Short & Sweet (Episode 4)

Random Forest Algorithm

Random Forest Algorithm is another **supervised machine learning algorithm** which is powerful yet simple to understand. Previously, we've seen Decision tree, hence it is inspired from that. First of all, let's see where decision tree is good but still not that much :-

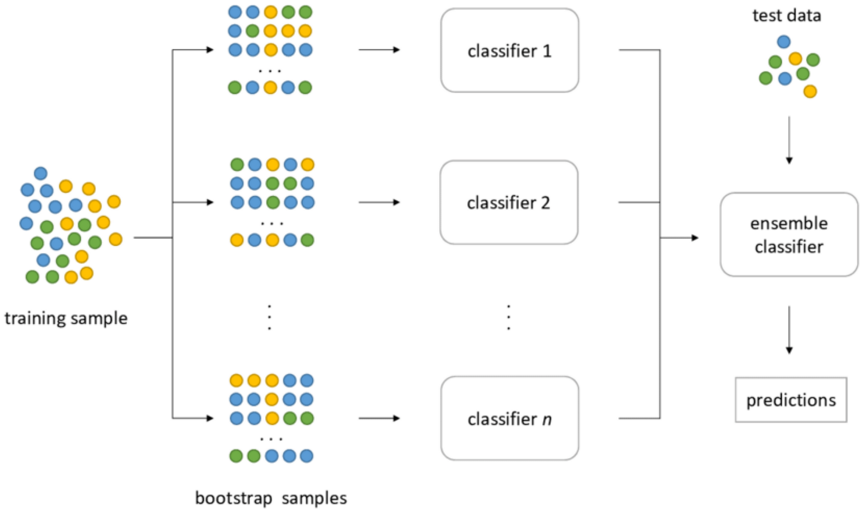
- **Overfitting**: Decision trees are prone to overfitting, especially when they are deep (have many levels or nodes) and capture noise in the data. This means they might perform very well on the training data but poorly on new, unseen data.
- **High Variance**: Small variations in the data can result in different decision tree structures. This high variance can make them less robust.
- **Bias**: Decision trees can be biased toward features with more levels or categories, which might lead to incorrect conclusions.

Random Forest Algorithm is a type of **ensemble learning** which means combining more than one model and aggregating the result of all models that we are going to use.

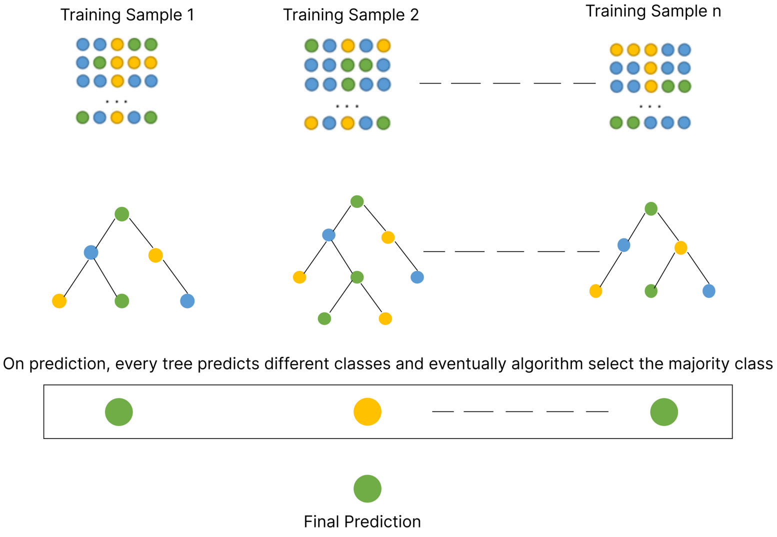
How It Works

As you know about **Decision Tree algorithm**, it analyses the dataset and tries to build the decision tree by selecting the best feature for particular decision tree such that it will give us the **pure split** quickly which we calculate using **Entropy** and **Information Gain**. So now only one tree is not enough as you've seen disadvantages above.

- We'll try to build numbers of decision tree and when we'll predict with our unknown data, random forest selects the majority class or value of all the trees.
- But let's see how this algorithm manages to make decision trees different from each other. For this, **bagging** concept comes into action.
- **Bagging** comes with two terms - **Bootstrapping** and **Aggregation.**
- **Bootstrapping** is a resampling technique in which we will generate some number of dataset by selection of rows randomly (rows can also itself) and for those datasets, algorithm will build various decision tree. Technically, It is a statistical resampling technique used to estimate the sampling distribution of a statistic by repeatedly resampling with replacement from the original dataset. It's a method used for estimating the uncertainty or variability of a sample statistic by generating multiple datasets (bootstrap samples) from the original data.



- **Aggregation** simply means aggregation of result of different decision trees . We'll use class of majority in terms of **classification** and calculate average in terms of **regression**.


Training Sample 1
Training Sample 2
Training Sample n
On prediction, every tree predicts different classes and eventually algorithm select the majority class
Final Prediction

Benefits of Random Forest Algorithm

Random Forests, on the other hand, are an ensemble method based on decision trees. They address some of the limitations of decision trees:

- **Reduced Overfitting**: Random Forests mitigate overfitting by combining multiple decision trees and aggregating their predictions, which helps generalize better to new data.
- **Improved Accuracy**: By averaging or voting on the results of many trees, random forests often provide more accurate predictions than individual decision trees.

- **Feature Importance**: Random Forests can provide information about feature importance, which helps in understanding which features are more influential in making predictions.
- **Less Sensitivity to Noise**: They are less sensitive to noise in the data compared to individual decision trees.