Random Forest Algorithm

Compiled by

Dr. Shashank Shetty

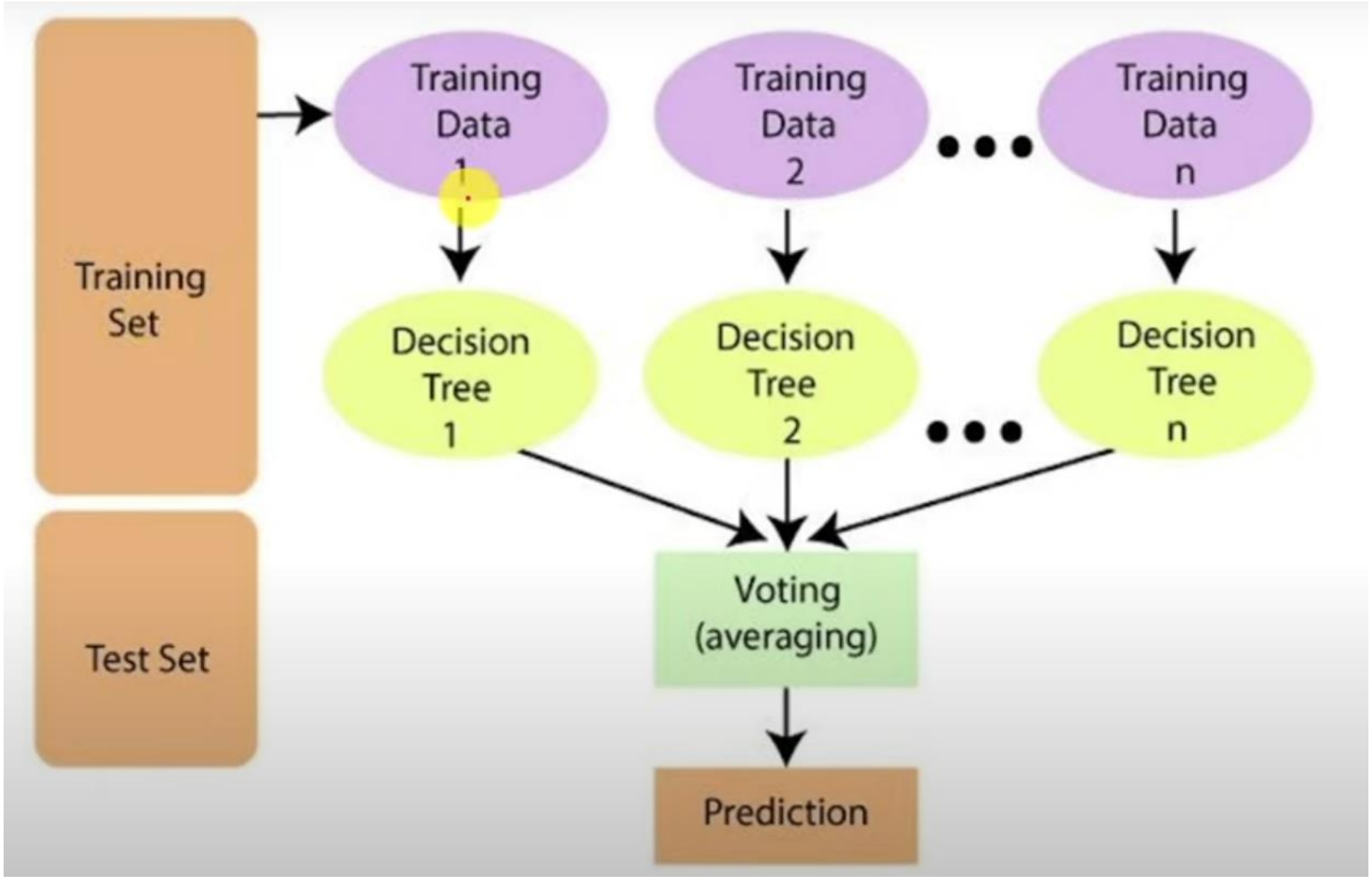
Random Forest Algorithm

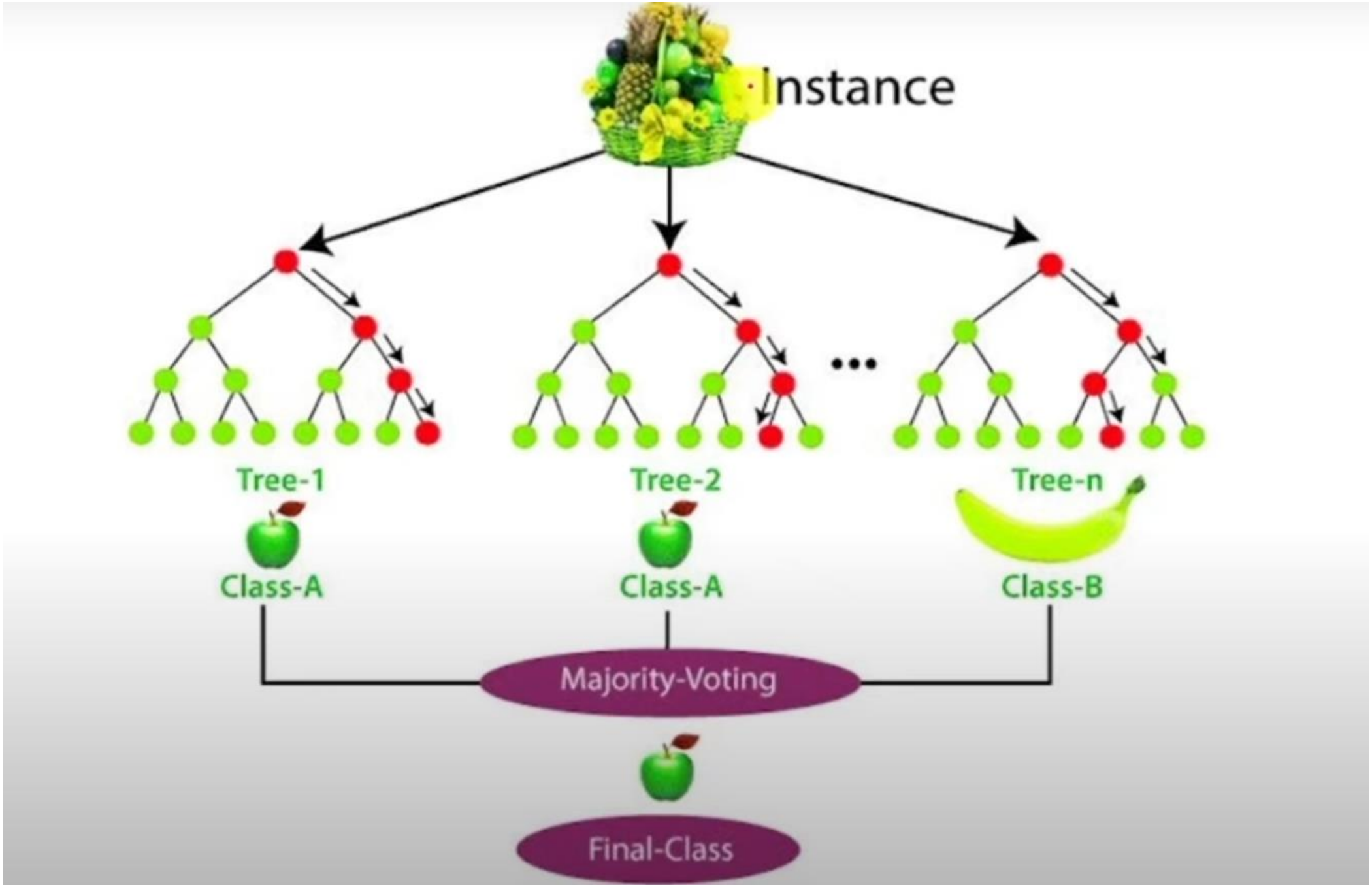
- Random forest is a commonly-used machine learning algorithm.
- A random forest is an ensemble learning method where multiple decision trees are constructed and then they are merged to get a more accurate prediction.
- Random forest became popular because of its ease of use and flexibility in handling both classification and regression problems.

Random Forest Algorithm - Steps

1. Build random forests :
   a) If the number of examples in the training set is N, take a sample of n examples at random - but with replacement, from the original data. This sample will be the training set for generating the tree.
   b) If there are M input variables, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the generation of the various trees in the forest.
   c) Each tree is grown to the largest extent possible.

2. 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**.


Training Set
Test Set
Training Data 1
Training Data 2
Training Data n
Decision Tree 1
Decision Tree 2
Decision Tree n
Voting (averaging)
Prediction


Instance
Tree-1
Tree-2
Tree-n
Class-A
Class-A
Class-B
Majority-Voting
Final-Class

Random Forest Algorithm - Strengths

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

Random Forest Algorithm - Weaknesses

1. A weakness of random forest algorithms is that when used for regression they cannot predict beyond the range in the training data, and that they may over-fit data sets that are particularly noisy.
2. The sizes of the models created by random forests may be very large. It may take hundreds of megabytes of memory and may be slow to evaluate.
3. Random forest models are black boxes that are very hard to interpret.