

Understanding the Difference between Decision Tree and Random Forest Algorithms

Machine learning is a significant field of artificial intelligence that enables computer systems to learn from data and make predictions or decisions without explicit programming. In supervised learning, classification techniques are used to assign data samples to predefined classes based on labeled training data. Among these techniques, Decision Tree and Random Forest algorithms are widely used due to their effectiveness, simplicity, and applicability across various real-world domains such as healthcare, finance, and weather prediction [1]. This activity aims to compare Decision Tree and Random Forest algorithms by examining their working principles, advantages, limitations, and practical use cases.

A Decision Tree is a supervised learning algorithm that represents decisions in a tree-like structure consisting of decision nodes, branches, and leaf nodes. It functions by repeatedly splitting the dataset based on selected attributes using measures such as entropy and information gain, until a final decision is reached. Decision Trees are easy to interpret and understand, as they follow clear rule-based logic that can be easily visualized. However, they are highly sensitive to training data and tend to overfit, which can reduce their performance on unseen datasets [2].

A Random Forest is an ensemble learning technique that builds upon Decision Trees by constructing multiple trees using random subsets of training data and features. Each tree makes a prediction, and the final output is determined through majority voting. This approach improves prediction accuracy and reduces overfitting compared to a single decision tree. Random Forest also provides feature importance measures, helping identify influential variables in a dataset. However, Random Forest models are more computationally expensive and less interpretable due to their complex structure [2].

The advantages of Decision Trees include simplicity, fast execution, and easy interpretation, making them suitable for transparent decision-making tasks. Their limitations include overfitting and lower accuracy on complex datasets. Random Forest offers higher accuracy, robustness, and better handling of large and noisy datasets, but requires more processing time and lacks interpretability. In practice, Decision Trees are preferred when model transparency and speed are important, whereas Random Forest is preferred for applications that require high predictive performance and reliability, such as medical diagnosis, fraud detection, and weather forecasting [2].

In conclusion, both Decision Tree and Random Forest algorithms play an important role in supervised machine learning. The choice between them depends on the nature of the dataset, the need for interpretability, computational resources, and the required level of prediction accuracy.

References

- [1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Southeast Asia Edition, Morgan Kaufmann.
- [2] P. T. R., "A Comparative Study on Decision Tree and Random Forest Using R Tool," International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 1, pp. 196-199, Jan. 2015, Doi: 10.17148/IJARCCE.2015.4142.

