

# Random Forests

🕒 Created	@April 28, 2022 3:02 PM
☰ machine learning	<span>Daily</span> <span>Work</span>

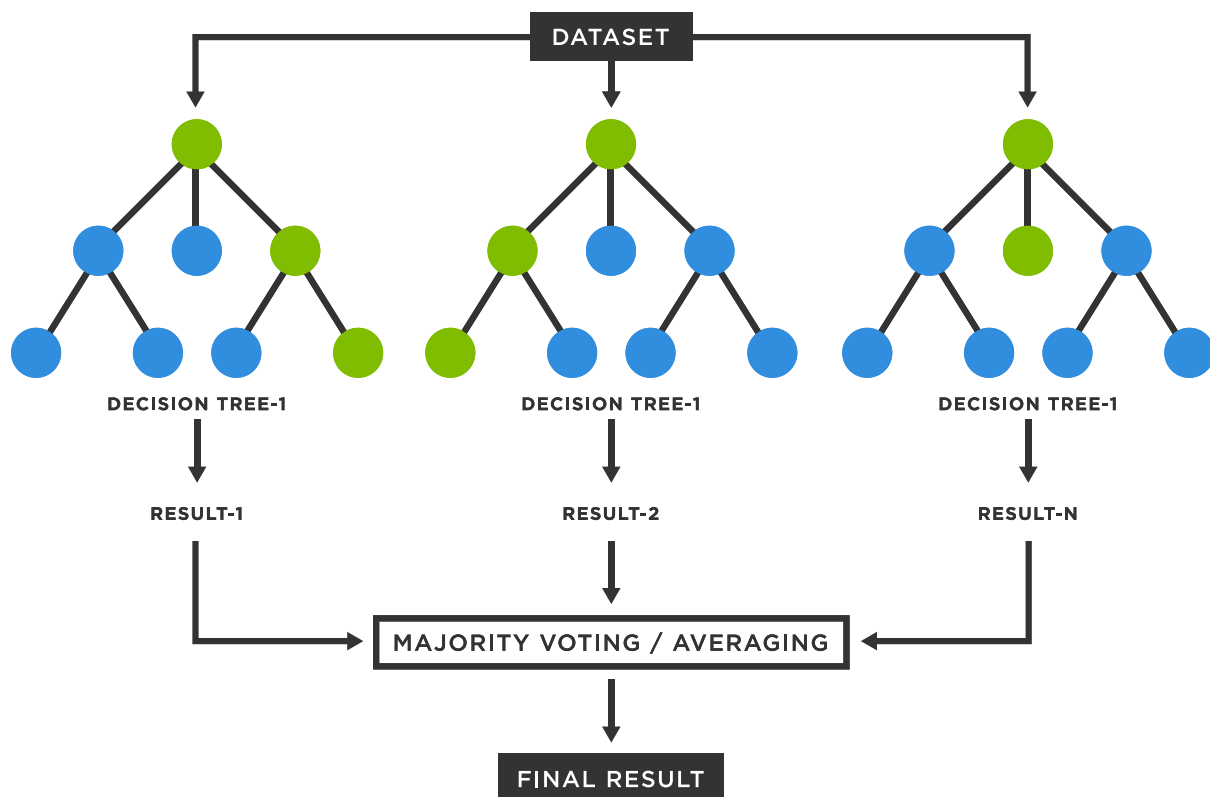
Random Forest is a supervised machine learning algorithm that consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest gives a prediction and the model prediction is chosen based on this information.

- In the case of classification problems, each individual tree's result is considered and the result with the majority of the votes becomes the model's prediction.
- When we are doing regression problems, the model chooses the mean or mode of prediction of individual trees.

Random forests consist of many individual trees that are formed simultaneously and independently from each other. The reason why this is possible is that Random forests use bootstrapped data for each tree.

**Bootstrapping** is picking random samples from the original dataset with replacement. Bootstrapped data may have some duplicated data and may not include all the samples from the original dataset. It usually includes 60% of the original data. The size of bootstrapped data may be different from the dataset, but all bootstrapped sets need to have the same length.

Bootstrapping the data and using the aggregate to make a decision tree is called **bagging**. Random forests allow each individual tree to randomly sample from the dataset with replacement and results in different trees.



Random forests is a very good and efficient model for working with large datasets. It is not prone to overfitting, mainly because a large number of relatively uncorrelated trees operating as a committee will work on making sure that predictions are accurate. The trees protect each other from their individual errors.

---

Strengths of Random forests:

- An all-purpose model that performs well - both classification and regression;
- It can handle noisy and missing data;
- It can select only the most important features;
- It model can be used efficiently with big datasets.


Weaknesses:

- It can be hard to interpret;
- Requires the model to be tuned for the data with Cross-validation, GridSearchCV, etc.

Sources:

#### `sklearn.ensemble.RandomForestClassifier`


A random forest classifier. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy

 <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>



#### `sklearn.ensemble.RandomForestRegressor`

A random forest regressor. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive

 <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>



### 1.11. Ensemble methods

In averaging methods, the driving principle is to build several estimators independently and then to average their predictions. On average, the combined estimator is usually better than any of the

 <https://scikit-learn.org/stable/modules/ensemble.html>

