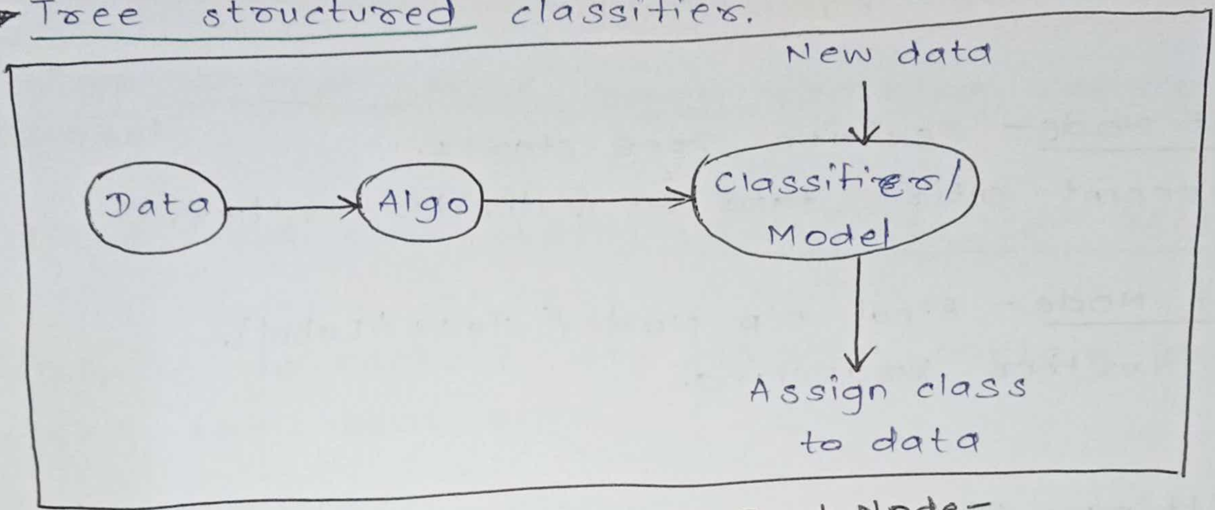
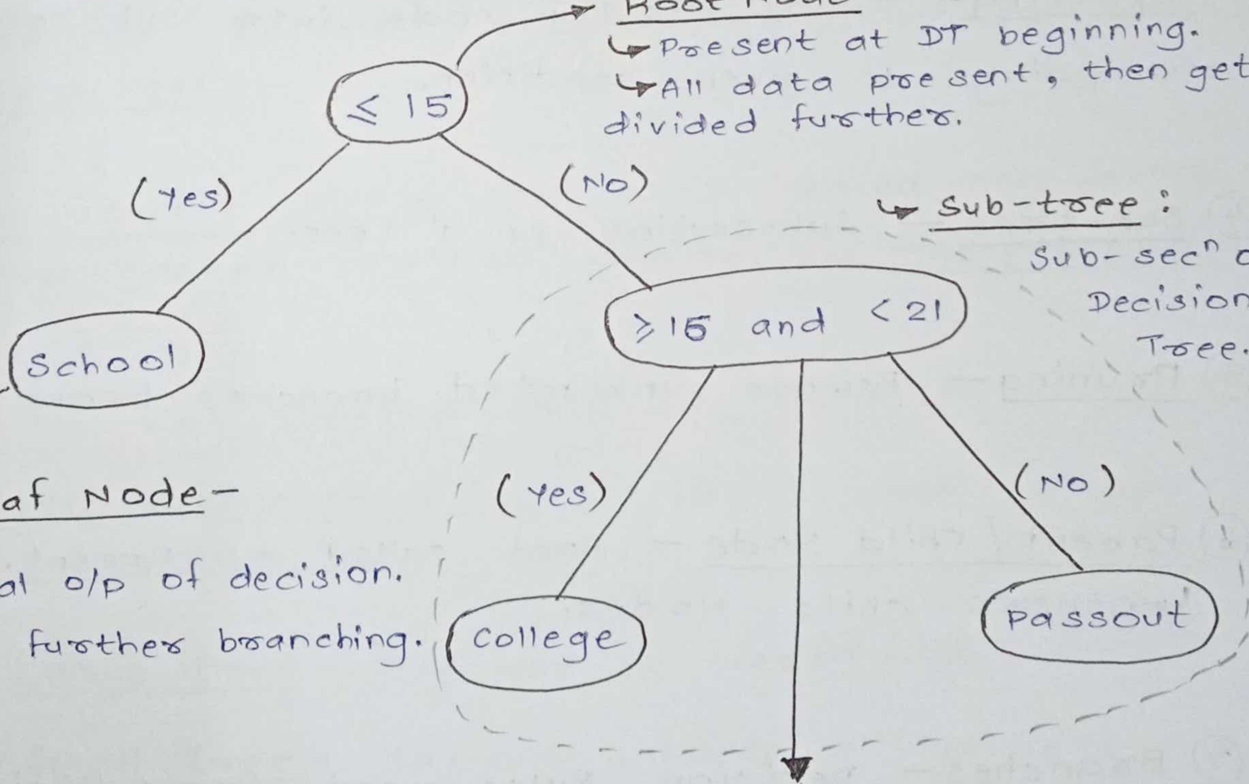
* Decision Tree *

