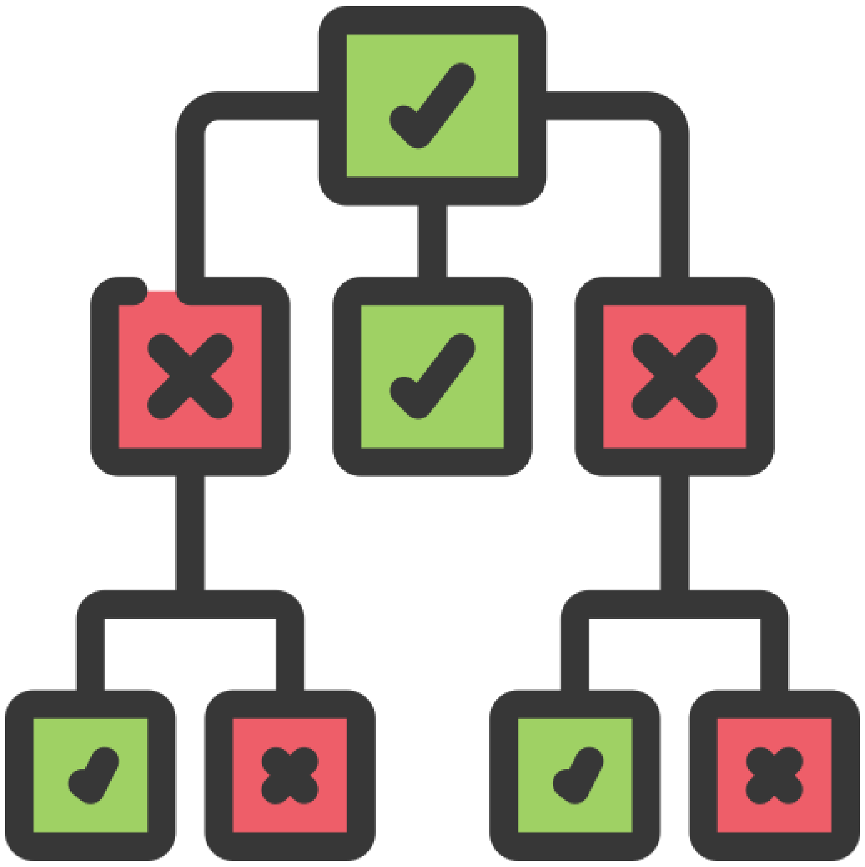
Understanding Decision Trees in ML



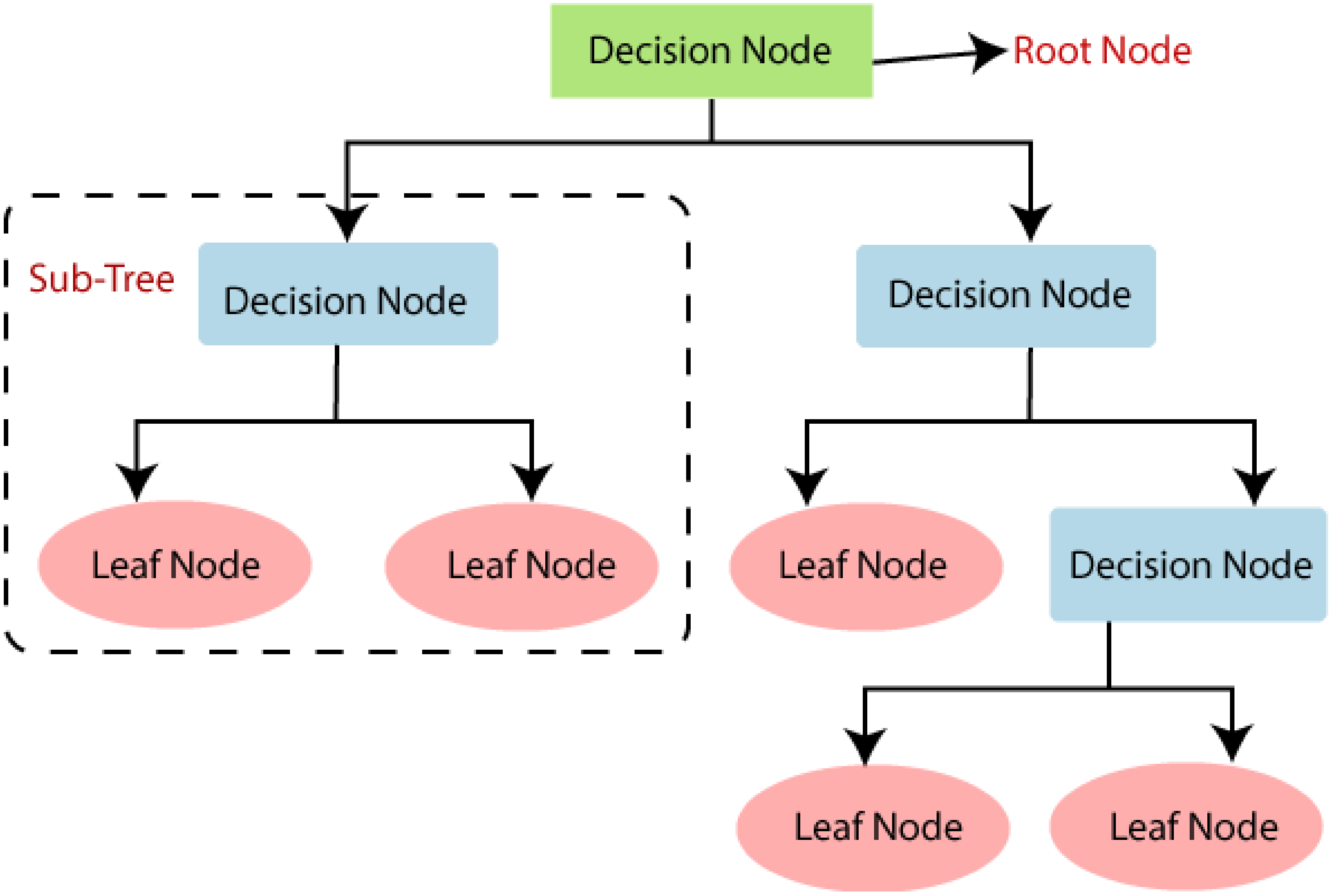
VÍCTOR VILORIA

in/vicvilo

1 . What Is a Decision Tree?

