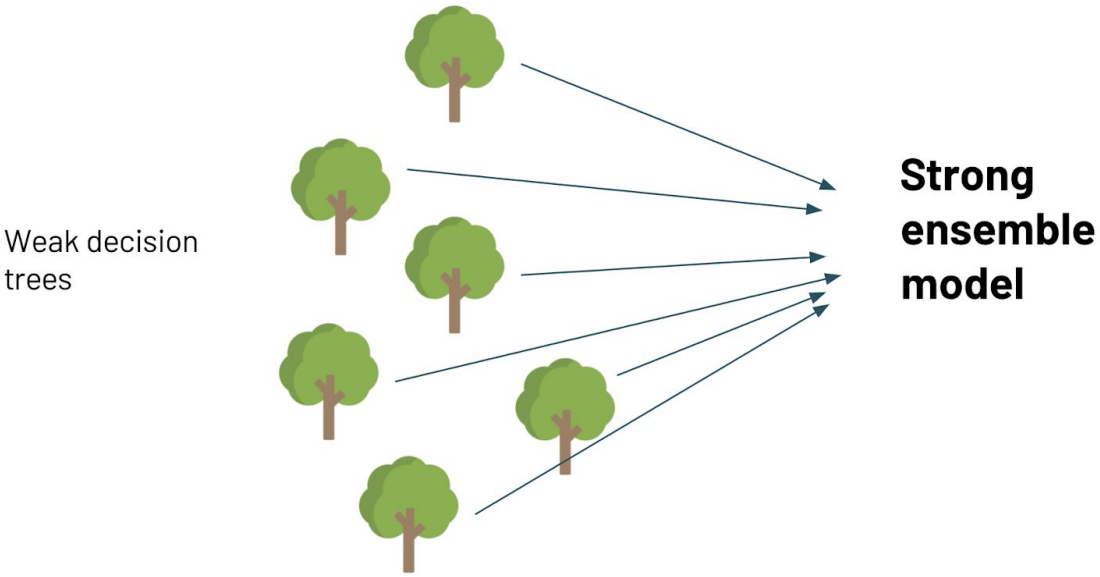
Random Forest Algorithm

Estimated time: 7 minutes

Random forest is an ensemble modeling algorithm that combines multiple weak decision trees into a strong learner.

The individual trees in any ensemble algorithm can be either shallow (not many depths of decisions) or deep (lots of depths). Shallow trees have less variance but higher bias, so they are generally a better choice for boosting or sequential methods that decrease bias. Conversely, Deep trees have low bias but high variance and are good for bagging or parallel methods that result in lower variance.



Random forest combines deep trees, fitted on bootstrap samples, to reduce variance. In addition to using bootstrapped samples for each individual decision tree, random forest employs a technique that makes each tree less correlated with the others and introduces randomness. When fitting each

tree, a sub sample is taken from the observations, or rows, in the dataset, with replacement so that each subsample can have repetitions.

Date	Humidity	Cloud coverage	Rain
2020-08-27	65%	82%	Yes
2020-08-28	62%	55%	Yes
2020-08-29	30%	12%	No
2020-08-30	25%	16%	No
2020-08-31	43%	24%	No
2020-09-01	51%	41%	Yes

In addition to the randomness introduced by sampling from the observations in the training dataset, within each decision tree, at every split or node the tree only considers a randomly selected subset of the features.

Date	Humidity	Cloud coverage	Rain
2020-08-27	65%	82%	Yes
2020-08-28	62%	55%	Yes
2020-08-29	30%	12%	No
2020-08-30	25%	16%	No
2020-08-31	43%	24%	No
2020-09-01	51%	41%	Yes

These two techniques are where the randomness in random forest comes from, and ensures that each of the decision trees in the forest are different, thus reducing variance.

There are some hyperparameters that are specific to random forest, such as the number of trees (or estimators) to build, and the maximum number of features to select from. In addition, the individual tree hyperparameters can be set when instantiating a random forest.