Decision Tree :-

- Supervised Learning technique.
- Used for both Regression & classification, but mostly classification.
- Tree structured classifier.





Root Node-
- Present at DT beginning.
- All data present, then gets divided further.

- **Sub-tree :** Sub-secⁿ of Decision Tree.

Leaf Node-
- Final o/p of decision.
- No further branching.

Decision Node-
- Used to make decisions.
- Test performed on the Feature / attr.
- Make decisions & have branches.

↳ Decision Tree: Graphical representation for getting all possible soln to problem / decision.

↳ CART Algorithm : Classification & Regression Tree Algo. Used to build Decision Tree.

- Decision Tree Terminology :-

① Root Node - Decision Tree starts.
↳ Represent entire data set & divide further.

② Leaf Node - Final o/p Node (class / Label).
↳ No further branching.

③ Splitting - Divide decision node into sub-nodes according to given condition.

④ Sub-tree — Sub-section of a tree.

⑤ Pruning — Remove unwanted branches from tree.

⑥ Parent / child Node — Root called as Parent.
↳ others child Nodes.

⑦ Branches - Decision Rules used for splitting.

- Why use Decision Tree:

① DT mimic human brain for decision-making ability, easy to understand.

② Easy to understand logic, bcz tree-like structure.

- How Decision Tree Algo works:

① Algo start from root, contains complete dataset.

② Find best attr using ASM (Attr selⁿ Measure).

③ Divide dataset, S into sub-sets contain possible values for best attr.

④ Generate Decision Tree node, which contains the best attr.

⑤ Recursively divide the tree using new attr, until we cannot further divide. c/a Final/ leaf node.

- Pruning (Get Optimal Decision Tree):-

↳ Delete unnecessary nodes from tree to get optial Decision Tree.

↳ • Large tree - May lead to overfitting.

• Small Tree - May not capture all Features.

↳ ↓es size of DT without reducing Accuracy.

↳ 2 types of technique for Pruning -

① Cost Complexity Pruning

② Reduced Error Pruning

Attribute Selection Measure (ASM):-

- Used to find best attr for root & sub-nodes.
- 2 Popular Techniques:
  1. Information Gain
  2. Gini Index (Gine Impurity)

  } Terminologies used to define pure/impure split & root node.

(1) Entropy: Measure impurity / randomness of split / node.

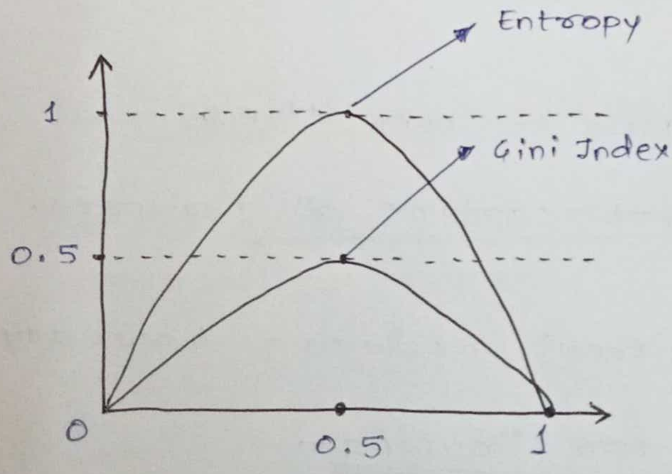
S → set of data points to consider.

$$\boxed{E(S) = -P_1 \log_2 (P_1) - P_2 \log (P_2)}$$

P_1 → Data points in S, belong to label1.

P_2 → Label 2.

