Machine Learning Algorithms
Bagging – Random Forest

Sundharakumar KB

Department of Computer Science and Engineering
School of Engineering

Shiv Nadar University Chennai

SHIV NADAR
UNIVERSITY
CHENNAI

Ensemble Methods

- Ensemble methods are meta-algorithms which combine several machine learning models together and creating a single predictive model out of it.
- The main goal of using ensemble methods are : to reduce variance (bagging), to reduce bias (boosting) and to improve the predictions.
- Stacking is also another method under ensemble methods but most widely used techniques are Bagging and Boosting.
- Bagging and boosting algorithms can be used in both regression problems and classification problems.
- Random Forest is one of the prominent bagging algorithms & Adaboost and Gradient Boosting are popular boosting methods.

Bagging – Random Forest

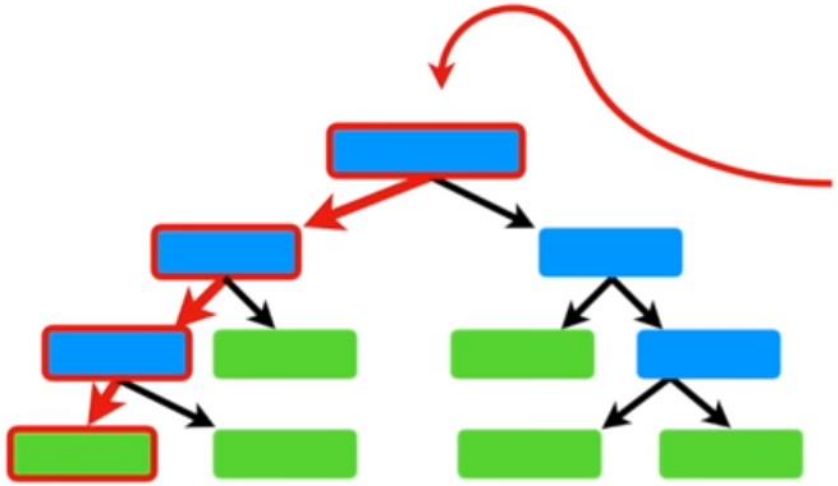
- Bagging is the short for Bootstrap Aggregation.
- Random Forest – is an extension to the decision tree algorithm.
- Constructing multiple decision trees creates the forest.
- Step 1: Create various bootstrapped datasets from the existing training data.
- Step 2: Create decision tree with the randomly selected bootstrapped data.
- Step 3: Repeat step 1 for “n” number of times and create “n” decision trees
- Step 4: Calculate the performance metrics for the dataset.

Random Forest – Intuition

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	No	168	

…and now we want to know if they have heart disease or not.

