RANDOM FOREST CLASSIFIER

What is Random Forest?

- Supervised learning algorithm

- Forest - Ensemble of decision trees, usually trained with the “bagging” method.

- **Builds multiple decision trees and merges them together to get a more accurate and stable prediction.**

Ensemble learning – What, Why & How?

- *What?*
  - *Ensemble models in machine learning combine the decisions from multiple models to improve the overall performance.*
- *Why?*
  - *Ensemble methods help to minimize errors in learning models due to* **noise, bias and variance**.
- *How?*
  - *Taking the mode of the results – majority voting*
  - *Taking weighted average of the results*

Bagging?

- Bootstrap AGGregatING
  - Create random samples of the training data set with replacement (sub sets of training data set).
  - Build a model (classifier or Decision tree) for each sample.
  - Combine the results of these multiple models using average or majority voting.

Random Forest Classifier

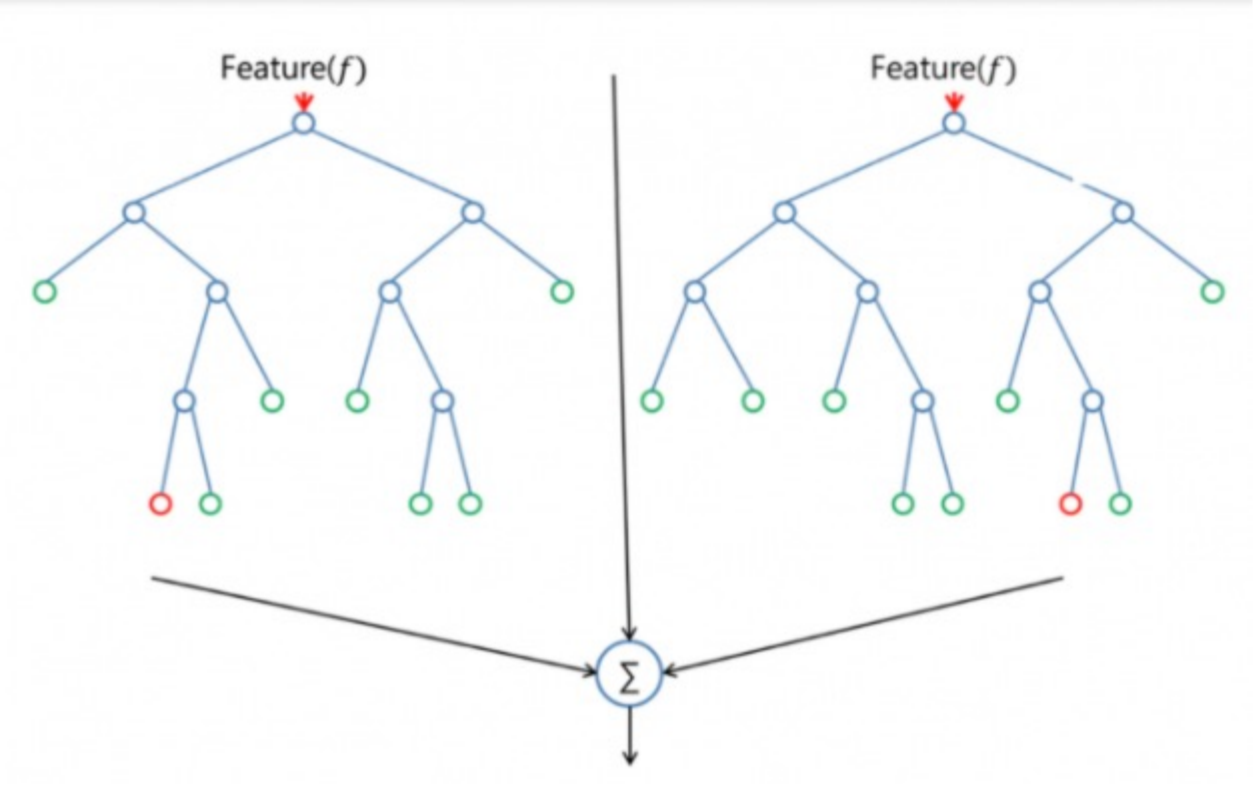
- Random forest adds additional randomness to the model, while growing the trees.

- Only a random subset of the features is taken into consideration by the algorithm for splitting a node.

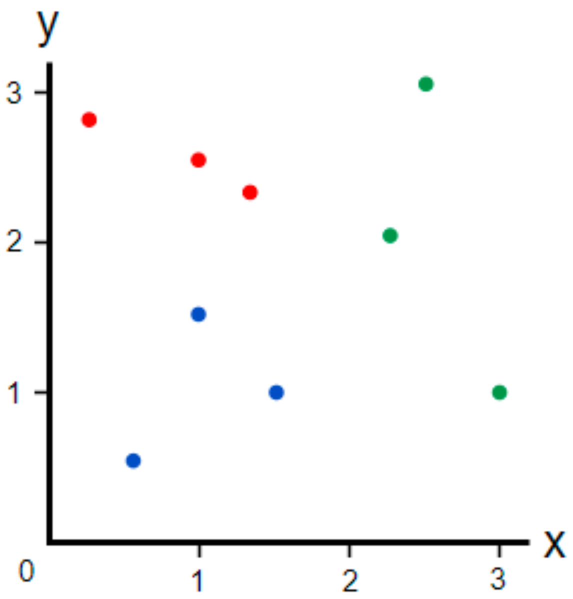
- Randomly selects observations and features to build several decision trees and then averages the results.