- Max Entropy - Data split in many labels / class.
  - i.e. Impure Split (1).
  - Max uncertainty.
- Min Entropy - Data belong to only 1 class / labels.
  - i.e. Pure Split (0).



Entropy range for Binary split is 0 to 1.

② Information Gain:

↳ Measurement of changes in entropy after segmentation of dataset based on attr.

↳ Calculate how much info a feature provides about a class.

↳ According to value of IG, we split the node & build Decision Tree.

↳ DT algo always try maximize IG value & node/attr with highest info gain split first.

$$\boxed{Gain(S, A) = En(S) - \sum_{val(A)} \frac{|S_v|}{|S|} \cdot En(S_v)}$$

③ Gini Index:

↳ Measure impurity of dataset, used when creating Decision Tree in CART algo.

↳ Low Gini Index attr is preferred.

↳ CART algo use Gini Index to create binary splits.

$$\boxed{Gini\ Index = 1 - \sum_v P_v^2}$$

↳ Lower GI – More homogeneous / Pure Distribution.

↳ Higher GI – Heterogeneous / Impure Distribution.

↳ GI faster to compute & more sensitive to probability changes.

- Example Dataset -

Day	Outlook	Temperature	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	NO
2	Sunny	Hot	High	Strong	NO
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	NO
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	NO
9	Sunny	Cool	Nomal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	NO

- Calculate Entropy of Initial Dataset -

$P(Yess) = \frac{9}{14}$ & $P(NO) = \frac{5}{14}$

$$E(S) = -P(Yes)\log_2(P(Yes)) - P(NO)\log_2(P(NO))$$
$$= -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$\boxed{E(S) = 0.940}$$

- Calculate Info Gain of Each Variable –

1) Outlook Feature –

Sunny - 5 (2 Yes, 3 No)   Rain - 5 (3 Yes, 2 No)   Overcast - 4 (4 Yes, 0 No)

$$E(\text{Sunny}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.971$$

$$E(\text{Rain}) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 0.971$$

$$E(\text{Overcast} = -\frac{4}{4}\log_2\left(\frac{4}{4}\right) = 0 = 0$$

$$IG(\text{outlook}) = E(S) - \sum_{val} E(S_{val}) * p(val)$$

$$= 0.940 - \left[\frac{5}{14} * 0.971 + \frac{5}{4} * 0.971 + 0\right]$$

$$= 0.240$$

2) Temperature Feature –

Hot - 4 (2 Yes, 2 No)   Mild - 6 (4 Yes, 2 No)   cool - 4 (3 Yes, 1 No)

$$E(\text{Hot}) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$$

$$E(\text{Mild}) = -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) = 0.91$$

$$E(\text{cool}) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.81$$

$$IG(\text{Temp}) = E(S) - \sum_{val} P(val) * E(S_{val})$$

$$= 0.940 - \left[\frac{4}{14} * 1 + \frac{6}{14} * 0.91 + \frac{4}{14} * 0.81\right]$$

$$= 0.032$$

3) ~~Humidity~~ wind Feature –

strong – 6 (3 Yes, 3 No)

weak – 8 (6 Yes, 2 No)

$$E(\text{strong}) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right) = 1$$

$$E(\text{weak}) = -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right) = 0.81$$

$$IG(\cancel{\text{Humidity}}\ \text{wind}) = E(S) - \sum_{val} P(val) * E(S_{val})$$

$$= 0.940 - \left[\frac{6}{14} * 1 + \frac{8}{14} * 0.81\right]$$

$$= 0.048$$

4) ~~wind~~ Humidity Feature –

High – 7 (3 Yes, 4 No)

Normal – 7 (6 Yes, 1 No)

$$E(\text{High}) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right) = 0.98$$

$$E(\text{Normal}) = -\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right) = 0.59$$

$$IG(\text{Humidity}) = E(S) - \sum_{val} P(val) * E(S_{val})$$

$$= 0.940 - \left[\frac{7}{14} * 0.98 + \frac{7}{14} * 0.59\right]$$

$$= 0.155$$

- Choose Highest Info Gain Feature –
  - ↳ High Info Gain Feature is "Outlook".
  - ↳ Choose this feature for split.

- Repeat : Repeat above steps until it reaches to terminal nodes.

Advantages:

1. Simple - because mimic human brain decision-making
2. Useful for decision-related problems.
3. Help think about all possible outcomes.
4. Less requirement of data cleaning.

Disadvantages:

1. Decision Tree contain lots of layers, complex.
2. May overfitting - solve by Random Forest.
3. More labels $\Rightarrow \uparrow$ computational complexity.