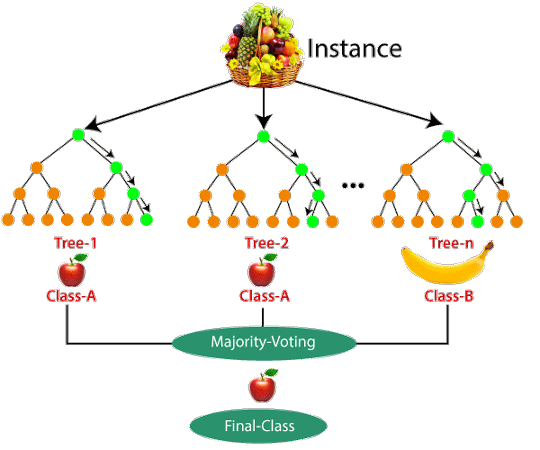
Steps Involved in Random Forest Algorithm

* **Step 1:**In the Random forest model, a subset of data points and a subset of features is selected for constructing each decision tree. Simply put, n random records and m features 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.

**For example**

Consider the fruit basket as the data as shown in the figure below. Now n number of samples are taken from the fruit basket, and an individual decision tree is constructed for each sample. Each decision tree will generate an output, as shown in the figure. The final output is considered based on majority voting. In the below figure, you can see that the majority decision tree gives output as an apple when compared to a banana, so the final output is taken as an apple.



Important Features of Random Forest

* **Diversity:**Not all attributes/variables/features are considered while making an individual tree; each tree is different.
* **Immune to the curse of dimensionality:** Since each tree does not consider all the features, the feature space is reduced.
* **Parallelization:**Each tree is created independently out of different data and attributes. This means we can fully use the CPU to build random forests.
* **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.
* **Stability:**Stability arises because the result is based on majority voting/ averaging.

Difference Between Decision Tree and Random Forest

Random forest is a collection of decision trees; still, there are a lot of differences in their behavior.

|  |  |
| --- | --- |
| **Decision trees** | **Random Forest** |
| 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; 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 rules to make predictions. | 3. Random forest randomly selects observations, builds a decision tree, and takes the average result. It doesn’t use any set of formulas. |

Thus random forests are much more successful than decision trees only if the trees are diverse and acceptable.

Important Hyperparameters in Random Forest

Hyperparameters are used in random forests to either enhance the performance and predictive power of models or to make the model faster.

Hyperparameters to Increase the Predictive Power

* **n\_estimators:** Number of trees the algorithm builds before averaging the predictions.
* **max\_features:** Maximum number of features random forest considers splitting a node.
* **mini\_sample\_leaf:** Determines the minimum number of leaves required to split an internal node.
* **criterion:** How to split the node in each tree? (Entropy/Gini impurity/Log Loss)
* **max\_leaf\_nodes:**Maximum leaf nodes in each tree

Hyperparameters to Increase the Speed

* ***n\_jobs:***it tells the engine how many processors it is allowed to use. If the value is 1, it can use only one processor, but if the value is -1, there is no limit.
* ***random\_state:***controls randomness of the sample. The model will always produce the same results if it has a definite value of random state and has been given the same hyperparameters and training data.
* ***oob\_score:****OOB* means out of the bag. It is a random forest cross-validation method. In this, one-third of the sample is not used to train the data; instead used to evaluate its performance. These samples are called out-of-bag samples.

Advantages and Disadvantages of Random Forest Algorithm

Advantages

* It can be used in classification and regression problems.
* It solves the problem of overfitting as output is based on majority voting or averaging.
* It performs well even if the data contains null/missing values.
* Each decision tree created is independent of the other; thus, it shows the property of parallelization.
* It is highly stable as the average answers given by a large number of trees are taken.
* It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.
* It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.
* We don’t have to segregate data into train and test as there will always be 30% of the data, which is not seen by the decision tree made out of bootstrap.

Disadvantages

* Random forest is highly complex compared to decision trees, where decisions can be made by following the path of the tree.
* Training time is more than 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.

Conclusion

Random forest is a great choice if anyone wants to build the model fast and efficiently, as one of the best things about the random forest is it can handle missing values. It is one of the best techniques with high performance, widely used in various industries for its efficiency. It can handle binary, continuous, and categorical data. Overall, random forest is a fast, simple, flexible, and robust model with some limitations.

**Key Takeaways**

* Random forest algorithm is an ensemble learning technique combining numerous classifiers to enhance a model’s performance.
* Random Forest is a supervised machine-learning algorithm made up of decision trees.
* Random Forest is used for both classification and regression problems.