- This results in a wide diversity that generally results in a better model.

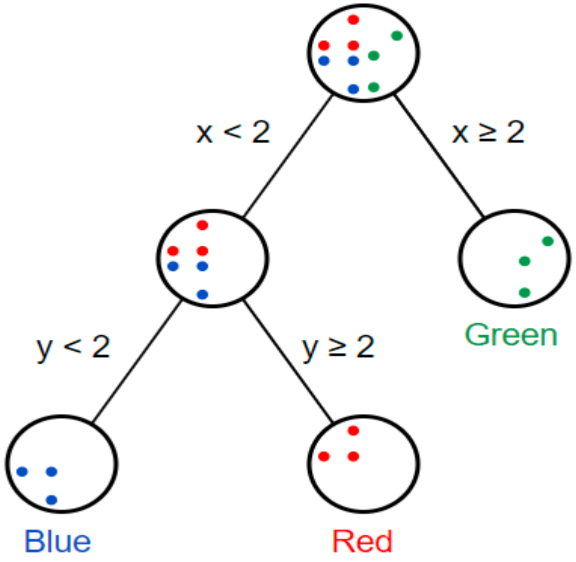
Random Forest Classifier



Decision Tree Vs. Random Forest

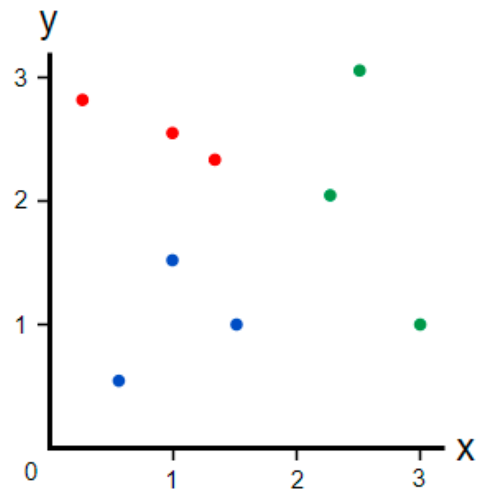


The Dataset v2

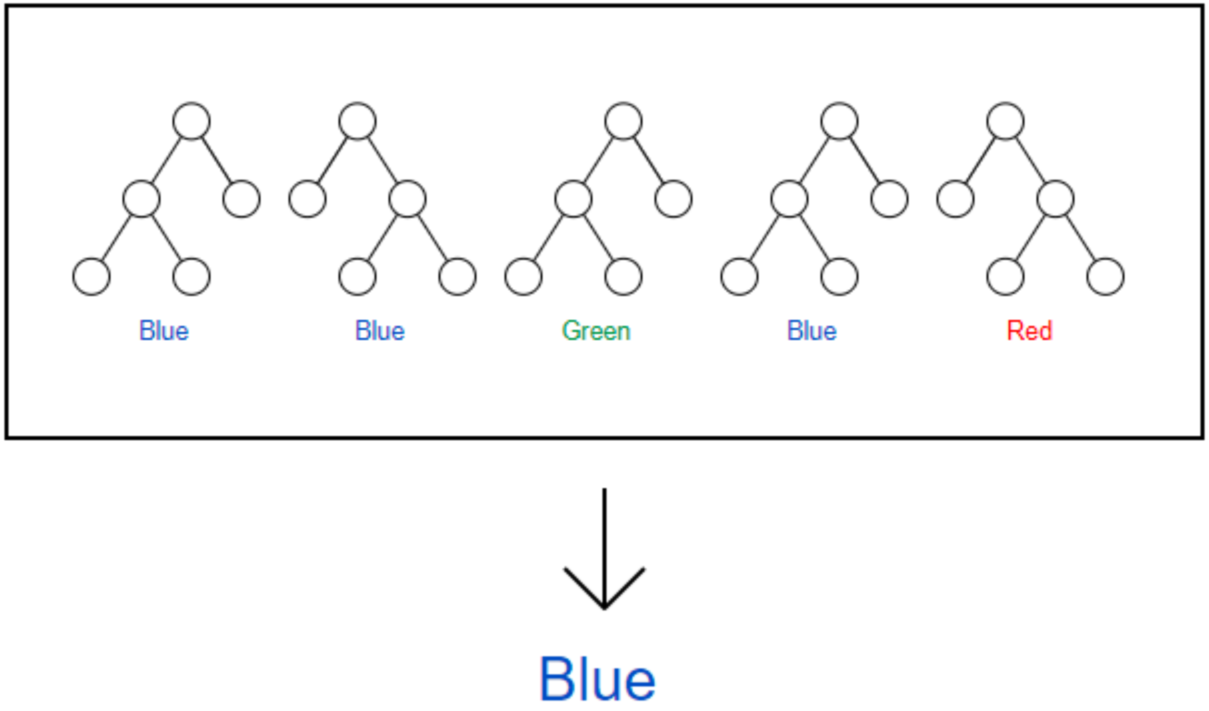


Decision Tree

Decision Tree Vs. Random Forest



The Dataset v2



Bagged Decision Trees predicting color

Bagging → Random Forest

- Has 2 parameters
  - A parameter to specify the number of trees
  - A parameter that controls **how many features to try when finding the best split**.

Pros & Cons

- **Pros**
- Versatility – used for both regression and classification models
- The default hyperparameters it uses often produce a good prediction result
- Because of enough trees in the forest, the classifier won't overfit the model.
- **Cons**
- Large number of trees can make the algorithm too slow and ineffective for real-time predictions.