A Decision Tree is a flowchart-like structure in Machine Learning, used for decision-making and predicting outcomes.

It consists of nodes and branches, each representing decisions based on data features and leading to final predictions or classifications.



2 . Key Concepts

- **Root Node:** This is where the decision tree starts. It represents the entire dataset being analyzed, from which the first and most significant decision is made based on a specific attribute.

- **Internal Nodes:** These nodes represent the points where the dataset is split further based on different attributes. Each internal node tests a condition, leading to further branches in the tree.

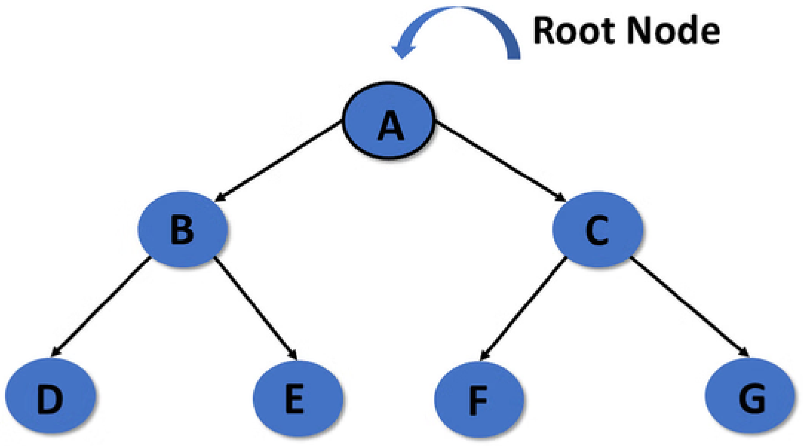
- **Leaf Nodes (or Terminal Nodes):** The endpoints of the tree, where no further splitting occurs. Each leaf node provides the final decision or outcome, such as a class label in classification tasks.

- **Branches:** The connections between nodes, representing the flow from one decision to the next. Each branch corresponds to a possible outcome of the test conducted at the internal node.

3 . How Is the Root Node Selected?

Selecting the most informative feature as the root node is crucial for optimizing the tree's performance. This selection is based on specific criteria that measure the feature's ability to segregate the data effectively. Two primary metrics are:

- **Gini Impurity:** A measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. The feature with the lowest Gini impurity is chosen for the root node to minimize misclassification.
- **Information Gain (Entropy):** This metric evaluates the reduction in entropy or disorder after a dataset is split on an attribute. The attribute that results in the highest information gain is selected, aiming to increase the predictability after the split.



4 . The Branching Process

After selecting the root node, the decision tree algorithm continues to split the dataset into smaller subsets using the remaining features.

- **Creating Internal Nodes:** Each internal node represents a decision point where the dataset is split based on the best feature at that level. The choice of feature is again determined by evaluating which split will most effectively increase the homogeneity of the resulting subsets.

- **Criteria for Splitting:** The algorithm evaluates each potential feature for splitting, calculating Gini impurity or information gain for each possible split. The feature that results in the most significant reduction of uncertainty or disorder is selected for the split at each internal node.

- **Resulting Branches:** Each split creates two or more branches leading to new nodes, which can be further internal nodes (if more splitting is required) or leaf nodes (if a conclusive classification can be made). This process repeats recursively until a stopping criterion is met, such as reaching a maximum depth or achieving sufficiently pure subsets.

5 . Determining Leaf Nodes and Final Decisions

What Defines a Leaf Node? A leaf node is reached when one of the following conditions is met:

- All data points at the node belong to a single class.
- No further attributes are available to split the data.
- Predefined stopping criteria, such as a maximum tree depth or a minimum number of samples at a node, are met.

Making the Final Decision: The decision or classification at a leaf node is determined by the majority class of the data points within that node. In cases of regression, it's typically the mean or median value of the target variable for the data points in the node.

The Role of Leaf Nodes: These nodes are crucial for understanding the decision tree's predictive power. They provide clear, actionable outcomes based on the input data's journey through the tree, reflecting the algorithm's ability to segment and classify the data effectively.

6 . Advantages

- **Interpretability:** Easily understood and visualized, even by those with little to no background in machine learning.

- **No Need for Data Preprocessing:** Can handle both numerical and categorical data and do not require normalization.

- **Handles Non-Linear Relationships:** Capable of capturing complex non-linear relationships between features and labels.

- **Feature Importance:** Naturally performs feature selection, highlighting the most significant features for classification or regression.

7 . Disadvantages

- **Overfitting:** Prone to overfitting, especially with complex trees. Regularization techniques like pruning are necessary to avoid this.

- **Instability:** Small changes in the data can lead to a completely different tree structure, making them sensitive to noise.

- **Biased Trees:** Decision trees can become biased if some classes dominate. Balanced datasets are essential for optimal performance.

- **Difficulty with Unseen Data:** May not perform well on unseen data, particularly if the data has a different distribution from the training set.

Follow me for More Trends, insights and tips in ML and Data Science

VÍCTOR VILORIA

in/vicviloria