MACHINE LEARNING

Ensemble Methods – Bagging and Boosting

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 algorithm & XGBoosting (Extreme Gradient Boosting) is one of the most widely used boosting algorithm. However Adaboost and Gradient Boosting are also used.

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.

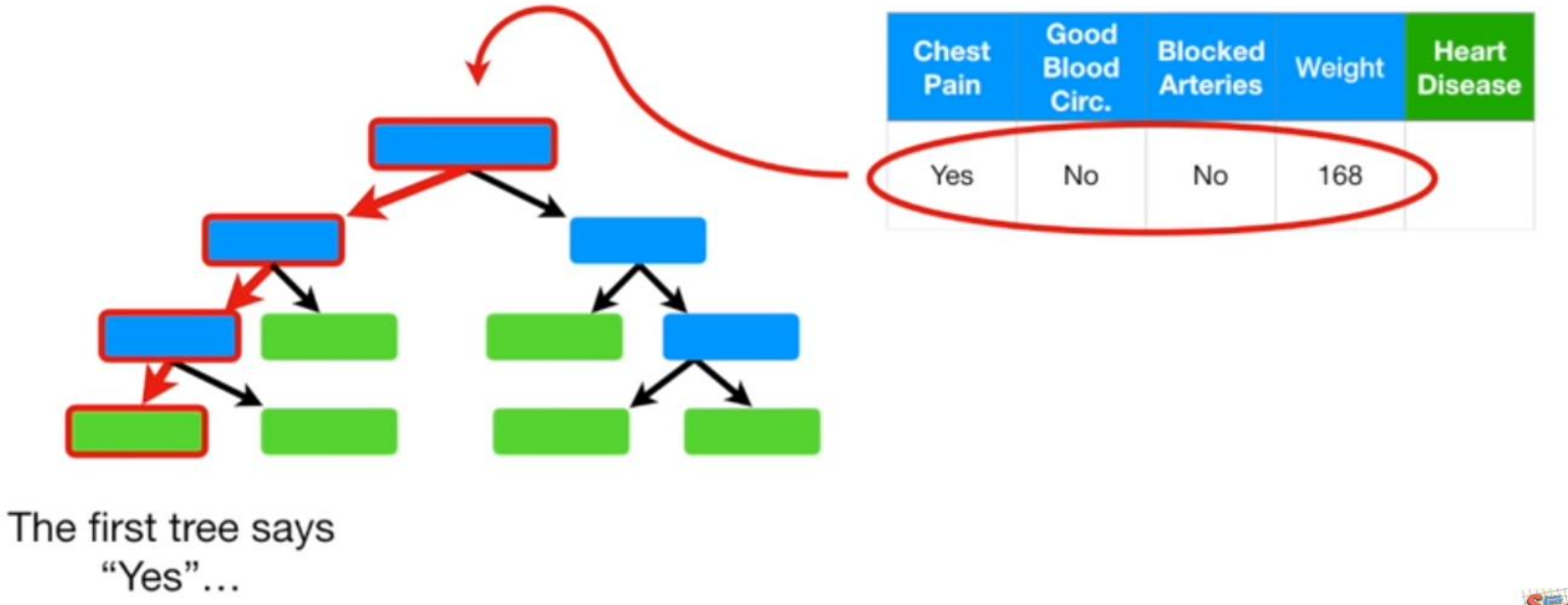
RANDOM FOREST - INTUITION

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

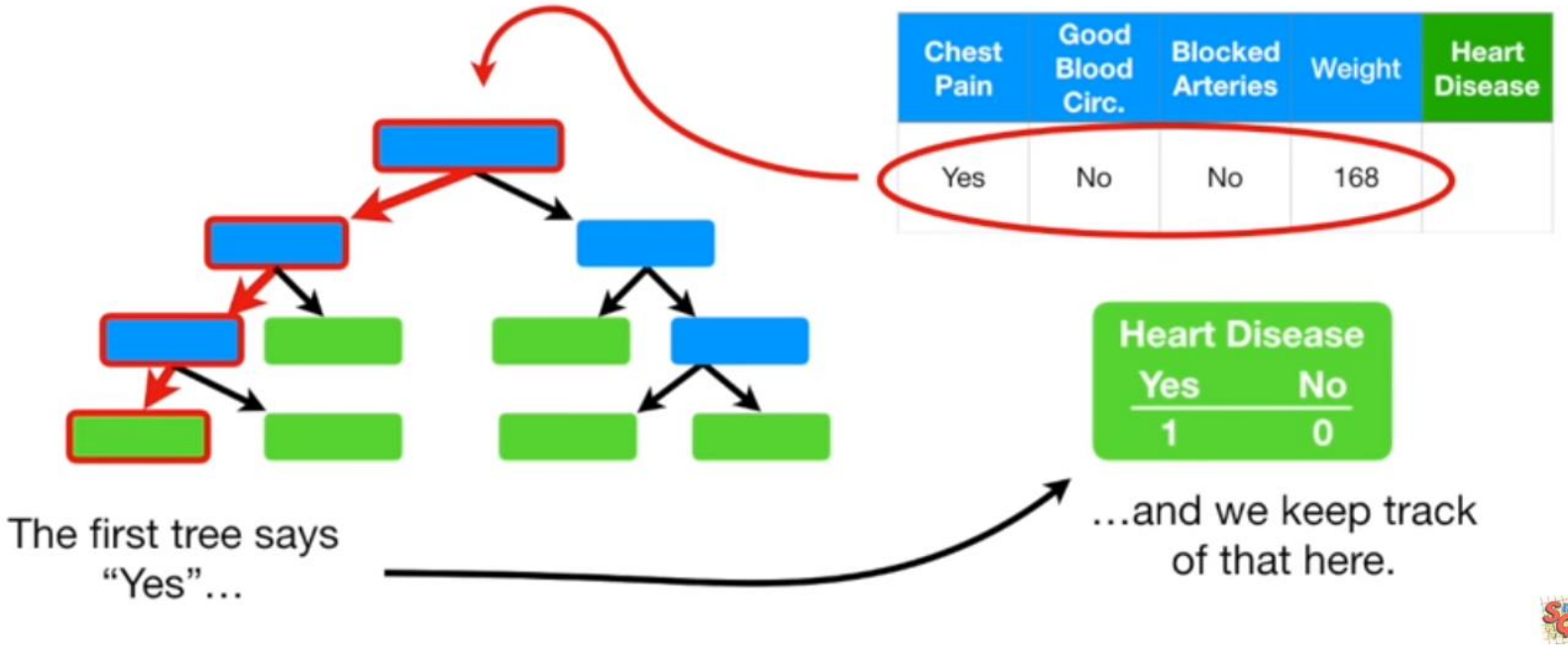
…and now we want to know if they have heart disease or not.

