Decision Tree

A Decision Tree is a decision based hierarchical model that uses a tree-like model of decisions and their possible consequences.

Decision Tree Classifier

i/p → o/p

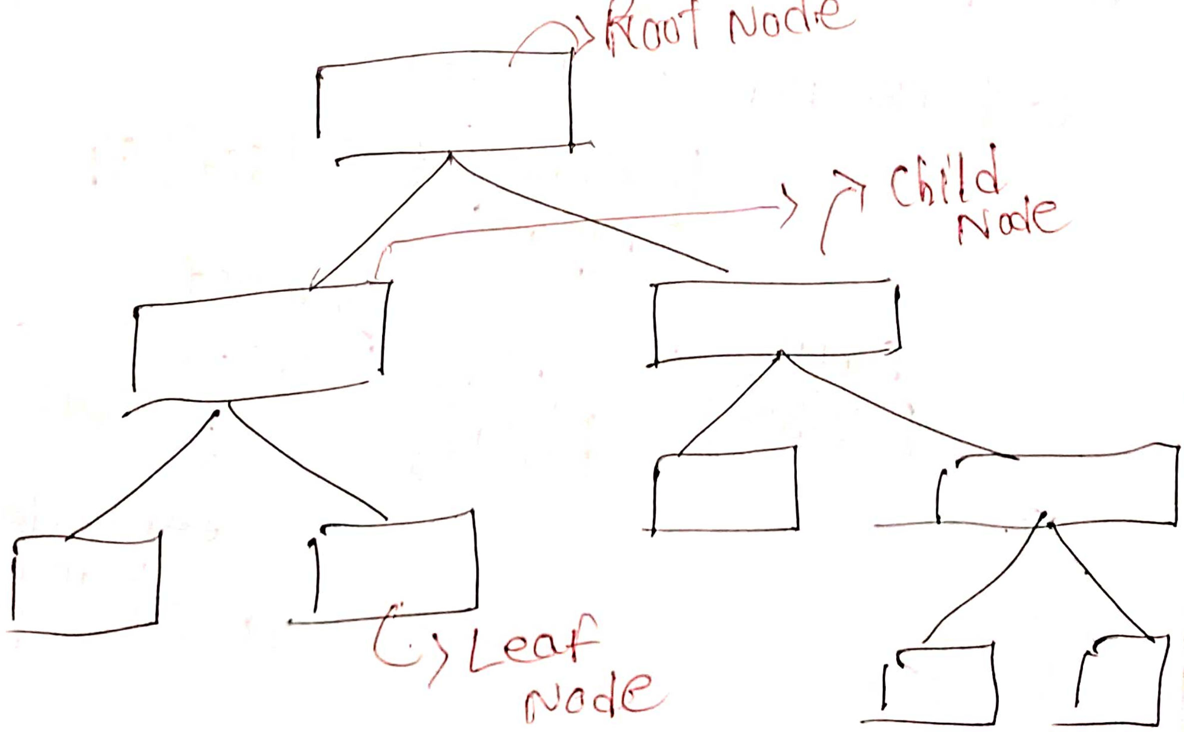
f_1 (Num / Cat) → categorical

Different forms of DTC (Decision tree classifier)

(1) ID_3 . (2) CART

ID_3	CART
(Iterative Dichotomiser 3)	(Classification And Regression Tree)
→ Entropy	→ Gini Impurity.

