1. Data ingested
2. EDA
3. Processing or FE
4. Model building
5. evalution

Core step

Probability

Naive bayes

Clustering

Neural net

Model building

1. Supervised ML → Regression
2. Unsupervised ML → Classification
   - Binary
   - multiclassification

1. Linear regression => Regression
   - L_1, L_2, L_1+L_2
   - lasso, ridge, elasticnet
2. logistic regression => Classification
3. SVM → regression, classification

Tree based approach (Condition based)

1. DT → DTC, DTR
2. RF
3. AB
4. GBM
5. XBG

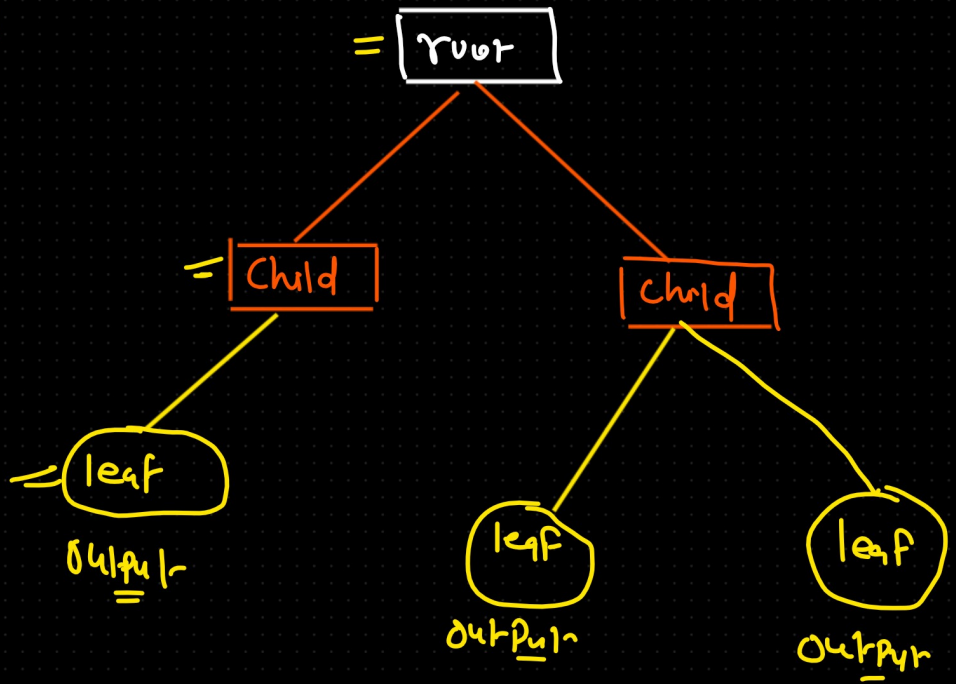
Decisontree
1. DTC
2. DTR

1. DTC
2. Entropy
3. Gini impurity | Gini coff | Gini index
4. IG
5. Pruning → Pre / Post-
6. DTR

ID3, CART, C4.5

```
age = 15        age = 17        age = 23

if (age ≤ 15):
    Print ("school")
elif (age > 15 and age ≤ 21):
    Print ("College")
else:
    Print ("working")
```



Decision-tree
1. ID3 — Iterative Dichotomiser 3 ⇒ C4.5
2. CART — Classification and regression tree

ID3

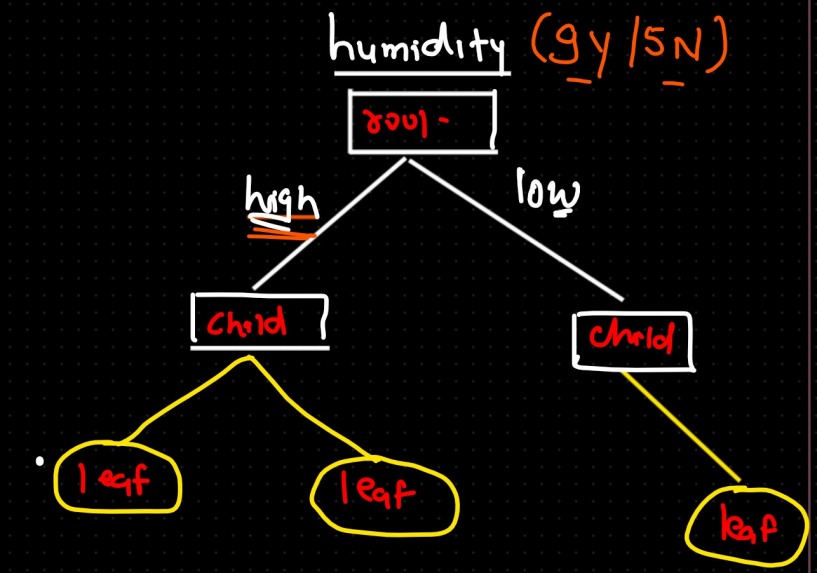
1. Entropy ⇒ Information gain (IG)
2. Categorical (classification)
3. it may divide the root feature more than 2 category