Image courtesy: Random Forest by Statquest

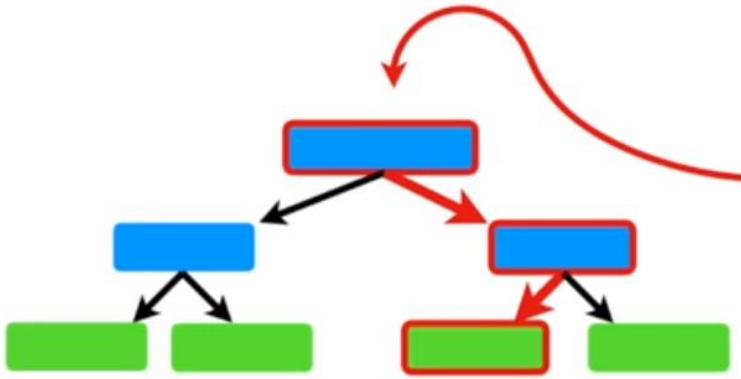
RANDOM FOREST - INTUITION



RANDOM FOREST - INTUITION



RANDOM FOREST - INTUITION

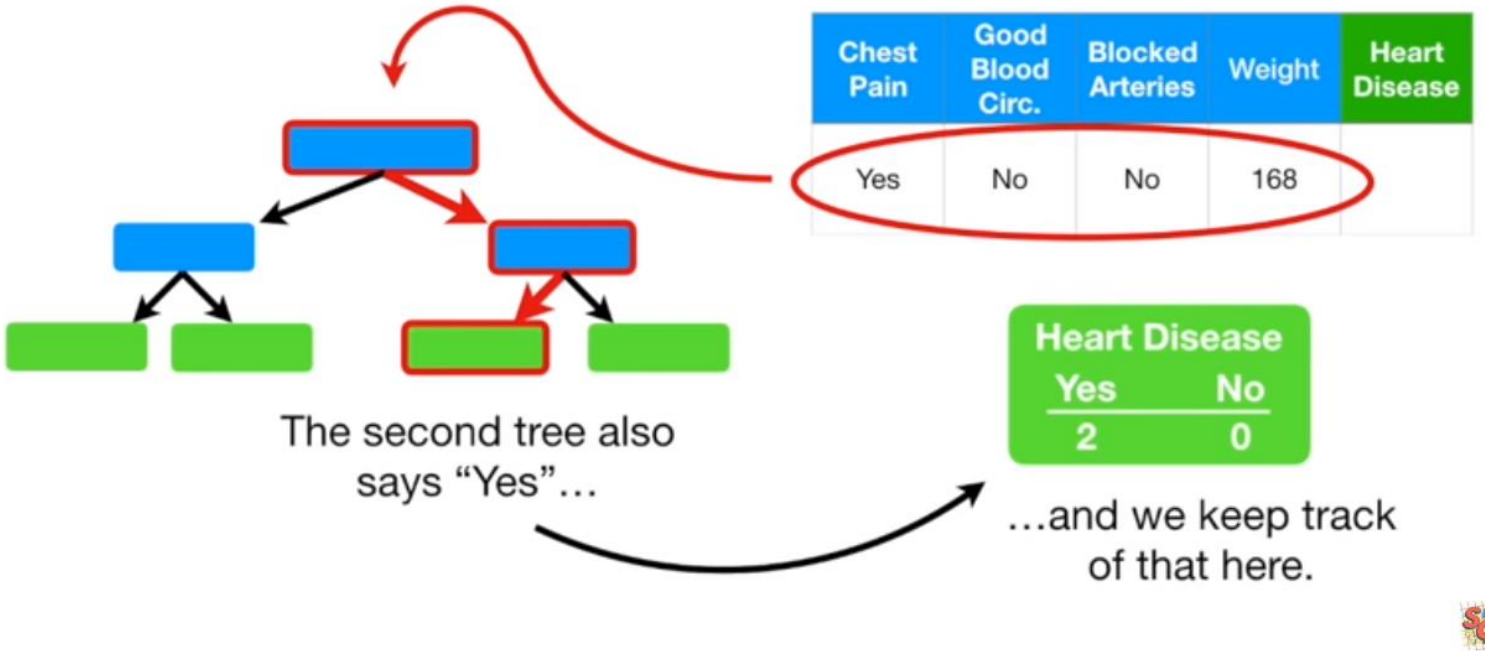


The second tree also says “Yes”…