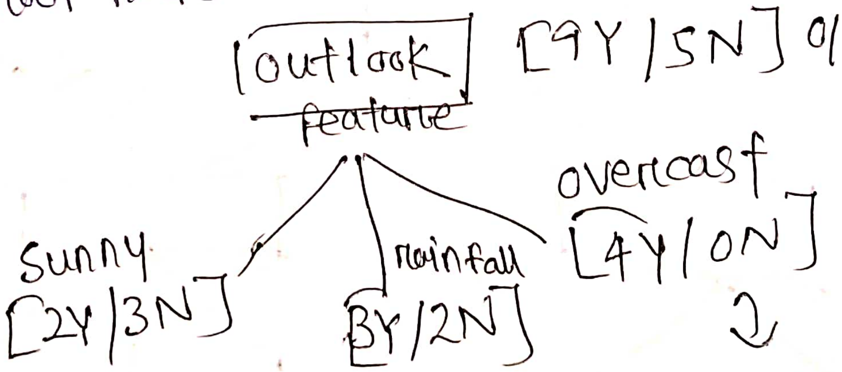
In case of Decision Tree



Day	f1 outlook	f2 Temperature	f3 humidity	f4 wind	Decision
1	sunny	hot	high	weak	NO
2	sunny	hot	high	strong	NO
3	overcast	hot	high	weak	Yes
4	rainfall	mild	~~normal~~ high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	strong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high.	strong	NO

Here
size(o/p) = 14 (9 yes, 5 NO)

Let root node is Outlook.



Leaf Node
because
It is a pure split

Pure split

If the split contain only one output feature its called pure split.

→ To find Impurity of a split / feature we use two method.

① Entropy ② ~~Information Gain~~ Gini Impurity

→ How to choose root node or child node?

Ans. we have to find the Impurity of all feature.

① Entropy

$$\boxed{-\sum_{i=1}^{n} P_i \times \log(P_i)}$$

② ~~Information Gain~~

② Gini Impurity/coeff

$$\boxed{1-\sum_{i=1}^{n} P_i^2}$$

For two class (Y/N)

$$\text{Entropy} = -P_y \log(P_y) - P_N \log(P_N)$$

For 3 class (C_1, C_2, C_3)

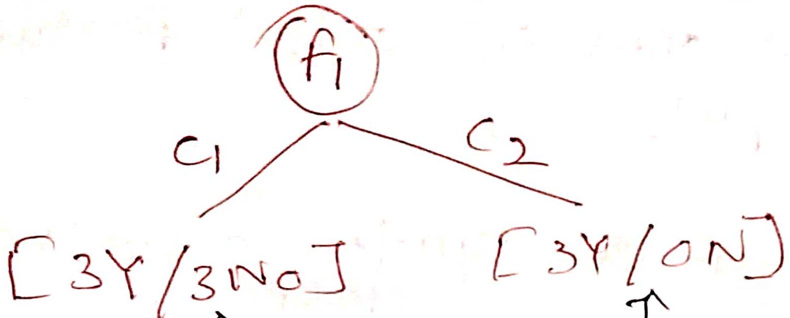
$$\Rightarrow -P_{c_1} \log(P_{c_1}) - P_{c_2} \log(P_{c_2}) - P_{c_3} \log(P_{c_3})$$

Gini Impurity (for 2 class (Y/N))