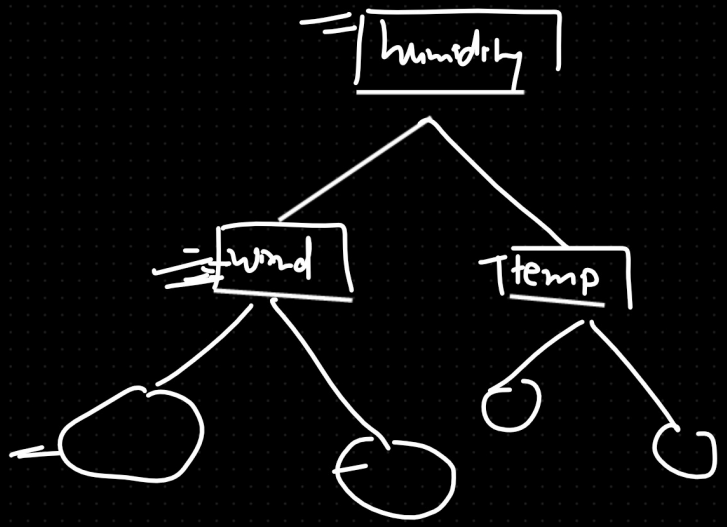
CART

1. Gini impurity
2. classification and regression
3. It always divide root-feature into two category.

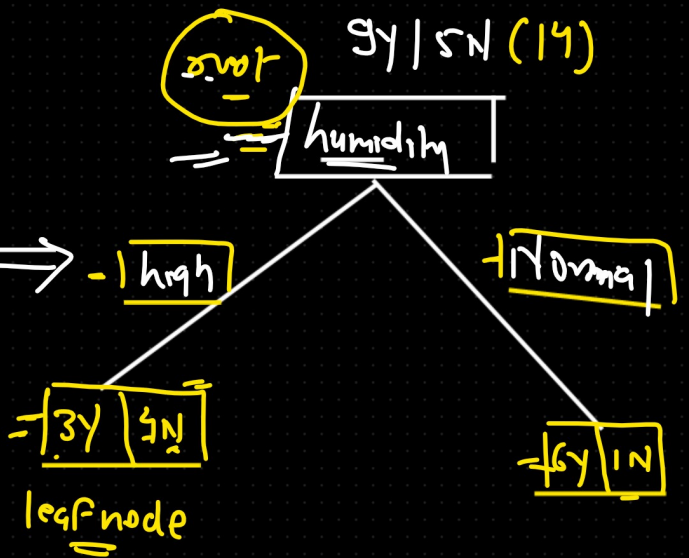
Outlook	Temperature	Humidity	Wind	Played football(yes/no)
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

