Definition & Structure

A decision tree is a supervised machine learning model that recursively partitions data into subsets based on feature values. It consists of:

Root Node: The topmost feature that splits the data.

Internal Nodes: Decision points based on features.

Leaf Nodes: Final classifications or regression outputs.

How It Works

Feature Selection: Algorithms choose the best feature to split the data using metrics like Gini impurity (classification) or variance reduction (regression).

Splitting: Data is divided into branches (e.g., "Age < 30" vs. "Age $\geq$ 30").

Stopping Criteria: Splitting stops when a node is pure (all samples belong to one class) or meets depth constraints.

Example: Medical Diagnosis

Features: Age, Blood Pressure, Cholesterol.

Splits:

Root: Blood Pressure > 120?

Yes → Check Cholesterol.

No → Classify as "Healthy."

Advantages

Interpretable (visualizable).

Handles both numerical and categorical data.

Limitations

Prone to overfitting (solved via pruning).

Sensitive to small data changes.