Decision Tree

Types

1. Decision tree classifier —Helps in→ [Classification Problems]
2. Decision tree Regressor —Helps in→ [Regression Problems]

Decision tree → It is a multination if else condition.

Decision tree classifier :-

- ID3 [Iterative Dichotomiser 3]
- CART [Classification and Regression tree]

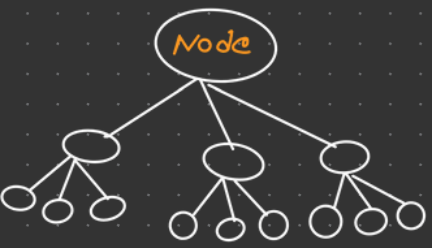
NOTE:- CART in the Sklearn library Solve the classification and Reg. Problems in Decision tree.

- **ID3 :-** When D.T. designed by ID3. We have a Root Node, we Split the Node. split can be binary or multiple split. whenever we split the root node in such a way that it has multiple child Node, more than binary and these further going on. it keeps getting split like this then we can call this technique as ID3 Technique child Node can be 2-3-4,.....,n in ID3 Technique.

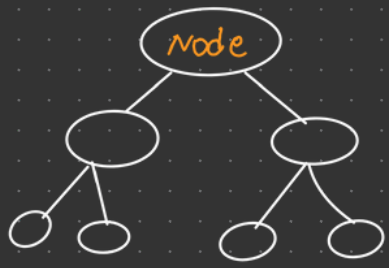
CART :-

CART is called classification and Regression Tree, the specific split done in it will be binary split.

ID3



CART



ex:-

Person, age = 14

if else condition

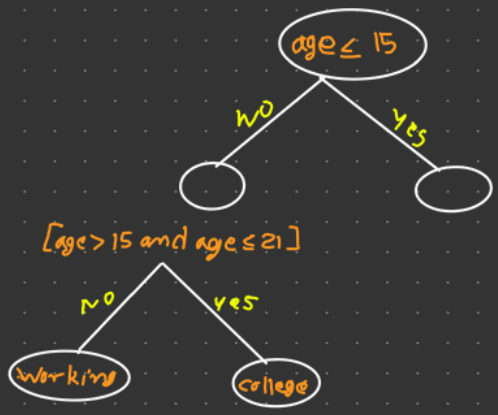
```
if (age ≤ 15)
    print ("School")

elif (age > 15 and age ≤ 21)
    print ("college")

else:
    Print ("Working")
```

Making this nested if else condition into decision Tree

age = 14



ex → if age = 21 → college

if age = 25 → Working

How Decision tree works in the dataset *

Dataset → To Predict play Tennis OR Not. (Goal)

DAY	Outlook	Temp.	Humidity	wind	play Tennis
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	Normal	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

→ We select one feature and after that we create categories in the specific feature present in dataset.

$\boxed{\text{Number of category} = \text{No. of Split}}$

Impure Split :-

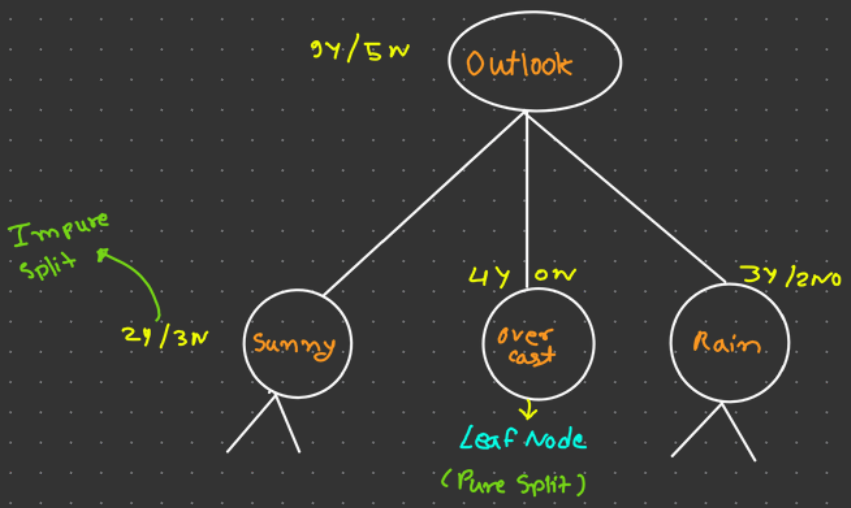
In the split of category when we get Yes OR No both it called Impure split. In case of Sunny we got 3 No and 2 Yes, So it is impure split.