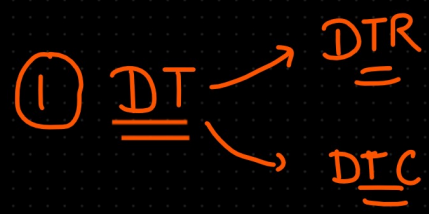
classification

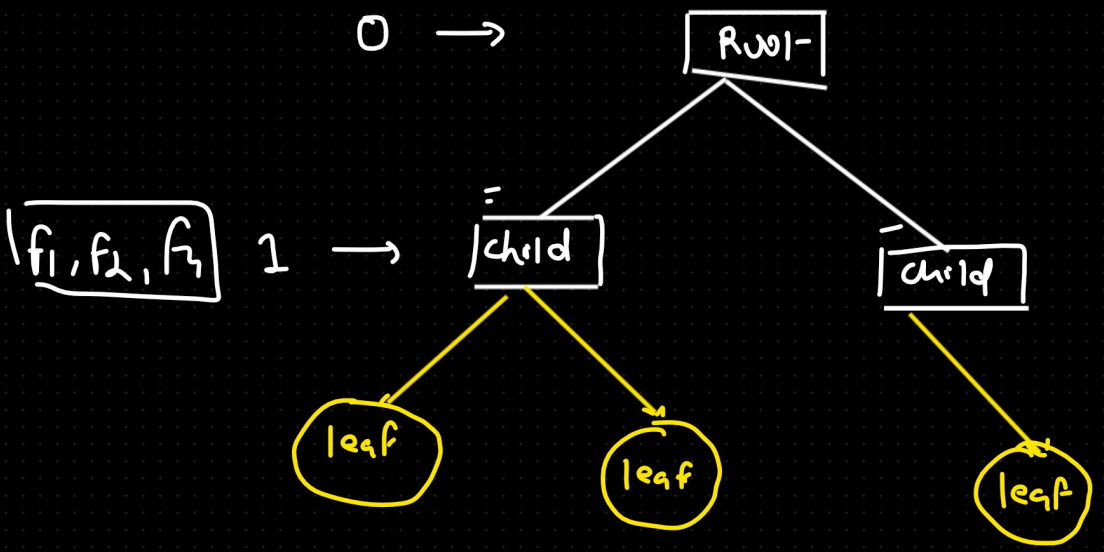




Stage 1 ⇒







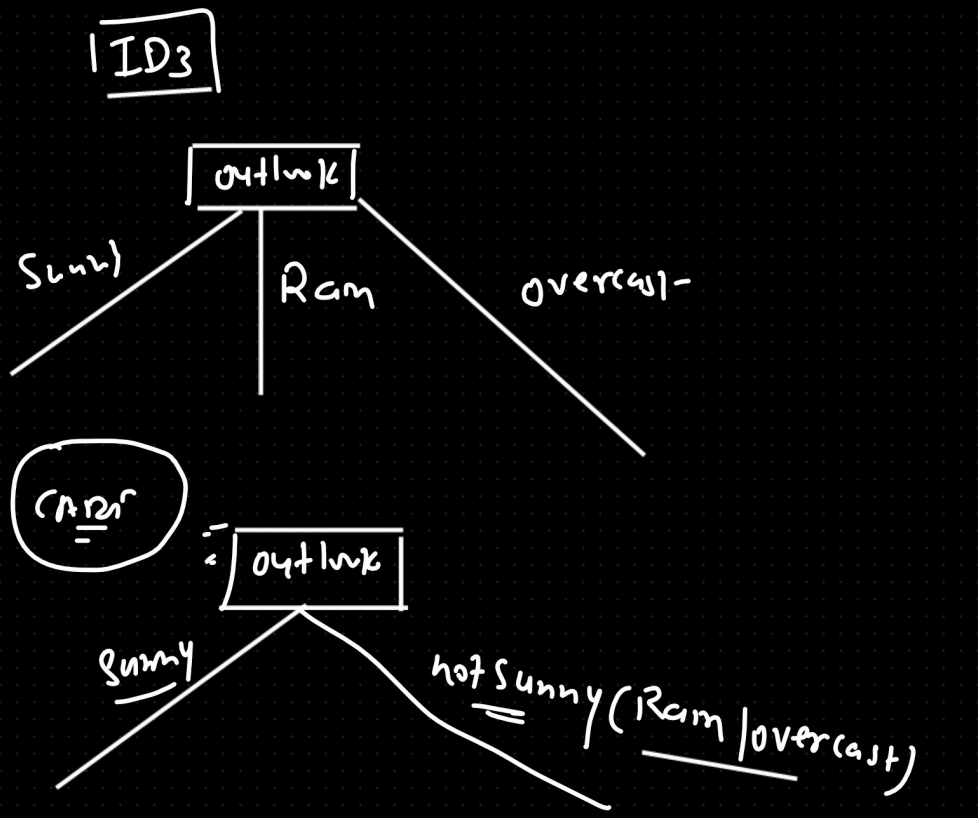
Data

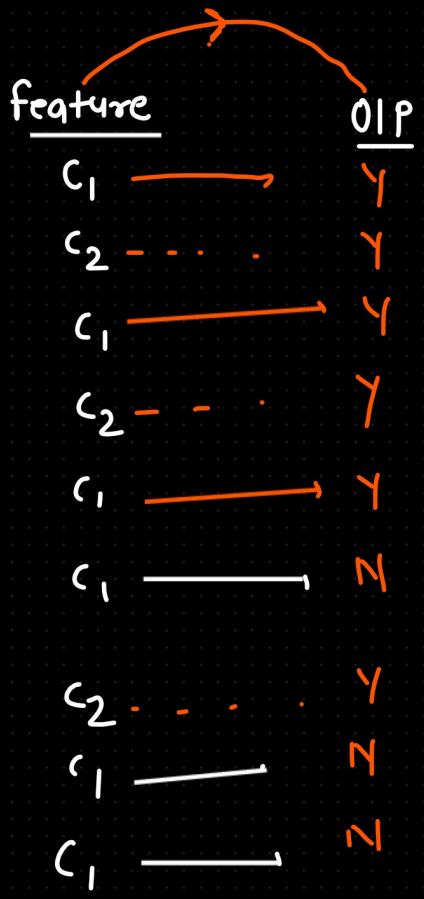
- f1 f2 f3 f4 O/P

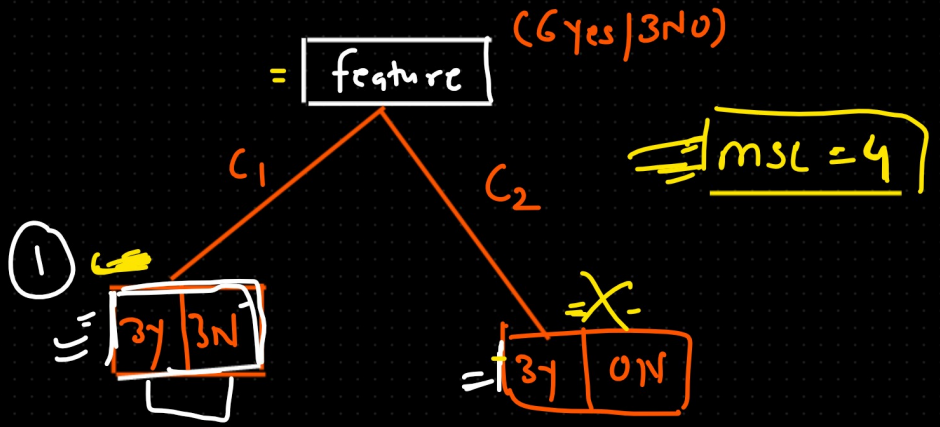
=ID3 ⇒ Entropy → Information gain

=CART ⇒ Gini impurity =

Outlook	Temperature	Humidity	Wind	Played football(yes/no)
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No







