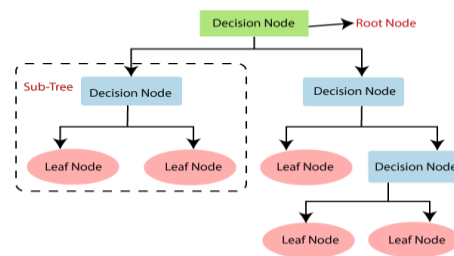


Understanding the Difference between Decision Tree and Random Forest algorithms.

Decision Tree: A decision tree is a flowchart-like tree structure where each internal node denotes a feature, branches denote the rules and the leaf nodes denote the result of the algorithm. It is a versatile supervised machine-learning algorithm, which is used for both classification and regression problems[3].



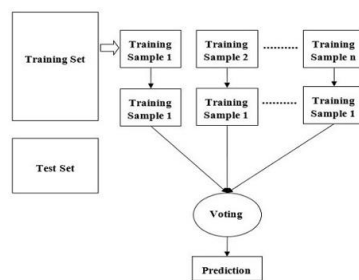
- **Root Node:** It is the topmost node in the tree, which represents the complete dataset.
It is the starting point of the decision-making process.
- **Decision/Internal Node:** A node that symbolizes a choice regarding an input feature.
Branching off of internal nodes connects them to leaf nodes or other internal nodes.
- **Leaf/Terminal Node:** A node without any child nodes that indicates a class label or a numerical value.
- **Splitting:** The process of splitting a node into two or more sub-nodes using a split criterion and a selected feature.
- **Branch/Sub-Tree:** A subsection of the decision tree starts at an internal node and ends at the leaf nodes.
- **Parent Node:** The node that divides into one or more child nodes. **Child Node:** The nodes that emerge when a parent node is split.

How do Decision Trees work

The decision to make strategic splits significantly impacts a tree's accuracy. The criteria for decision-making differ between classification and regression trees[3].

Decision trees utilise multiple algorithms to determine whether to split a node into two or more sub-nodes. The creation of sub-nodes enhances the homogeneity of the resulting sub-nodes. In other words, we can say that the purity of the node increases in relation to the target variable. The decision tree splits the nodes based on all available variables and then selects the split that produces the most homogeneous sub-nodes[4].

Random Forest: 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[1]. 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[1].



Random forests can be used for either a categorical response variable, such as "classification", or a continuous response, referred to as "regression". Similarly, expectation variables can be either categorical or continuous[2].

From a computational standpoint, random forests are appealing because they:

- Naturally handle both regression and classification (multilayer).
- Are relatively quick to train and predict.
- Depend on only one or two tuning parameters.
- Provide a built-in estimate of the generalisation error.
- Can be directly applied to high-dimensional problems.
- Can be easily implemented in parallel.

Statistically, random forests are attractive due to the additional features they offer, such as:[5]

- Metrics of variable importance.
- Weighting the differential layer.
- Calculation of missing values.
- The perception

How Random Forest Works:

Random forest is a machine learning technique that combines numerous decision trees to reduce the correlation among feature data. Simultaneously, the computational cost of RF is $O(n)$ (where n represents the number of samples) when dealing with huge amounts of data. Additionally, the technique is executed in parallel due to this integration, resulting in increased speed[4].

Comparison between the Decision Tree and Random Forest

Features	Decision Tree	Random Forest
Model Type	Single Predictive model	Ensemble of multiple decision Trees
Bias/Variance	High variance, Low bias(prone to overfitting)	Low variance, moderate bias(reduces overfitting)
Interpretability	Highly interpretable(white box)	Less interpretable(black box)

References:

- [1].B. Leo, "Random Forests," in *Machine Learning*, vol. 45, pp. 5-32, 2001.
- [2].A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forest," in *Machine Learning Mag.*, DOI: 10.1007/978-1-4419-9326-7_5, Jan. 2011.
- [3].Charbuty, Bahzad, and Adnan Abdulazeez."Classification based on decision tree algorithm for machine learning." *Journal of applied science and technology trends* 2.01 (2021): 20-28.
- [4].S. Sah, "Machine Learning: A Review of Learning Types,.doi:10.20944/preprints202007.0230.v1, 2020.
- [5].A. Abdi, "Three types of Machine Learning Algorithms," DOI: 10.13140/RG.2.2.26209.10088, 2016.
- [6]"Decision Trees in Machine Learning," [Online]. Available: <https://pdf.co/blog/decision-trees-in-machine-learning>.