Impure split give access to split again. when we got 50% Yes and 50% NO, then it is also consider Impure Split.

Impure = 1 → in 50% Yes, 50% No

Pure Split :- In the split of category when we got only Yes OR only No. It called Pure split. In the cage of overcast we got Yes and No. so it is a pure Split.

Pure split don't give access to split again, so it is called "Leaf Node".



Entropy Vs Gini Impurity

- whenever dataset is small ⇒ Go with Entropy (1000 – 2000 Records)
- Whenever huge is Huge ⇒ Go with Gini Impurity (Millions, 100K Records)

How to Check Purity In ML *

Purity check

Pure Split OR Impure Split

we use two Techniques

- Entropy
- Gini Impurity

Used for measure of Purity

What feature we need to Select to start the split.?

for this we can use Information Gain Technique.

For Binary Classification

① Entropy :

$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

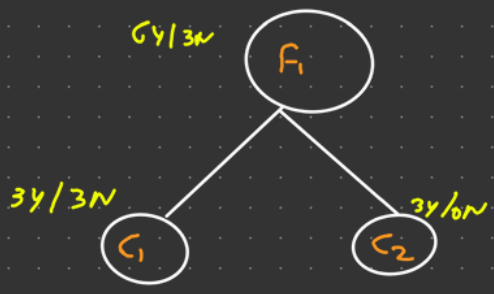
② Gini Impurity

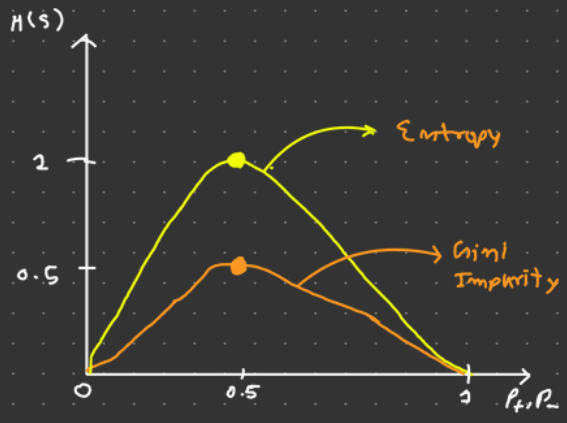
$$G.I. = 1 - \sum_{j=1}^{2} (P)^2$$

Here $P_+ \Rightarrow$ Probability of +ve category

$P_- \Rightarrow$ Probability of -ve category

ex





Entropy

① $-P_+ \log_2 P_+ - P_- \log_2 P_-$

$$H(C_1) = -\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$$

$$\boxed{H(C_1) = 1}$$ → Impure split

② $-P_+ \log_2 P_+ - P_- \log_2 P_-$

$$H(C_2) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

$$H(C_2) = \left(\frac{-3}{0}\right)\log_2\left(\frac{-3}{0}\right) - \left(\frac{0}{3}\right)\log_2\left(\frac{0}{3}\right)$$

$$\boxed{H(C_2) = 0}$$ → Pure split

for Multiclass classification

C_1, C_2, C_3

yes / No / maybe

Entropy

$$\boxed{H(S) = -P_{C_1}\log_2 P_{C_1} - P_{C_2}\log_2 P_{C_2} - P_{C_3}\log_2 P_{C_3}}$$