ID_3 = entropy

CART = Gini coff

$$\text{ENTROPY} = -\sum_{i=1}^{N} P_i \times \log_2(P_i) \quad (2 \text{ class}) \quad i = 2$$

$$H(S) = -P_Y \times \log_2(P_Y) - P_N \times \log_2(P_N)$$

$\log_e^{!} e = 1$

$$H(S)(3y|3N) = -\frac{3}{6} \times \log_2\left(\frac{3}{6}\right) - \frac{3}{6} \log_2\left(\frac{3}{6}\right)$$

$$= \frac{-1}{2} \times \log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right)$$

$$= \frac{-1}{2}\left[\log_2(1) - \log_2(2)\right] - \frac{1}{2}\left[\log_2(1) - \log_2(2)\right]$$

$$= \frac{-1}{2}[0-1] - \frac{1}{2}[0-1]$$

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

$$H(S) = \frac{1}{2} + \frac{1}{2} = \boxed{1}$$

$(3Y|3N)$

3Y | 3N

Y|N → X (1)

4Y | 2N → (Y) (2)

1Y | 5N → (N) (3)

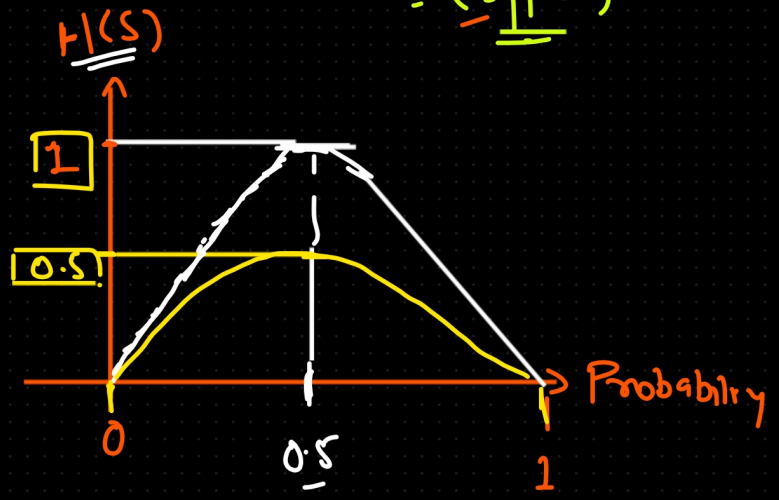
6Y | 0N → (Y) (4)

$$H(S)(3Y|0N) = -\frac{3}{3}\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right)$$

$$H(S) = -\frac{3}{3}\log_2(1) = \boxed{0}$$

$= (3Y|0N)$

feature with less impurity is considered to be the root feature?



