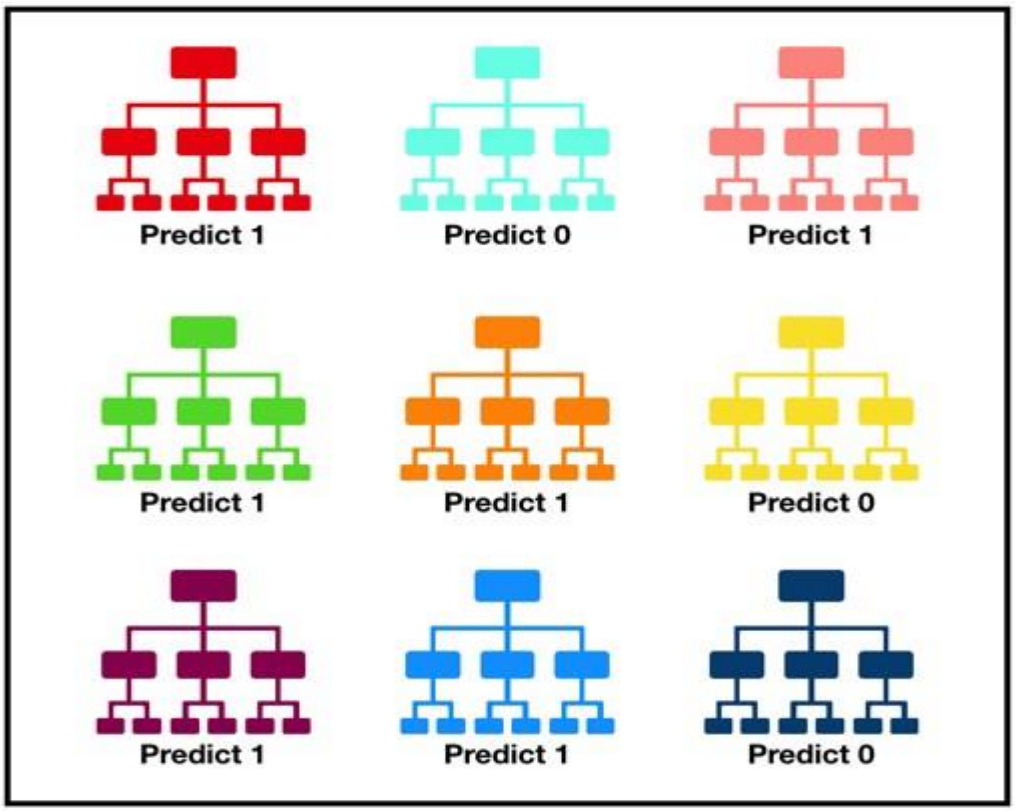
Lecture 6

Classification Trees and Random Forest

The random forest consists of many individual decision trees that operate as an ensemble. For classification tasks, the output of the random forest is the class selected by most trees.



Source: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Each tree is grown as follows:

STEP 1. Random Record Selection

Each tree is trained on 63.2% of the total training data. Cases are drawn at random with replacement from the original data.

STEP 2. Random Variable Selection

m predictor variables are selected at random out of all the predictor variables, and the best split on this m is used to split the node. By default, m is a square root of the total number of all predictors for classification.

STEP 3. Out Of Bag (OOB) Error Rate

Using the leftover (36.8%) data, calculate the misclassification rate for each tree.

STEP 4. Ensemble

Aggregate error from all trees to determine overall OOB error rate for the classification. The forest chooses the classification having the most votes over all the trees in the forest.

Random forests are frequently used as "blackbox" models. Random forests can be used to rank the **importance of variables** in a classification problem. Higher the value of mean decrease accuracy or mean decrease gini score, higher the importance of the variable in the model:

- Mean Decrease Accuracy – How much the model accuracy decreases if we drop that variable.
- Mean Decrease Gini – Measure of variable importance based on the Gini impurity index used to calculate splits in trees.

Task 1. Titanic Data

Predict the chance of surviving the Titanic sinking using:

- rpart
- tree
- Random Forest

Check the results depending on the size of the training set (70%, 80%, 90%).

Task 2. Wine

In the file https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv, there are data about red wine characteristics. Each wine in this dataset is given a quality score between 0 and 10.

Predict wine quality based on:

- rpart
- Random Forest

As an explained variable, choose:

a) quality score
b) classification:
    - the wines ranked as 5 and 6 as "normal",
    - the lower ranked wines as "bad",
    - the wines rated above as "good".