② Gini Impurity

$$\boxed{G.I. = 1 - \sum_{j=1}^{n} p^2}$$

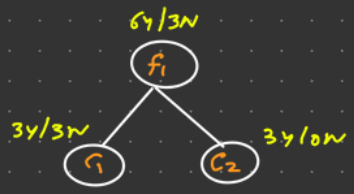
$$GI(C_1) = 1 - [(P_+)^2 + (P_-)^2]$$

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

$$1 - \frac{1}{4} + \frac{1}{4}$$

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

$$\boxed{GI(C_1) = 0.5}$$ → Impure split



$$GI(C_2) = 1 - [(P_+)^2 + (P_-)^2]$$

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

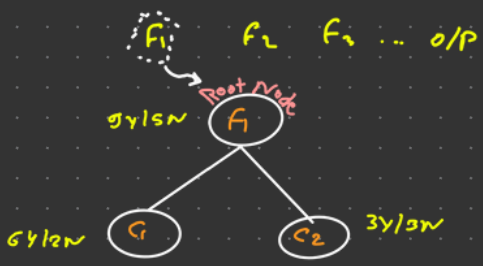
$$1 - 1$$

$$\boxed{GI(C_2) = 0}$$ → Pure Split

Information Gain *

$$\boxed{Gain(S, f_i) = H(S) - \sum_{v \in Val} \frac{|S_v|}{|S|} H(S_v)}$$

ex,



⇒ Entropy of Root Node

➤ $H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$

$= -\left(\frac{9}{14}\right)^2 \log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) \log_2\left(\frac{5}{14}\right)$

$\boxed{H(S) \cong 0.94}$

➤ $H(C_1) = -\left(\frac{6}{8}\right)^2 \log_2\left(\frac{6}{8}\right) - \frac{2}{8} \log \frac{2}{8}$

$\boxed{H(C_1) \cong 0.81}$

➤ $H(C_2) =$ 50% yes and 50% no

$\boxed{H(C_2) = 1}$ → Impure split

$$\boxed{Gain(S, f_1) = H(S) - \sum_{v \in val} \frac{|S_v|}{|S|} H(S_v)}$$

Here

- $|S_v|$ = In the category total output
- $|S|$ = Root Node total output value.
- $H(S_v)$ = Entropy

$$Gain(S, f_1) = 0.94 - \left[\frac{8}{14} \times 0.81 + \frac{6}{14} \times 1\right]$$

$$\boxed{Gain = 0.049}$$

f_1 feature's total Gain = 0.049

Let's

$\Rightarrow$ f_2 features total Gain = 0.051

To choosing what feature to select for split we take a Gain comparision of each feature. The feature with more information Gain, Selected for the spliting.

$$\boxed{Gain(S, f_2) = 0.051} > \boxed{Gain(S, f_1) = 0.049}$$

Information Gain is more when we split Using f_2, So we go with f_2 feature for the Root Node and spliting.

Decision tree for Numerical Split.

→ Sorting the feature : (in assending order value setle)

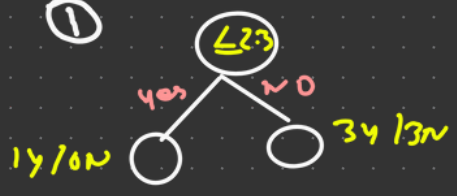
Suppose

f_1	O/P
2.3	yes
3.6	yes
4	NO
5.2	NO
6.7	yes
7.8	NO
9.0	yes

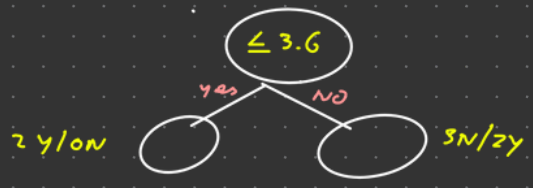
→ Threshold Set

Sorting with Threshold = 2.3

①



② Threshold = 3.6



After that 2.3 Purity check with Entropy then check Information gain, After that next value with same step followed till the end.

③ Threshold = 4

... ④, ⑤, ⑥, .. ⑦

→ Same with all feature

At the end feature selected whose information Gain is Higest.

Disadvantage :-

when we have Huge Dataset (Millons) then ⇒ Time complexbility ↑↑↑