(2Y | 3N) $H(S) = ?$ $-\sum_{i=1}^{N} P_i \times \log_2(P_i)$

$2+3=5$ 0.97 → 1

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

$\Rightarrow -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right)$

$\Rightarrow$ 0.97

Gini coff or Gini impurity $\Rightarrow$ $\boxed{1 - \sum_{i=1}^{h} (P_i)^2}$

$\boxed{G.I. \Rightarrow 1 - \left[P_Y^2 + P_N^2\right]}$

① 3Y | 3N ⇒

② 3Y | 0N ⇒

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

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

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

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

$$\Rightarrow 1 - \left[\frac{2}{4}\right]$$

$$\Rightarrow 1 - \frac{1}{2} \Rightarrow \boxed{0.5}$$

$$1 - [(1)^2 + 0]$$

$$1 - 1 = 0$$

Sol: 3Y | 3N

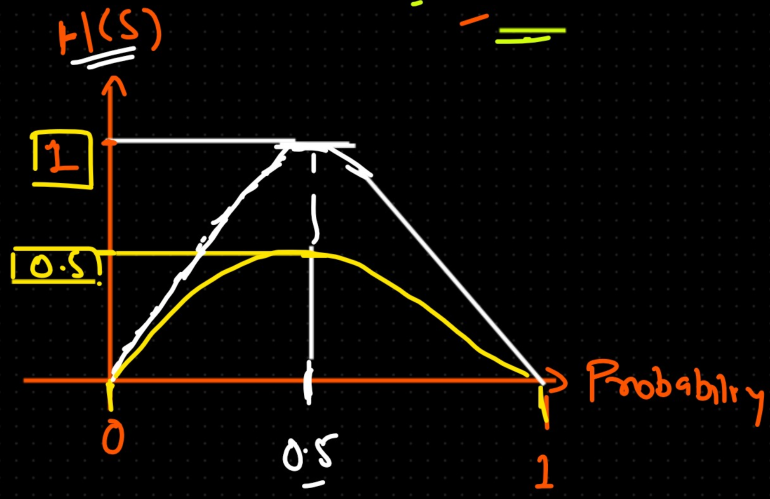
impure → 0.5

entropy → 1

3Y | 0N ⇒ (4) ⇒ Impurity = 0

Gini = 0

entropy = 0



Entropy = 1 — slower

gini-index = 0.5 — faster

Which one should use

ID3 = Entropy

CART = Gini-index

large Dataset → ~~ID3~~ ⇒ Slow

CART

Gini coff

question
(1) 4Y | 8N ⇒ 0.44
(2) 8Y | 2N ⇒ 0.32

$$\Rightarrow \quad 1 - \sum_{i=1}^{n} (P_i)^2$$

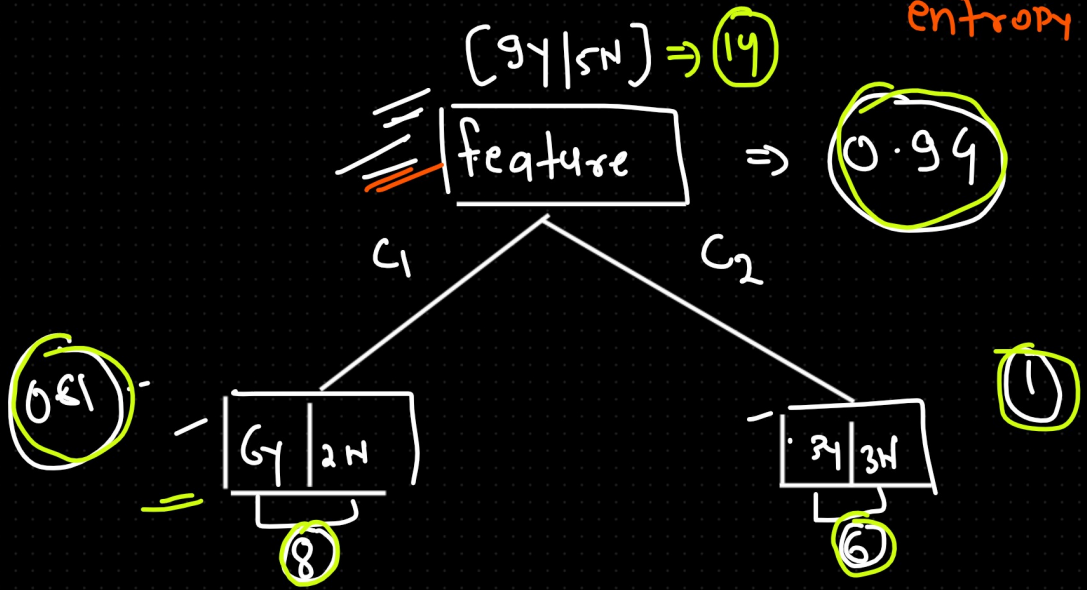
Information gain

$$G.I. \Rightarrow 1 - \left[P_Y^2 + P_N^2 \right]$$

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

root Node entropy

child nod. entropy or entropy after split



$H(S)$ = root feature entropy $= -P_Y \log_2(P_Y) - P_N \log_2(P_N)$

$= -\frac{9}{14}\log_2(9/14) - \frac{5}{14}\log_2(5/14)$

$= -(0.64)\log_2(0.64) - 0.35 \log_2(0.35)$

$H(S)(root) \approx 0.9402$

$6Y|2N = -\frac{6}{8}\log_2(6/8) - \frac{2}{8}\log_2(2/8) = \underline{0.81}$

$8Y|3N = -\frac{3}{6}\log_2(3/6) - 3/6\log_2(3/6) = \underline{1}$

$$\text{gain} \Rightarrow H(S) - \sum_{V \in Val} \frac{|S_v|}{|S|} H(S_v)$$

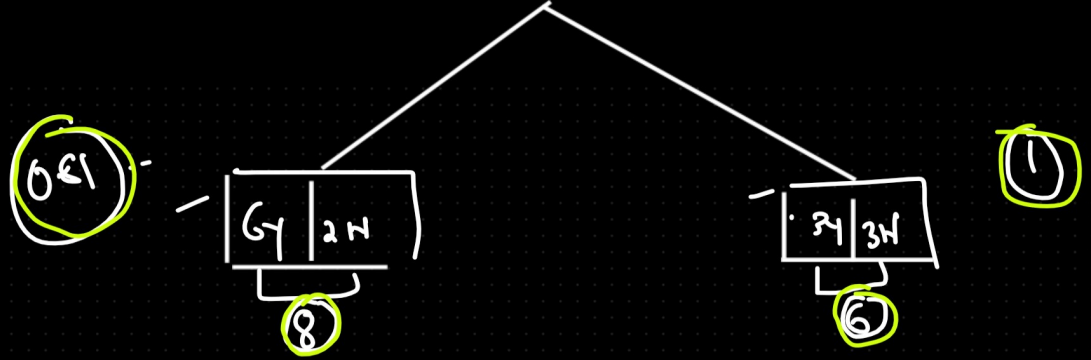
$$\Rightarrow 0.94 - \left[\frac{8}{14} \times 0.81 + \frac{6}{14} \times 1\right]$$

