Decision Tree in Machine Learning

Definition & Architecture

A decision tree is a supervised machine learning algorithm used for both classification and regression tasks . It features a hierarchical tree structure consisting of a root node, branches, internal nodes, and leaf nodes . Operating much like a flowchart, it enables step-by-step decision-making .

- **Root Node:** The top-most node representing the entire dataset, which splits into further branches based on the primary feature .
- **Internal Nodes:** Represent attribute condition tests (e.g., 'Is Income > $50K?') .
- **Branches:** Represent individual attribute values or test outcomes .
- **Leaf Nodes:** Represent final class labels (classification) or continuous numerical values (regression) .

Visual Architecture of a Decision Tree

The diagram below illustrates the hierarchical decision path from root node to terminal leaf predictions

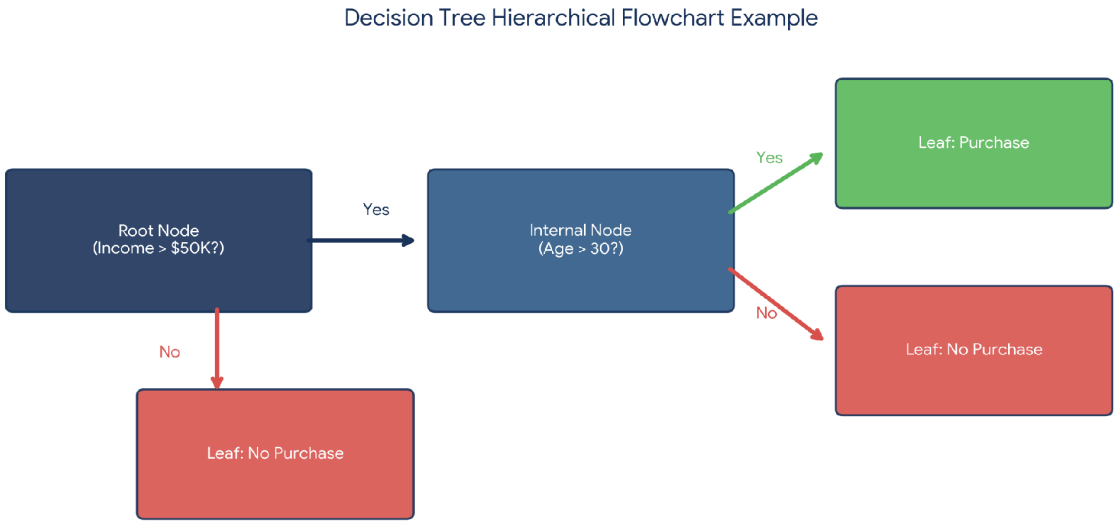


Figure 1: Decision Tree Hierarchical Flowchart Architecture

Attribute Selection Measures (Splitting Criteria)

To determine the optimal feature to split on at each node, decision trees utilize mathematical impurity measures .

1 Information Gain (Entropy)

Information gain measures how much uncertainty decreases after a dataset is split on an attribute . It relies on entropy, which characterizes data impurity .

> **Entropy & Information Gain Formulas .** *Entropy(S) = - ∑ p_i * log2(p_i)*
> *Gain(S, A) = Entropy(S) - ∑ (|S_v| / |S|) * Entropy(S_v)*

- **Selection Rule:** The feature yielding the highest information gain is chosen as the splitting node .

2 Gini Index

The Gini index measures the probability of a randomly chosen element being incorrectly categorized if it were randomly labeled according to the distribution of labels in the subset .

> **Gini Index Formula .** *Gini = 1 - ∑ (p_i)^2*

- **Interpretation:** A Gini index of 0 indicates a completely pure node (all items belong to a single class), whereas higher values indicate greater impurity .

Comparison Table: Information Gain vs. Gini Index

Metric Property	Information Gain (Entropy)	Gini Index
Mathematical Basis	Based on information theory and entropy calculations .	Based on squared class probabilities .
Computational Speed	Slower to compute due to logarithmic ($\log_2$) operations .	Faster to compute with simple algebraic squared terms .
Split Tendency	Tends to favor attributes with many distinct values .	Tends to favor creating equally sized child splits .

Advantages & Limitations of Decision Trees

- **Interpretability:** Highly intuitive and easy to visualize, interpret, and explain to non-technical stakeholders .
- **Low Preprocessing:** Requires minimal data preprocessing; handles both numerical and categorical data without rigorous scaling .
- **Overfitting Risk:** Unpruned trees easily grow overly complex, capturing training noise and leading to severe overfitting .