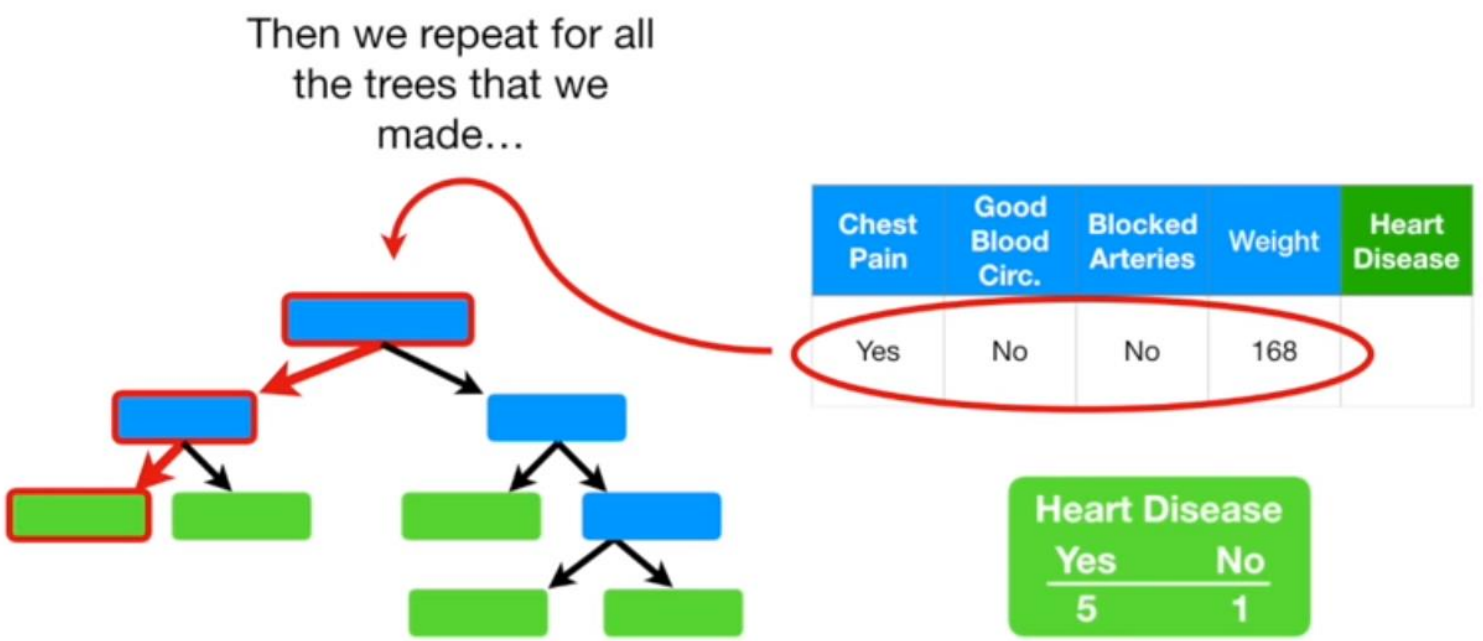
Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	No	168	

Heart Disease	
Yes	No
1	0

RANDOM FOREST - INTUITION



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

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.