

# Advance Ensemble Techniques

## 1) BAGGING

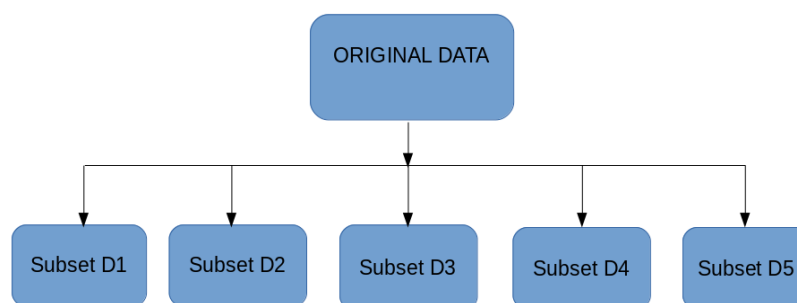
The idea behind bagging is combining the results of multiple models (for instance, all decision trees) to get a generalized result. Here's a question: If you create all the models on the same set of data and combine it, will it be useful? There is a high chance that these models will give the same result since they are getting the same input. So how can we solve this problem? One of the techniques is bootstrapping.

Bootstrapping is a sampling technique in which we create subsets of observations from the original dataset, **with replacement**. Bagging also called **Bootstrap Aggregating** technique uses these subsets (bags) to get a fair idea of the distribution (complete set). The size of subsets created for bagging may be less than the original set.

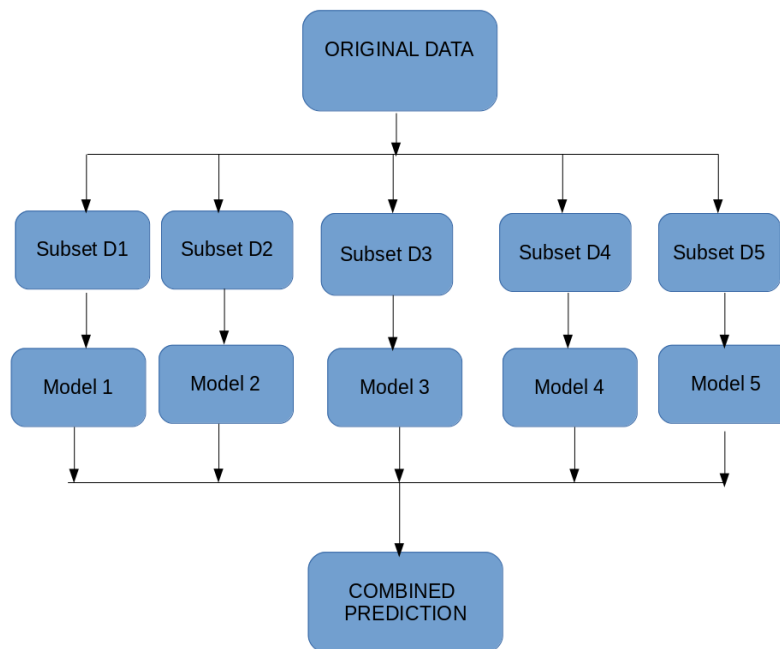
Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as **row sampling**. This step of row sampling with replacement is called **bootstrap**.

Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as **aggregation**.

- **Creates a different training subset from sample training data with replacement**
- **The final output is based on majority voting. For example, Random Forest.**



1. Multiple subsets are created from the original dataset, selecting observations with replacement.
2. A base model (weak model) is created on each of these subsets.
3. The models run in parallel and are independent of each other.
4. The final predictions are determined by combining the predictions from all the models.



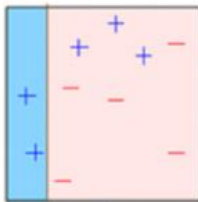
## 2) Boosting

It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST

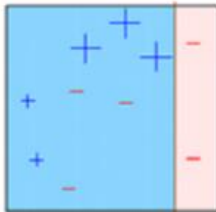
If a data point is incorrectly predicted by the first model, and then the next (probably all models), will combining the predictions provide better results? Such situations are taken care of by boosting.

Boosting is a sequential process, where each subsequent model attempts to correct the errors of the previous model. The succeeding models are dependent on the previous model. Let's understand the way boosting works in the below steps.

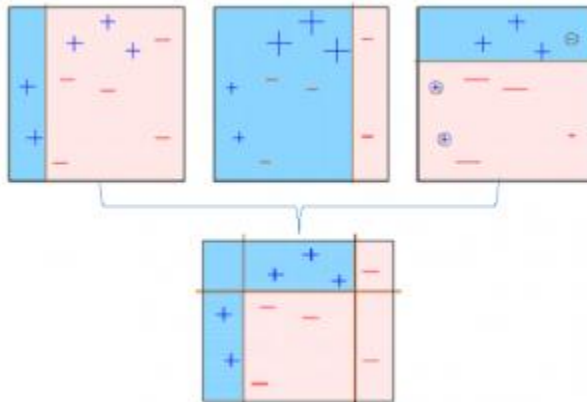
1. A subset is created from the original dataset.
2. Initially, all data points are given equal weights.
3. A base model is created on this subset.
4. This model is used to make predictions on the whole dataset.



