Random Forest Classification

Definition - Random Forest is an **ensemble learning** method used for classification (and regression) that builds multiple decision trees and combines their predictions to improve accuracy and robustness. It creates a "forest" of decision trees, where each tree is trained on a random subset of the data and features. The final prediction is made by aggregating the predictions from all the trees, usually through majority voting for classification tasks.

How It Works

1 - **Creating Multiple Trees**: Creating multiple trees in a Random Forest involves generating several decision trees, each trained on a different subset of the training data. This subset is created through bootstrap sampling, where each sample is drawn randomly with replacement from the original dataset. As a result, each tree is exposed to a slightly different version of the data, which helps in capturing various patterns and reducing the risk of overfitting. This diversity among the trees is crucial for the overall performance of the Random Forest.

2 - **Random Feature Selection**: In Random Forest, random feature selection is used to introduce additional diversity among the decision trees. At each node of a tree, a random subset of features is chosen to evaluate possible splits, rather than considering all features. This randomness ensures that the trees are less correlated with each other and that they explore different aspects of the data. It helps to prevent individual trees from becoming too similar and reduces the overall variance of the model.

3 - **Making Predictions**: To make predictions, a Random Forest aggregates the outputs from all the decision trees in the forest. For classification tasks, each tree provides a class label, and the final prediction is determined by **majority voting**, where the class with the most votes is chosen. For regression tasks, the final prediction is the **average** of the predictions from all the trees. This aggregation leverages the collective knowledge of multiple trees to produce a more accurate and robust prediction compared to any single tree.