

# Random Forest

Epoch IIT Hyderabad

Himani Agrawal  
MA22BTECH11008

## Contents

- **Introduction**
- **Working of random forest algorithm**
- **Assumptions and use**

## Introduction

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve predictive accuracy and reduce overfitting. Random Forests are powerful and versatile algorithms known for their robustness and ability to handle various types of data. They are widely used for both classification and regression tasks in machine learning.

## Working of Random forest Algorithm

Random Forest works in two phases. The first phase is to create the random forest by combining  $N$  decision trees, and the second phase is to make predictions for each tree created in the first phase. The working process can be explained in the below steps:

1. Select random  $K$  data points from the training set.
2. Build the decision trees associated with the selected data points (Subsets).
3. Choose the number  $N$  for decision trees that you want to build.
4. Repeat Step 1 & 2.

5. For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

## Assumptions and Use

There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result. The predictions from each tree must have very low correlations.

Random Forest takes less training time as compared to other algorithms. It predicts output with high accuracy, even for the large dataset it runs efficiently. It can also maintain accuracy when a large proportion of data is missing

## Key formulas used

Bootstrap Aggregating (Bagging):

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set  $X = x_1, \dots, x_n$  with responses  $Y = y_1, \dots, y_n$ , bagging repeatedly ( $B$  times) selects a random sample with replacement of the training set and fits trees to these samples.

Majority Voting:

For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes. This is not a formula per se, but a rule used to determine the final prediction.