$$\Rightarrow 1 - [P_Y^2 + P_N^2)$$

Data 1

f_1	o/p
c_1	Y
c_2	Y
c_1	Y
c_2	Y
c_1	Y
c_1	~~Y~~ NO
c_2	Y
c_1	~~Y~~ NO
c_1	NO



50% Y
50% No
Entropy = 1

100% Y
pure split
Entropy = 0

entropy for c_1 $= -\sum_{i=1}^{n} P_i \log(P_i)$

$$= -P_Y \log(P_Y) - P_N \log(P_N)$$

$$= -\frac{3}{6}\log(3/6) - \frac{3}{6}\log(\frac{3}{6})$$

$$= -\frac{1}{2}[\log(1/2) + \log(1/2)]$$

$$= -\frac{1}{2}[\cancel{\log(1)} - \log_2 2 + \cancel{\log(1)} - \log_2 2]$$

$$= -\frac{1}{2}[-2] = 1$$

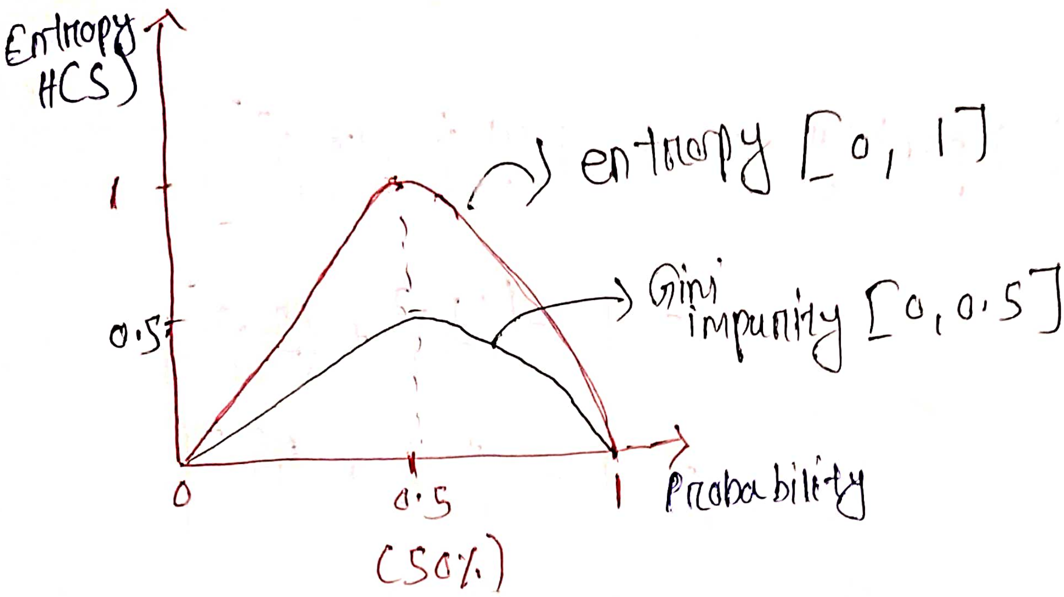
entropy for c_2

$$H(S) = -P_Y \log(P_Y) - P_{No}\log(P_{No})$$

$$= -\frac{3}{3}\log(\frac{3}{3}) - \frac{0}{3}\log(0/3)$$

$$= -1\log(1)$$

$$= 0$$



$H(S) = 1$, very impure split

$H(S) = 0$, pure split.

Let o/p 2 yes / 3 No

$$H(S) = -\frac{2}{5}\log_2(2/5) - \frac{3}{5}\log_2(3/5)$$

$$= 0.97$$

By using Gini-Impurity (For Data)

$$\text{Gini Impurity} = 1 - \sum_{i=1}^{n}(p_i^2) \quad (\text{For } C_1)$$

$$= 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right]$$

$$= 1 - \left[\frac{1}{4} + \frac{1}{4}\right]$$

$$= 1 - \frac{1}{2} = 0.5$$

For C_2

$$\text{gini impurity} = 1 - \left[\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right]$$

$$= 1 - 1 = 0$$

Q. For [4Y / 8N] gini impurity ?

Ans

$$\text{Gini impurity} = 1 - \left[\left(\frac{4}{12}\right)^2 + \left(\frac{8}{12}\right)^2\right]$$

$$= 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right]$$

$$= 1 - \left[\frac{1}{9} + \frac{4}{9}\right]$$

$$= 1 - \frac{5}{9} = \frac{4}{9} = 0.44$$

Q. [8Y / 2N]

$$\text{gini impurity} = 1 - \left[\left(\frac{8}{10}\right)^2 + \left(\frac{2}{10}\right)^2\right]$$

$$= 1 - \left[\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right]$$

$$= 1 - \frac{17}{25}$$

$$= \frac{8}{25} = 0.32$$

→ feature 1, feature - 2, feature - 3
which feature to choose first ?