$$\Rightarrow 0.94 - [0.462 + 0.42] = 0.0485 \approx \boxed{0.049}$$

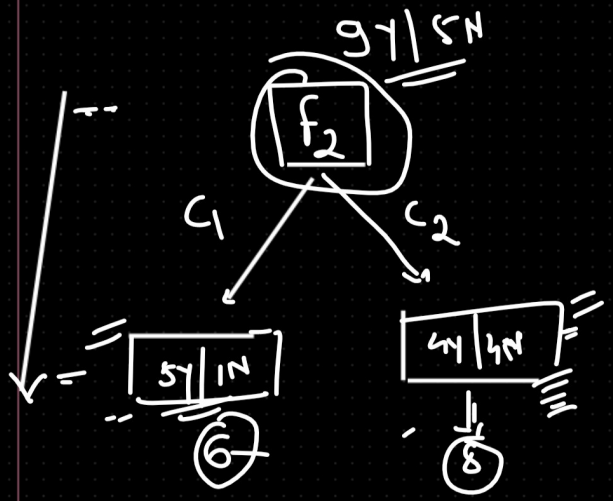
$[9Y|5N] \Rightarrow (14)$

feature $\Rightarrow$ 0.94

C_1 C_2



$$\text{gain} \Rightarrow H(S) - \sum_{v \in \text{val}} \frac{|S_v|}{|S|} H(S_v)$$



$$\Rightarrow \quad 0.94 - \left[\frac{6}{14} \times 0.650 + \frac{8}{14} \times 1 \right]$$

$$\Rightarrow \quad 0.94 - \left(\left[\frac{6 \times (0.650)}{14} + \frac{8}{14} \right] \right)$$

3.9 ± 8

0.94 - 0.85

$f_2 \Rightarrow 0.09$

$f_1 = 0.049$

$f_2 > f_1$ root ?

f_2 will be my root feature = ?

Outlook	Temperature	Humidity	Wind	Played football(yes/no)
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

(1)

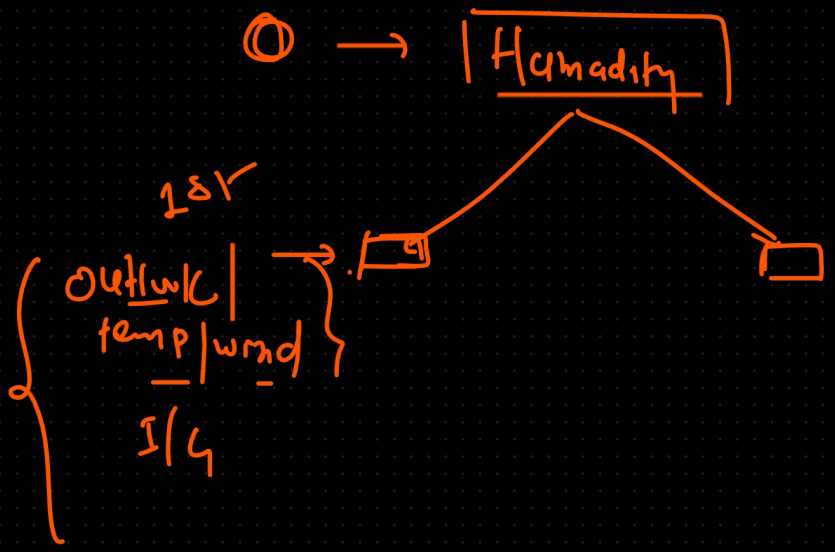
Task

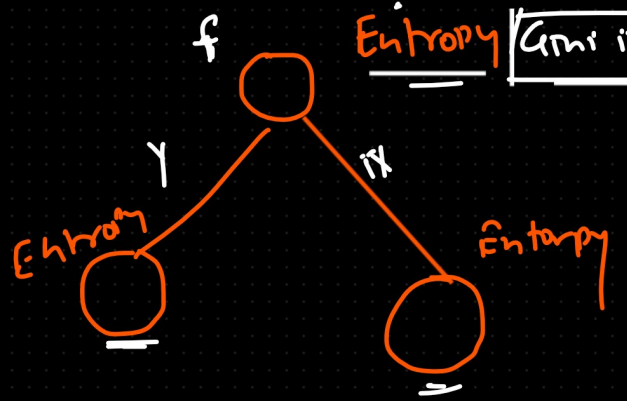
(1) Build the complete DT on top of this Data.

(2) C4.5 algo it's update version of ID3

IG

- Outlook = 0.01
- temp = 0.5
- Humidity = 0.67
- wind = 0.012





Entropy | Gini impurity

ID3

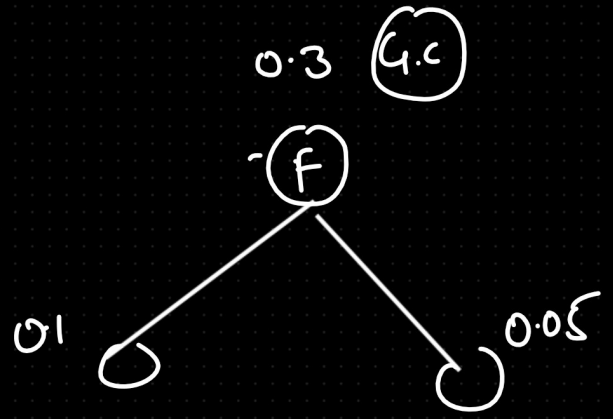
Gain = Root entropy − [Child entropy]

~~Gain = Root-Gini coff − [Gini coff child]~~

Gain = Root-Gini coff − [Gini coff child]