5. Errors are calculated using the actual values and predicted values.
6. The observations which are incorrectly predicted, are given higher weights.  
(Here, the three misclassified blue-plus points will be given higher weights)
7. Another model is created and predictions are made on the dataset.  
(This model tries to correct the errors from the previous model)



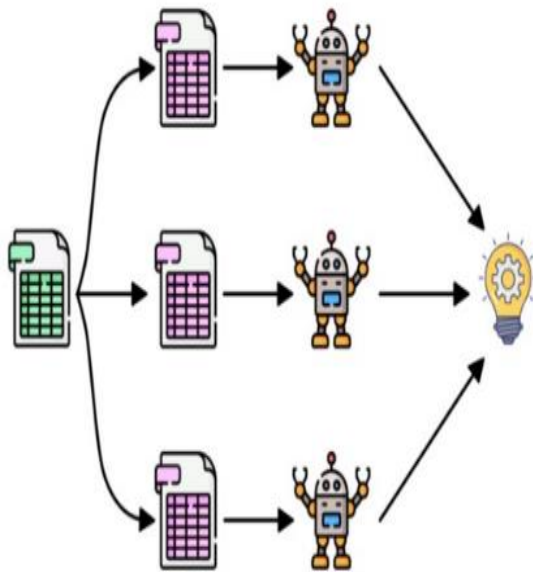
8. Similarly, multiple models are created, each correcting the errors of the previous model.
9. The final model (strong learner) is the weighted mean of all the models (Weak learners).



Thus, the boosting algorithm combines a number of weak learners to form a strong learner. The individual models would not perform well on the entire dataset, but they work well for some part of the dataset. Thus, each model actually boosts the performance of the ensemble.

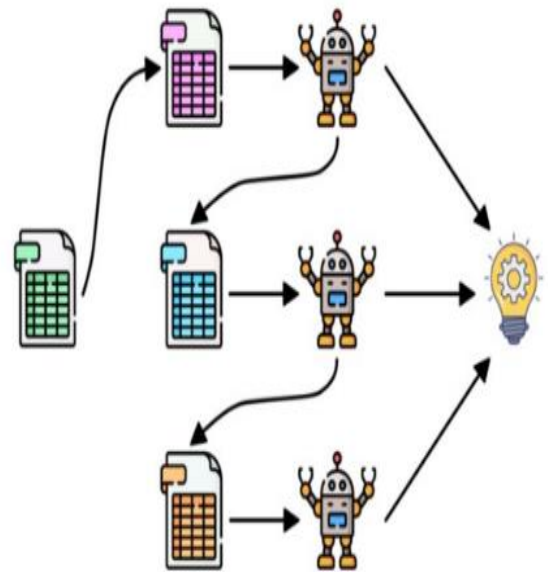


## Bagging



## Parallel

## Boosting



## Sequential

**Random forest** is a **Supervised Machine Learning Algorithm** that is **used widely in Classification and Regression problems**. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing **continuous variables** as in the case of

regression and ***categorical variables*** as in the case of classification. It performs better results for classification problems.

