sports) for the response variable, the sports analyst would use a multinomial logistic regression model.

Ordinal Logistic Regression

Ordinal logistic regression models are a type of logistic regression in which the response variable can belong to one of three or more categories and there *is* a natural ordering among the categories. Here are a couple examples:

Example 1: School Ratings

Suppose an academic advisor wants to use the predictor variables (1) GPA, (2) ACT score, and (3) SAT score to predict the probability that an individual will get into a university that can be categorized into "bad", "mediocre", "good", or "great." Since there are more than two possible outcomes (there are four classifications of school quality) for the response variable and there *is* a natural ordering among the outcomes, the academic advisor would use an ordinal logistic regression model.

Example 2: Movie Ratings

Suppose a movie critic wants to use the predictor variables (1) total run time and (2) genre to predict the probability that a given movie will receiving a rating between 1 and 10. Since there are more than two possible outcomes (there are 10 possible ratings) for the response variable and there *is* a natural ordering among the outcomes, the movie critic would use an ordinal logistic regression mode

3.5 *Random Forest Algorithm*

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The

greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The below diagram explains the working of the Random Forest algorithm:
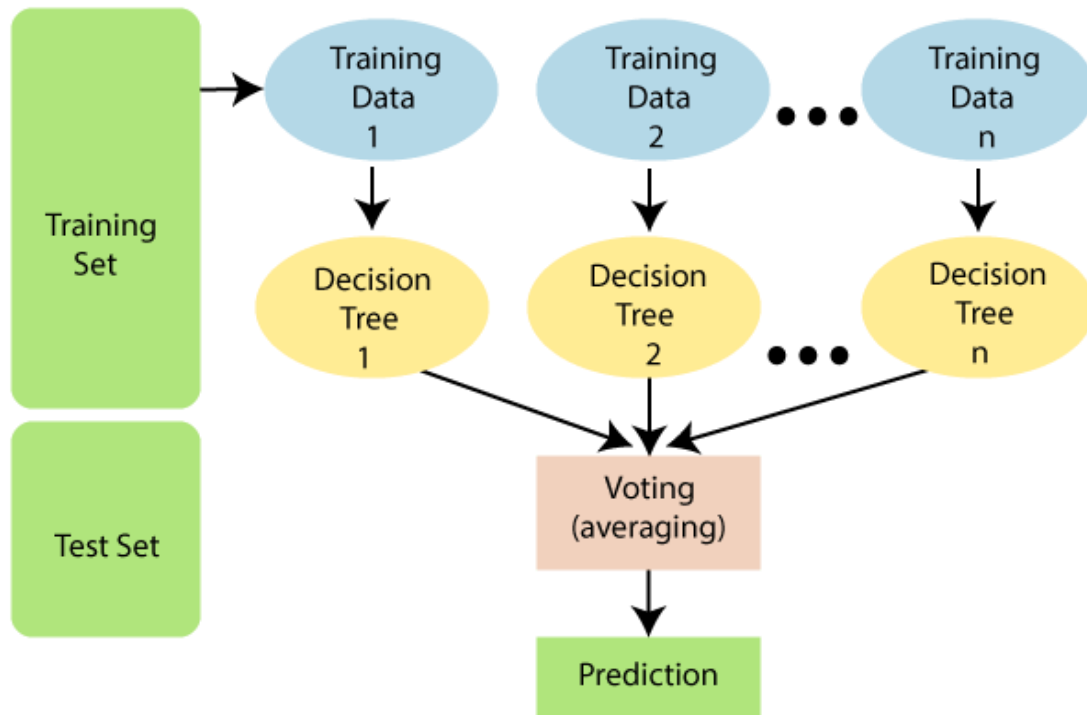
Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

- 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.

Why to use Random Forest?

Below are some points that explain why we should use the Random Forest algorithm:

- It 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.

How does Random Forest algorithm work?

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

**Step-2:** Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-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.

The working of the algorithm can be better understood by the below example:

Example: Suppose there is a dataset that contains multiple fruit images. So, this dataset is given to the Random forest classifier. The dataset is divided into subsets and given to each decision tree. During the training phase, each decision tree produces a prediction result, and when a new data point occurs, then based on the majority of results, the Random Forest classifier predicts the final decision. Consider the below image:
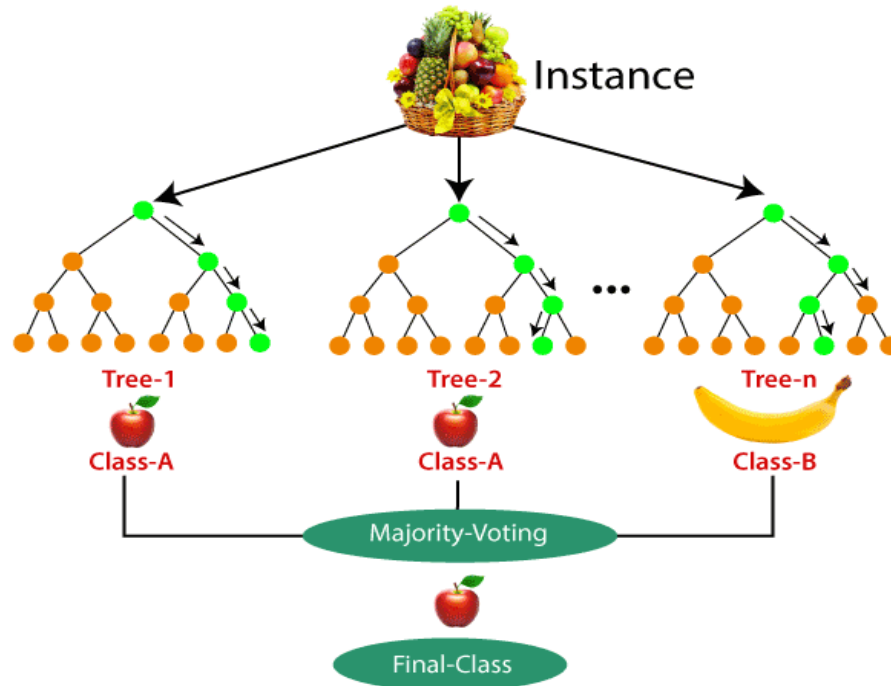
Applications of Random Forest

There are mainly four sectors where Random forest mostly used:

1. **Banking:** Banking sector mostly uses this algorithm for the identification of loan risk.

2. **Medicine:** With the help of this algorithm, disease trends and risks of the disease can be identified.

3. **Land Use:** We can identify the areas of similar land use by this algorithm.

4. **Marketing:** Marketing trends can be identified using this algorithm.

Advantages of Random Forest
- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.

Disadvantages of Random Forest

Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.