CART | Calculation of all Gini coff

Summation



⇒ 0.3 + 0.1 + 0.05 = 0.45

0.4 + 0.05

G.C

$f_1 = 0.45$

$f_2 = 0.35$

$f_3 = 0.01$ ⟵

less impurity

1. DTC ⇒ implementation
2. DTR –
3. PrePruning | Post-Pruning

1. DTC

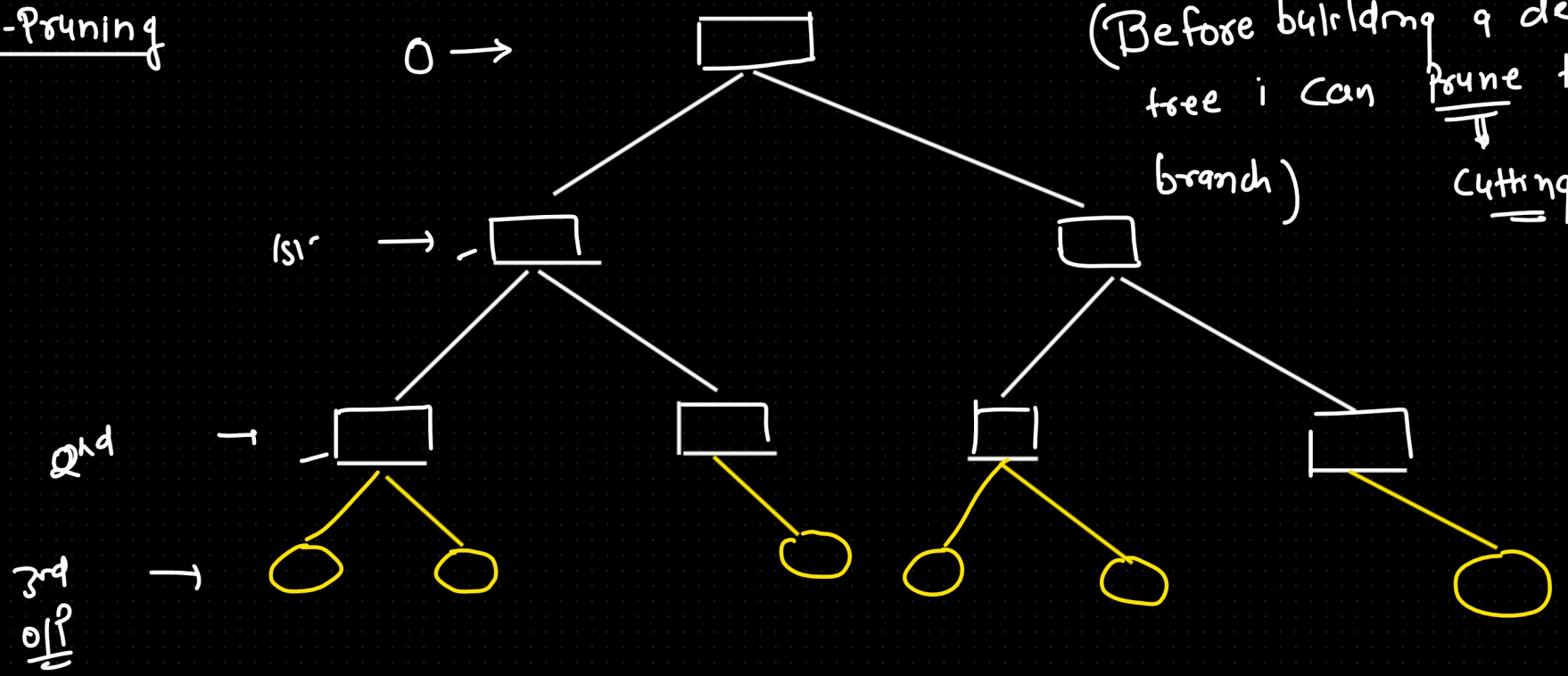
1. ID3 ⇒ Entropy | IG
2. CART ⇒ Gini coeff

Entropy ⇒ formula
Gini formula
I.G. formula

Assignment. ⇒ Build a complete DT on the given Data

Explore C4.5 algo

Pre-Pruning



(Before building a decision tree i can prune the branch)

Pruning ⇒ Cutting

if we going to build DT till complete Depth = ?

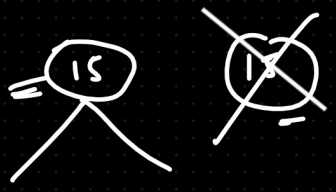
1. Overfitting (train accuracy will be good test accuracy will be bad)
2. Computational expensive

