Notes by :- JP Web developers

Decision Tree

- A **Decision Tree** is a supervised learning algorithm used for both classification and regression tasks.

- It is a tree like model used to make decisions or classify data.
  It splits data into branches based on feature values and leads to a final decision.
  Here :-
  - Root node :- first decision.
  - Branches :- Rules based on features.
  - Leaf nodes :- Final o/p (like Yes/No).

#JPNotes

Key Concepts

Term	Description
• Root Node	The top node, represents the entire dataset.
• Decision Node	A node where the data splits based on a feature.
• Leaf / Terminal Node	final node with a Prediction. (final o/p (e.g. Yes/No).
• Branches	Paths from decision node to next node.

- Splitting — It is the process of diving the root node into sub nodes
- Pruning — It is the process of removing the unwanted branches from the tree.

How it Works (Step-by-Step)

1. Start with the full dataset.
2. Choose the best feature to split on using:-
   - Gini index
   - Entropy / information Gain
3. Split the dataset into groups.
4. Repeat steps 2 and 3 for each group.
5. Stop when:
   - All samples in a node are of the same class, or
   - Tree reaches max depth.

Splitting Criteria — Attribute Selection Measure

When building a Decision Tree, we need to choose the best feature to split the data

#JP Web developers

This is done using Attribute Selection Measures

Name	What is Does	Used in
Gini Index	Measures Impurity (Lower is better)	CART algorithm
Entropy	Measures disorder (We want to reduce it)	ID3 algorithm
Information Gain	How much a split improves purity	ID3

#JPNotes

Information Gain (Used in ID3)

It tells us how much a feature helps to reduce uncertainity (entropy).

formula

IG = Entropy (Parent) - weighted Avg Entropy (children)

Entropy Formula

Entropy (S) = $-P(Yes) \log_2 P(Yes) - P(no) \log_2 P(no)$

P(Yes) - Proportion of +ve samples

P(no) - " " " -ve " "

Gini index (Used in CART algorithm)

It measures how often a randomly chosen element would be incorrectly classified.

Lower Gini = better Purity

formula

Gini index = $1 - \sum P_j^2$

where

P_j = Probability of class j

Pruning

To avoid overfitting, trees are pruned.

- **Pre-Pruning**: Stop early (e.g. max depth = 3)
- **Post-Pruning**: Remove unnecessary branches after full tree is built.

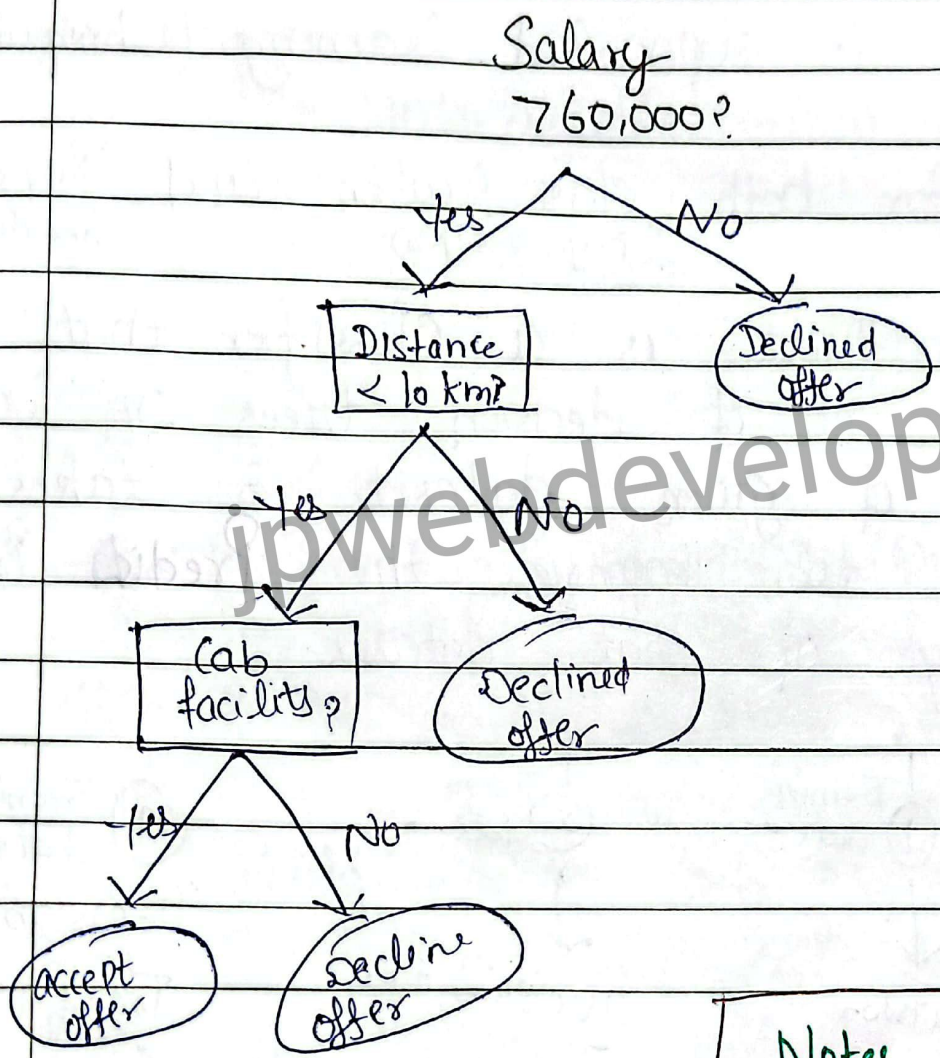
Example: Should I Accept the Job Offer?

A candidate has a Job offer and is confused whether to accept or reject it. A Decision Tree helps him decide step by step based on important features like salary, distance and Cab facility.

Decision Criteria

Attribute	Condition	Decision Taken
Salary	> 60,000	Go to next step: check Distance
	≤ 60,000	✗ Decline Offer

Distance	< 10 km	Go to next step: check cab facility
	≥ 10 km	✗ Decline Offer
Cab Facility	Available	✓ Accept offer
	Not Available	✗ decline "



Notes by :-
@jpwebdevelopers.
#JPNotes
JPWebdevelopers.