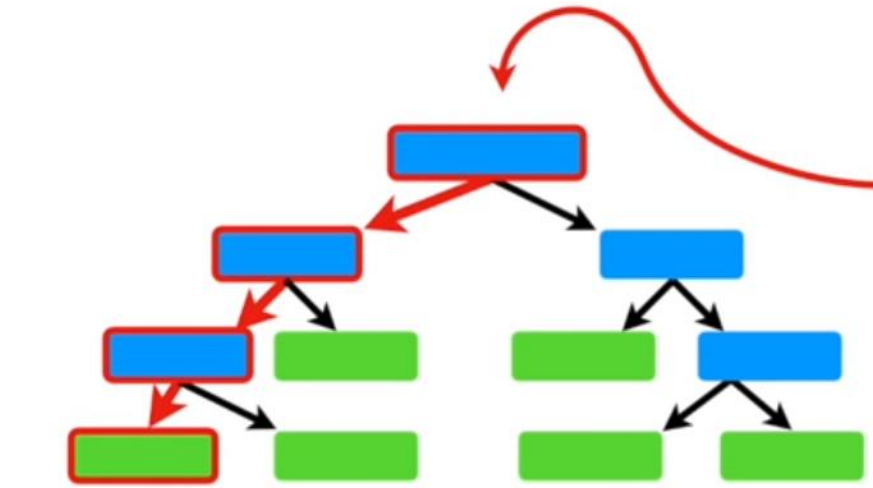
Random Forest – Intuition



Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	No	168	

The first tree says "Yes"…

Random Forest – Intuition



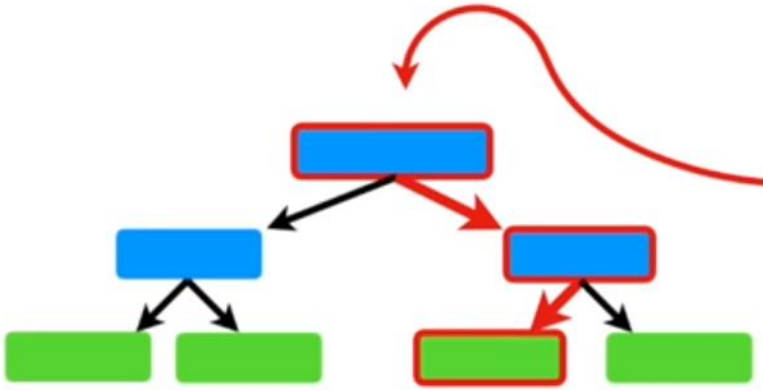
Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	No	168	

Heart Disease	
Yes	No
1	0

The first tree says "Yes"…

…and we keep track of that here.

Random Forest – Intuition

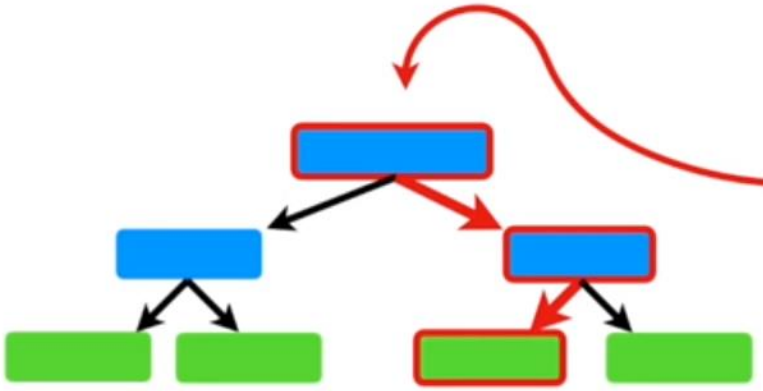


The second tree also says "Yes"…

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	No	168	

Heart Disease	
Yes	No
1	0

Random Forest – Intuition



Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	No	168	

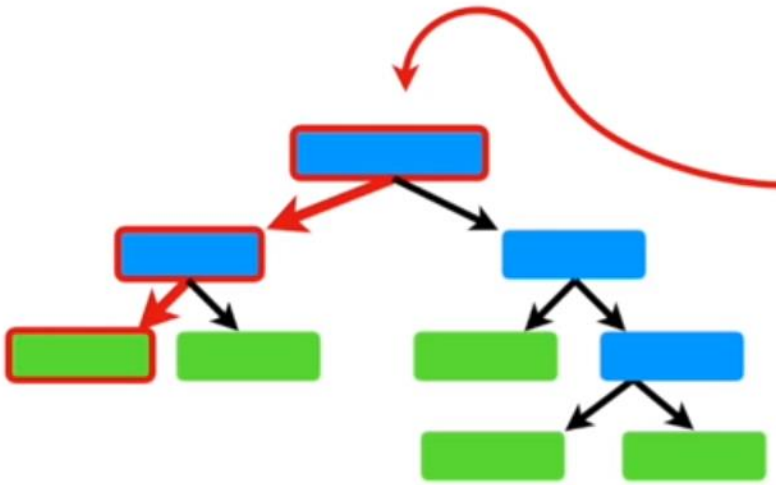
The second tree also says "Yes"…

Heart Disease	
Yes	No
2	0

…and we keep track of that here.

Random Forest – Intuition

Then we repeat for all the trees that we made…



Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	No	168	

Heart Disease	
Yes	No
5	1

Random Forest – Intuition

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	No	168	

After running the data down all of the trees in the random forest, we see which option received more votes.

Heart Disease	
Yes	No
5	1

SHIV NADAR UNIVERSITY CHENNAI

Random Forest – Summary

- More the number of trees, more will be the accuracy. Having exceedingly higher number of trees can lead to overfitting.
- For every stage of boostrapped data, we will have few data left behind. They are called as Out-of-bag samples (OOB).
- The trees generated will be functioning in parallel.