Preprune ⇒

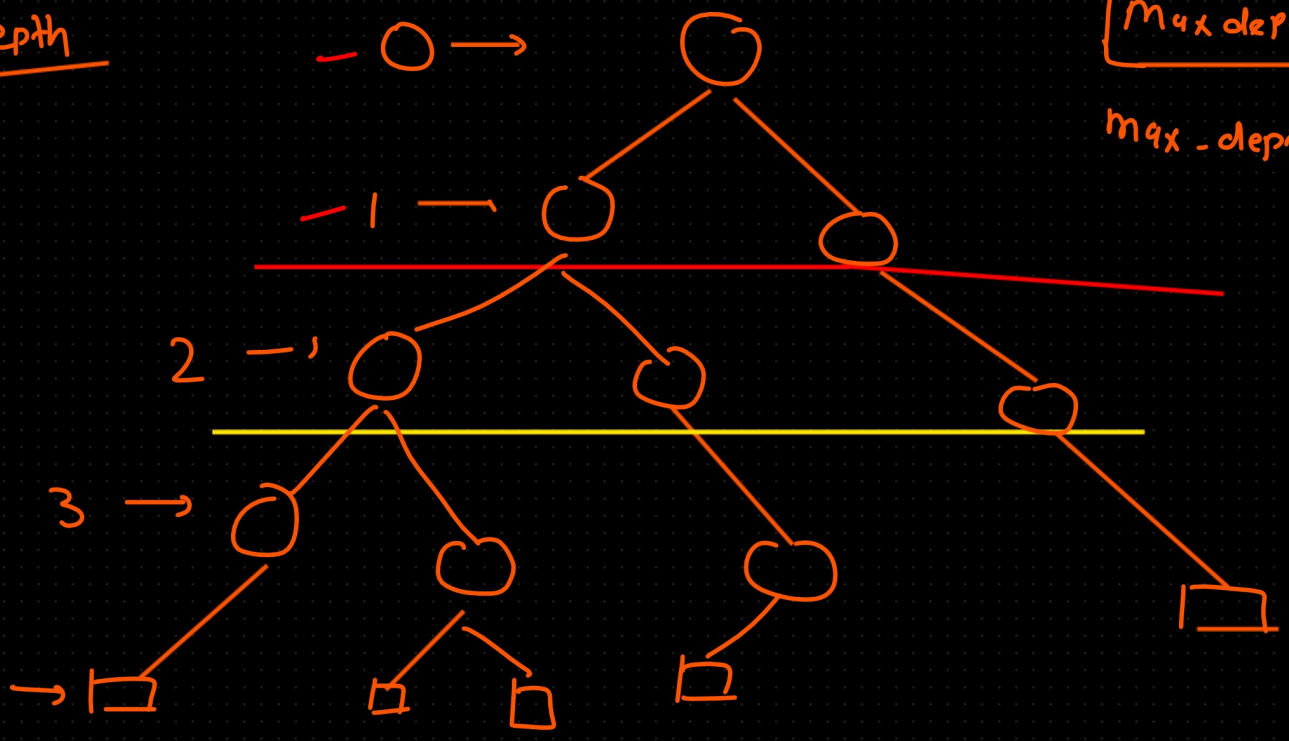
1. Max_Depth
2. Minimum Sample leaf
3. minimum Sample split
4. max_feature

MSS = 10

mss = 18



① Max-Depth



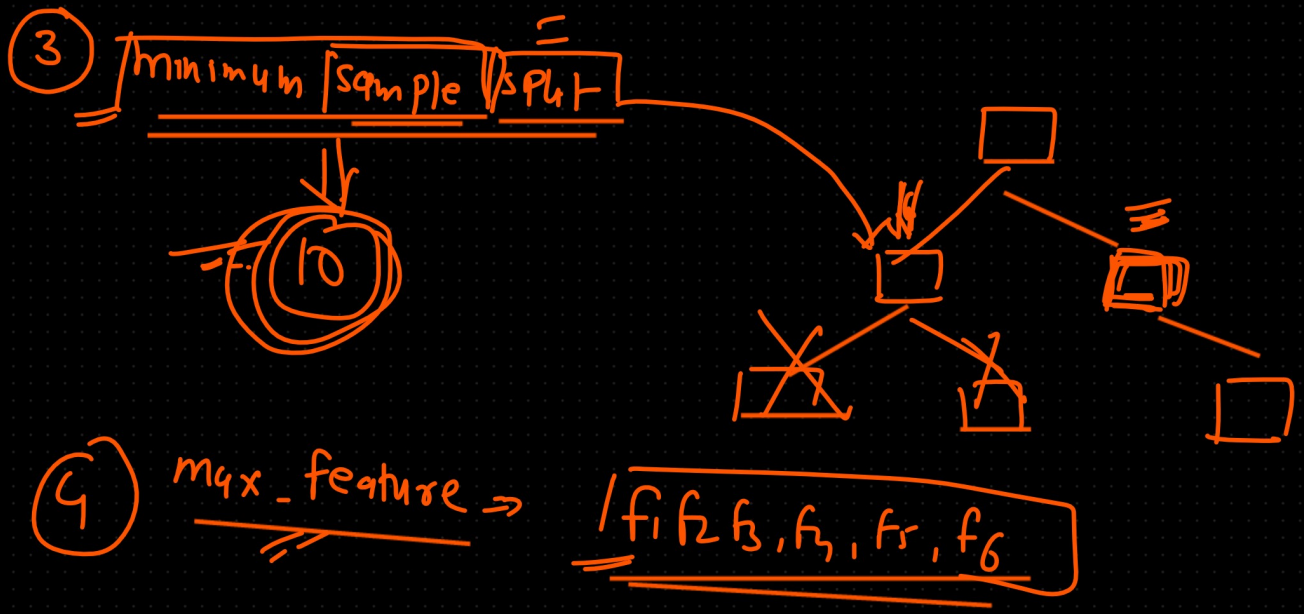
max depth = 2

max_depth = 3

② minimum sample-leaf

Constraint-

msl = 14


3
minimum sample split
10
4
max_feature ⇒
f1 f2 f3, f4, f5, f6