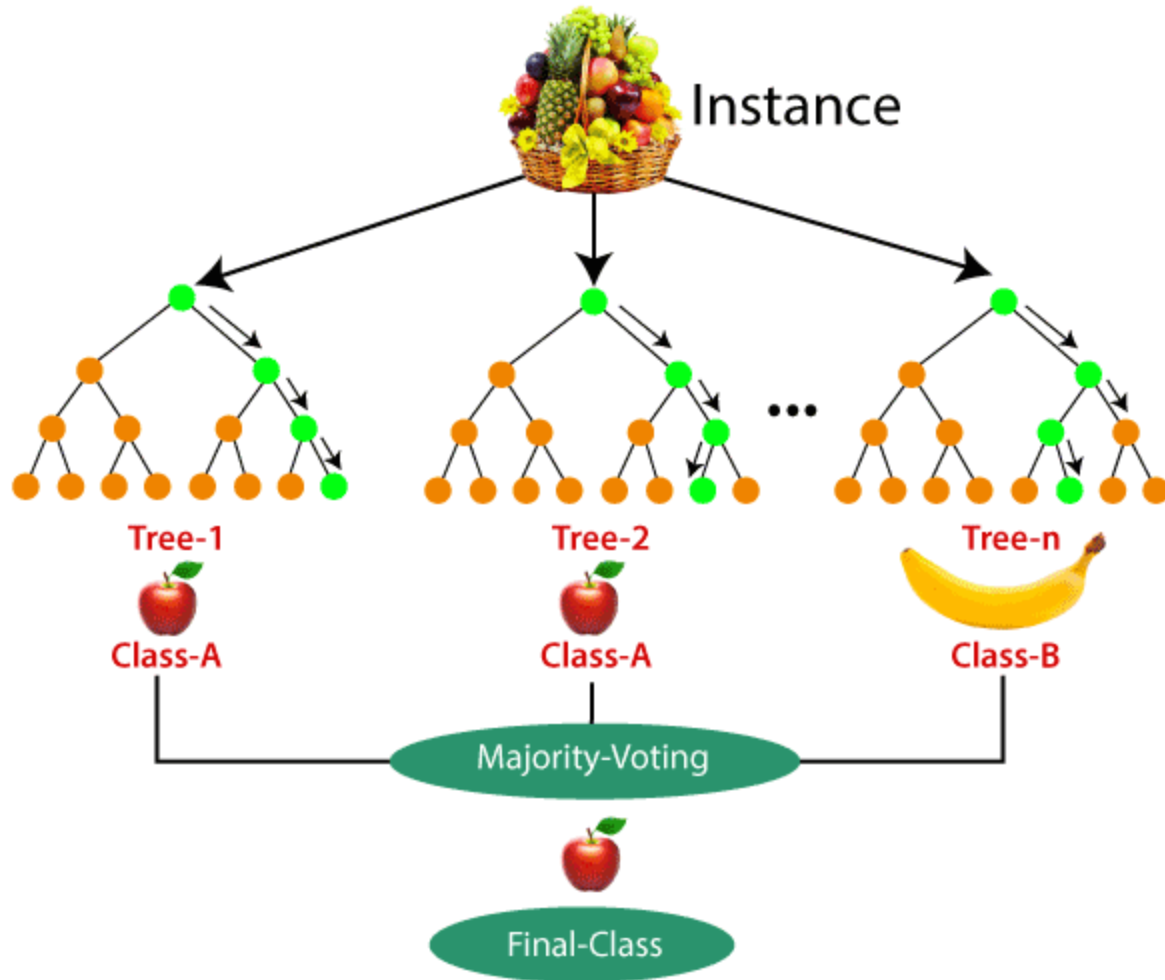
### **Steps involved in random forest algorithm:**

Step 1: In Random forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on ***Majority Voting or Averaging*** for Classification and regression respectively.



## Important Features of Random Forest

- 1. Diversity-** Not all attributes/variables/features are considered while making an individual tree, each tree is different.
- 2. Immune to the curse of dimensionality-** Since each tree does not consider all the features, the feature space is reduced.
- 3. Parallelization-** Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.

**4. Train-Test split-** In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.

**5. Stability-** Stability arises because the result is based on majority voting/averaging.

### **Difference Between Decision Tree & Random Forest**

<b>Decision trees</b>	<b>Random Forest</b>
1. Decision trees normally suffer from the problem of overfitting if it's allowed to grow without any control.	1. Random forests are created from subsets of data and the final output is based on average or majority ranking and hence the problem of overfitting is taken care of.
2. A single decision tree is faster in computation.	2. It is comparatively slower.
3. When a data set with features is taken as input by a decision tree it will formulate some set of rules to do prediction.	3. Random forest randomly selects observations, builds a decision tree and the average result is taken. It doesn't use any set of formulas.

## **Use Cases of Random Forest**

This algorithm is widely used in E-commerce, banking, medicine, the stock market, etc. For example: In the Banking industry it can be used to find which customer will default on the loan.



# Advantages and Disadvantages of Random Forest Algorithm

## Advantages

1. It can be used in classification and regression problems.
2. It solves the problem of overfitting as output is based on majority voting or averaging.
3. It performs well even if the data contains null/missing values.
4. Each decision tree created is independent of the other thus it shows the property of parallelization.
5. It is highly stable as the average answers given by a large number of trees are taken.
6. It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.

## Disadvantages

1. Random forest is highly complex when compared to decision trees where decisions can be made by following the path of the tree.
2. Training time is more compared to other models due to its complexity. Whenever it has to make a prediction each decision tree has to generate output for the given input data.

## Differences Between Bagging and Boosting

NO.	Bagging	Boosting
1.	The simplest way of combining predictions that belong to the same type.	A way of combining predictions that belong to the different types.
2.	Aim to decrease variance, not bias.	Aim to decrease bias, not variance.
3.	Each model receives equal weight.	Models are weighted according to their performance.
4.	Each model is built independently.	New models are influenced by the performance of previously built models.
5.	Different training data subsets are selected using row sampling with replacement and random sampling methods from the entire training dataset.	Every new subset contains the elements that were misclassified by previous models.
6.	Bagging tries to solve the over-fitting problem.	Boosting tries to reduce bias.
7.	If the classifier is unstable (high variance), then apply bagging.	If the classifier is stable and simple (high bias) then apply boosting.
8.	In this base classifiers are trained parallelly.	In this base classifiers are trained sequentially.
	Example: The Random forest model use Bagging.	Example: The AdaBoost uses Boosting techniques