Ans. for checking impurity

ID3 [Entropy ↓ Info. gain] or [Information Gain ↓ Information Gain] CART

Information Gain

$$\boxed{Gain(S, f_1) = H(S) - \sum \frac{|S_v|}{|S|} H(S_v)}$$

f_1 (c_1, c_2) [9Y, 5N]

c_1 c_2

[6Y | 2N] [3Y / 3N]

$H(S) \Rightarrow$ root feature entropy

$$H(S) = -P_Y \log(P_Y) - P_N \log(P_N)$$

$$= -\left(\frac{9}{14}\right) \log\left(\frac{9}{14}\right) - \frac{5}{14} \log\left(\frac{5}{14}\right)$$

$$= -(0.64) \log(0.64) - 0.35 \log(0.35)$$

$$\approx 0.94$$

for c_1

$$H(S) = -\frac{6}{8} \log(6/8) - \frac{2}{8} \log(2/8)$$

$$= 0.81$$

for c_2

$$H(S) = -\frac{3}{6} \log(3/6) - \frac{3}{6} \log(3/6) = 1$$

$$\text{Gain}(S, f_1) = 0.94 - \Big(\frac{8}{14} \times \overbrace{0.81}^{H(S)(C_1)} + \frac{6}{14} \times \overbrace{1}^{H(S)C_2}\Big)$$

$$= 0.94 - (0.462 + 0.42)$$

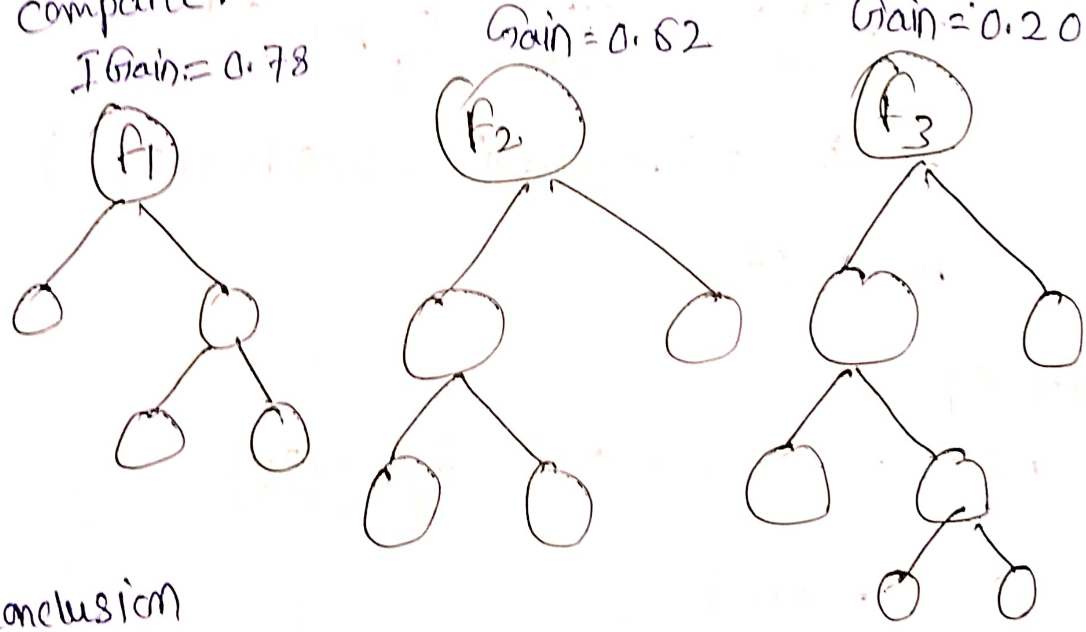
$$= 0.049$$

$|S_v|$ = total no. of sample after spliting

$|S|$ = before splitting total no. of sample.

Similarly we can calculate Info. Gain for all feature. Then we need to compare.

Let



conclusion

-> f1 is best.

we will select the feature which has highest information gain then we split the node based on that feature.

Note

Decision tree is robust to outliers and no scaling of feature is required. because Decision tree works of condition based approach.

Q. when to stop spliting?

In real-world datasets have large no. of features which results in large no. of splits. which in turn gives a large tree. Such trees take time to build and can lead to overfitting. (training accuracy very high and test accuracy very low)

There are many ways to tackle this problem (hyperparameter tuning)

- max_depth (no. of depth split should be)
- min_samples_split (split the node if it contains samples more then min_samplesplit)
- max_feature, min_sample_leaf
- Pruning (cutting the tree)

Pruning → Pre pruning, post-pruning

- More we will discuss in Decision tree regressor