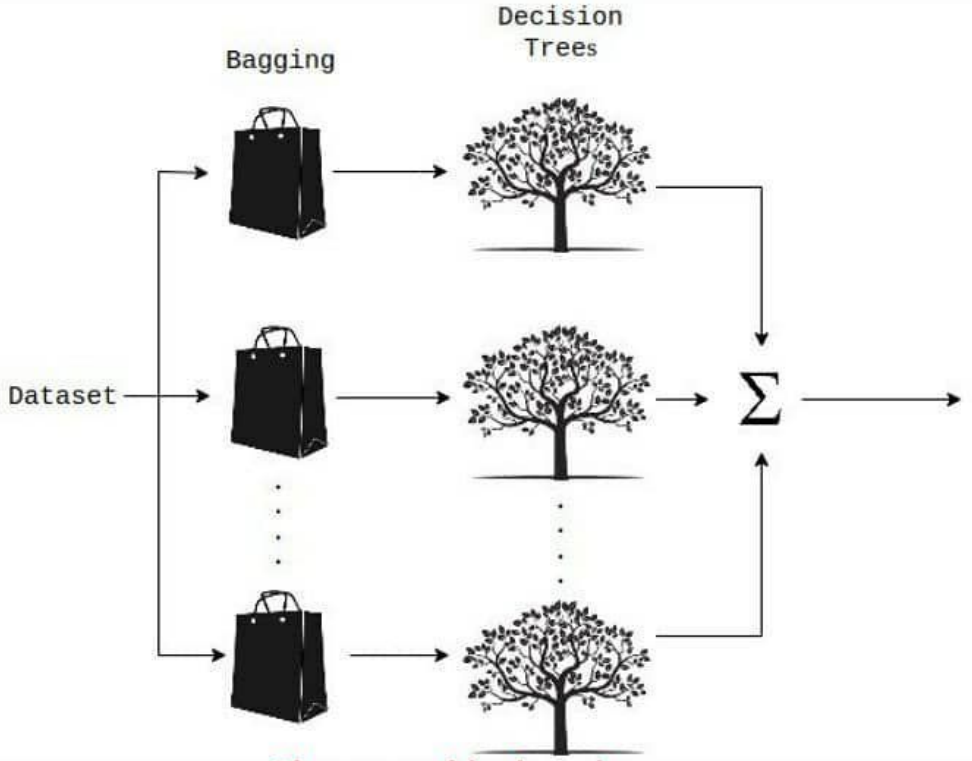
Random Forest

- Random forest algorithm can use both for classification and the regression kind of problems.
- Random forest algorithm is a supervised classification algorithm.
- Random forest is a type of ensemble machine learning algorithm called Bootstrap Aggregation or bagging. We already discussed Bootsratping in our statistic series.
- Bagging is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model.



Random Forest

- Bootstrap Aggregation is a general procedure that can be used to reduce the variance(overfit) for that algorithm that has high variance(like Decision trees).
- The disadvantage of Decision trees is they are sensitive to the specific data on which they are trained. If the training data is changed the resulting decision tree can be quite different and in turn, the predictions can be quite different.
- When bagging with decision trees, we are less concerned about individual trees overfitting the training data. For this reason and for efficiency, the individual decision trees are grown deep and the trees are not pruned. These trees will have both high variance and low bias. These are important characterize of sub-models when combining predictions using bagging.
- The only parameters when bagging decision trees is the number of samples and hence the number of trees to include. This can be chosen by increasing the number of trees on run after run until the accuracy begins to stop showing improvement. Very large numbers of models may take a long time to prepare, but will not overfit the training data.

Random Forest

- The 2 key concepts in Random forest which play major role in increasing the accuracy and performance of the model.
- A random sampling of training data points when building trees, Random subsets of features considered when splitting nodes
- **A random sampling of training observations:** When training, each tree in a random forest learns from a random sample of the data points(which includes random rows and random columns). The idea is that by training each tree on different samples, although each tree might have high variance with respect to a particular set of the training data, overall, the entire forest will have lower variance but not at the cost of increasing the bias.
- **Random Subsets of features for splitting nodes:** The other main concept in the random forest is that only a subset of all the features is considered for splitting each node in each decision tree. Generally, this is set to sqrt(n_features) for classification meaning that if there are 16 features, at each node in each tree, only 4 random features will be considered for splitting the node.

Random Forest

- The random forest combines hundreds or thousands of decision trees, trains each one on a slightly different set of the observations, splitting nodes in each tree considering a limited number of the features. The final predictions of the random forest are made by averaging the predictions of each individual tree.
- For each bootstrap sample taken from the training data, there will be samples left behind that were not included. These samples are called Out-Of-Bag samples or OOB.
- The performance of each model on its left out samples when averaged can provide an estimated accuracy of the bagged models. This estimated performance is often called the OOB estimate of performance.
- These performance measures are reliable test error estimate and correlate well with cross-validation estimates.

Random Forest

- **Advantages of Random Forests**
- **It can be used for both classification and regression problems**
- **Reduction in overfitting by averaging several trees and random row sampling and column sampling, there is a significantly lower risk of overfitting.**
- **Random forests make a wrong prediction only when more than half of the base classifiers are wrong**
- **It is very easy to measure the relative importance of each feature on the prediction.**
- **The random forest algorithm can be used for feature engineering. Which means identifying the most important features out of the available features from the training dataset.**

Random Forest

- **DisAdvantages of Random Forests**
- **Random forests have been observed to overfit for some datasets with noisy classification/regression tasks.**
- **It's more complex and computationally expensive than the decision tree algorithm.**