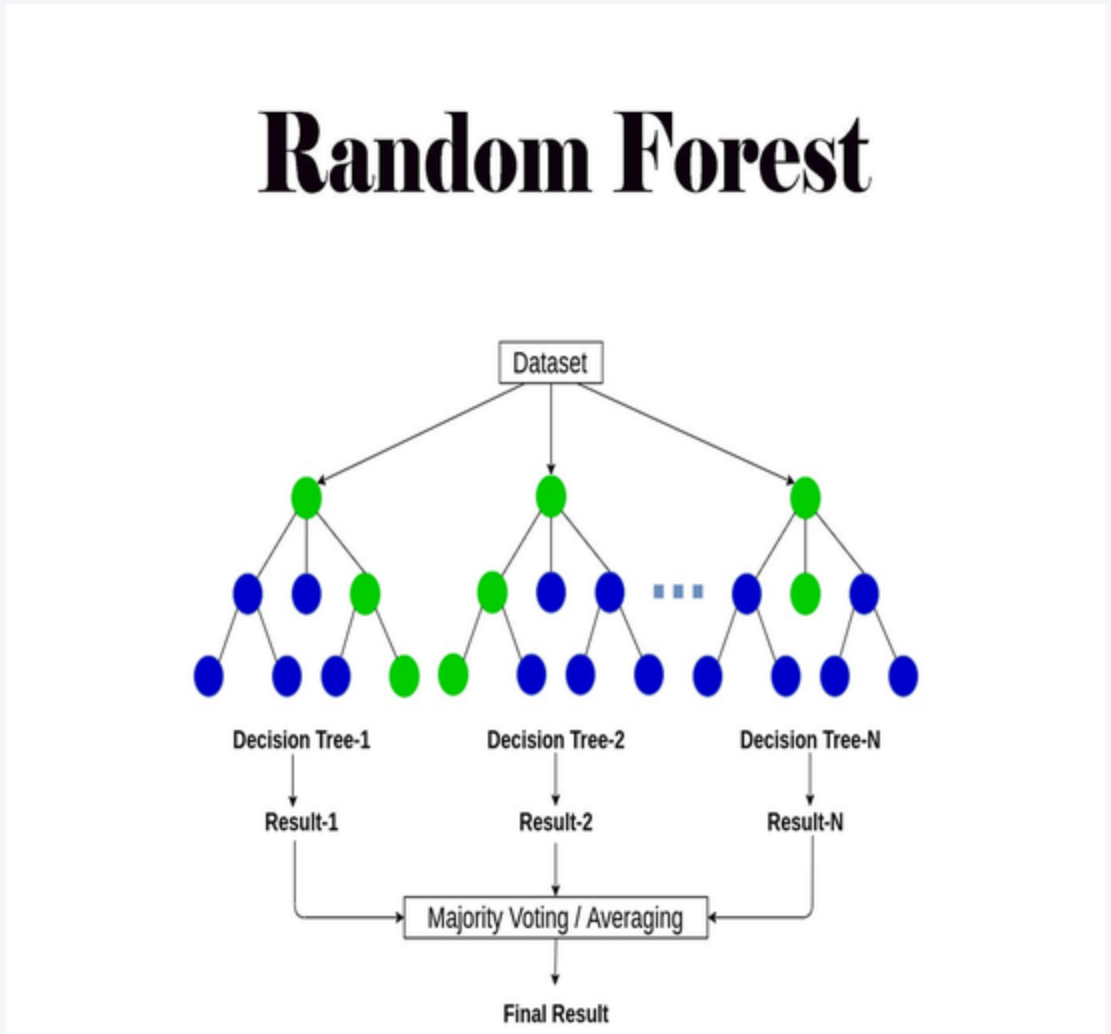
Random Forests — the power of multiple decision trees.



The Weakness of One

A single decision tree can be a powerful predictor. It asks a series of yes/no questions to classify data or predict values. However, this strength is also its biggest weakness.

A single tree often becomes too complex and tailored to the training data. It memorizes the noise and specific details instead of learning the general pattern. This leads to poor performance on new, unseen data, a problem known as overfitting.

Strength in Numbers

The core idea of a Random Forest is simple yet powerful. Instead of relying on one tree, we build an entire forest of them. Each tree is trained on a slightly different subset of the data and asks different questions.

When making a prediction, every tree in the forest gets a vote. For classification, the final prediction is the majority vote from all trees. For regression, the prediction is the average of all the tree outputs.

This collaborative approach cancels out individual errors. While one tree might be wrong, the collective wisdom of hundreds of trees is remarkably accurate and stable.

Two Key Ingredients

Random Forests use two main techniques to ensure each tree is unique. The first is Bootstrap Aggregating, or Bagging. Each tree is trained on a random sample of the original data, drawn with replacement.

The second technique is feature randomness. When a tree is splitting a node, it is not allowed to choose from all features. Instead, it must select from a random subset of features, typically the square root of the total number.

This combination ensures diversity in the forest. The trees are decorrelated, meaning they make different mistakes, which is crucial for the ensemble to work effectively.

Building the Forest

The algorithm starts by creating multiple bootstrap samples from the training data. Each sample is the same size as the original dataset but will contain some duplicates and miss some original points.

For each bootstrap sample, a decision tree is grown. The tree is built by recursively splitting nodes. The split at each node is found using only the predefined random subset of features.

The trees are grown deeply and are not pruned. We allow them to overfit on their specific data sample. Their high variance is what the ensemble method will later average out.

Making a Prediction

For a new data point, we run it through every single tree in the forest. Each tree will output its own individual prediction based on its unique structure.

For a classification task, we collect the predicted class from every tree. The final forest prediction is the class that receives the most votes from all the trees. This is called majority voting.

For a regression task, we collect the predicted value from every tree. The final forest prediction is simply the average of all these individual tree predictions.

Why It Works

Random Forests are robust because of the "wisdom of the crowd" principle. A large number of uncorrelated models operating together will outperform any single constituent model.

The model is highly resistant to overfitting. The bagging and feature randomness introduce stability, and the averaging process smooths out the noise learned by individual trees.

It also handles missing data well and provides good performance without extensive parameter tuning. It is less sensitive to the specific training data than a single decision tree.

Key Advantages

One major advantage is high accuracy. They are among the most accurate learning algorithms available for many types of problems. They often achieve excellent results right out of the box.

They can handle large datasets with higher dimensionality very efficiently. The built-in feature randomness helps in managing a large number of input features.

They also provide built-in estimates of feature importance. The model can calculate which features were most influential in making predictions, offering valuable insights.

Wrapping Up

Random Forests harness the collective power of many weak learners, decision trees, to create a strong and robust ensemble model. They are a cornerstone of modern machine learning.

Their power comes from introducing randomness through bagging and feature subsets. This randomness creates diversity, which is averaged out to produce a superior result.

They are versatile, working for both classification and regression tasks. They are a powerful tool that every data scientist should have in their toolkit for its predictive performance and reliability.

Follow for More

Like this post? Repost!

Share your thoughts in the comments!

I share little hacks and tips I actually use in Data & AI — follow if you wanna see more!