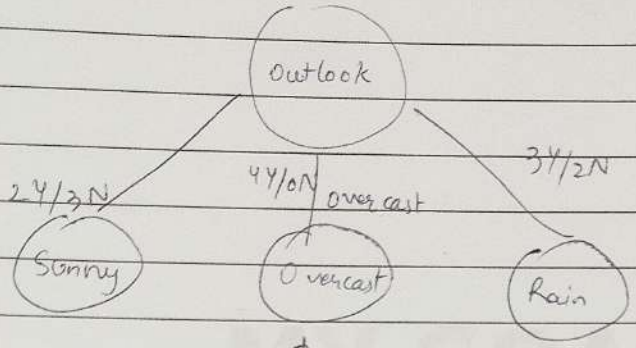
RANKA
DATE / /
PAGE

DECISION TREE



↑
Pure leaf Node. ⟶ Either with 0N/0Y.

(1) How can we say if it is a pure split →
- Entropy
- Gini Coeff.

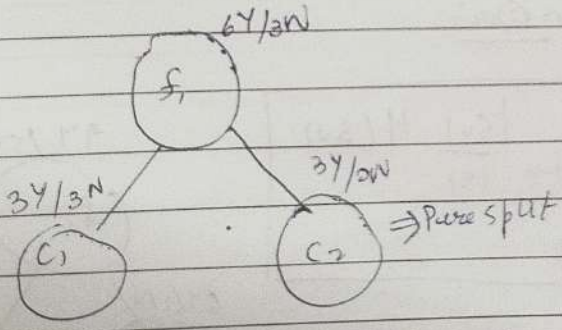
(2) How features are selected
- Information Gain

ENTROPY

$$H(S) = -\sum_{i=1}^{n} P_i \log_2 P_i \rightarrow H_S = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

GINI IMPURITY

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



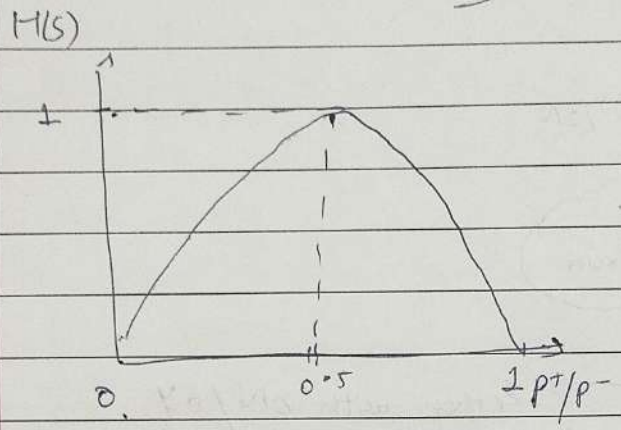
Entropy right node $H(S) = -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}$

$= 0$

Entropy of left node = $-0.5 \log_2 0.5 \; -0.5 \log_2 0.5$

$= -1 \log_2 0.5$

$= (1)$

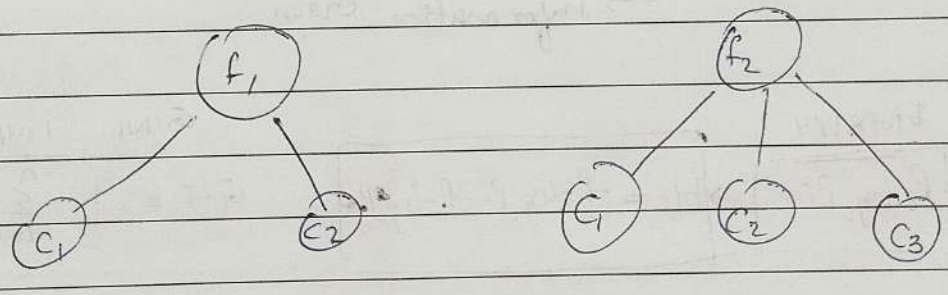


* Entropy B/w [0-1]
* Using Purity Test,

If $H(S) = 0 \rightarrow$ Pure split

$H(S) = 1 \rightarrow$ Impure split

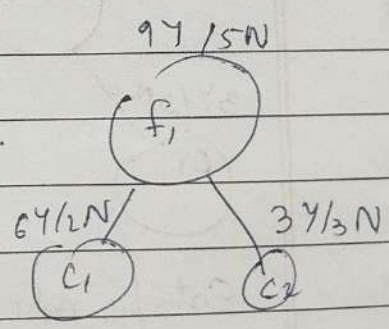
(2) which feature to take to split?



→ Here we use Information Gain.

$$\text{Gain}(S, f_1) = H_S - \sum_{v \in \text{value}} \frac{|S_v|}{|S|} H(S_v)$$

S → Sample, f_1 → feature 1, v → belong



$H(S) \rightarrow$ Root node $= -P_+ \log_2 P_+ - P_- \log_2 P_-$

$= -\frac{9}{14} \log_2 \frac{9}{14} \; -\frac{5}{14} \log_2 \left(\frac{5}{14}\right) = 0.94$

$H_{SV}(C)_1 = -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{6}{8}\right) = 0.81$

$H_{SV}(C_2) = -\frac{3}{6}\log_2(0.5) - \frac{3}{6}\left(\log_2(0.5)\right) = 1$

$\text{Gain}(S, F_1) = 0.94 - \left[\frac{8}{14} \times 0.81 + \frac{6}{14} \times 1\right]$

$= 0.94 - \left[0.574 + 0.4628 + 0.4285\right)$

$= 0.94 - \left[0.8913\right) = 0.049$

$\text{Gain}(S, F_2) = 0.051$ → Using which feature should I start splitting first.

$\text{Gain}(S, F_2) >> \text{Gain}(S, F_1)$

⇒ We will use feature 2 (first) to split data.

GINI IMPURITY

$G.I = 1 - \sum_{i=1}^{n}(P)^2$ → Gini value $[0 - 0.5]$

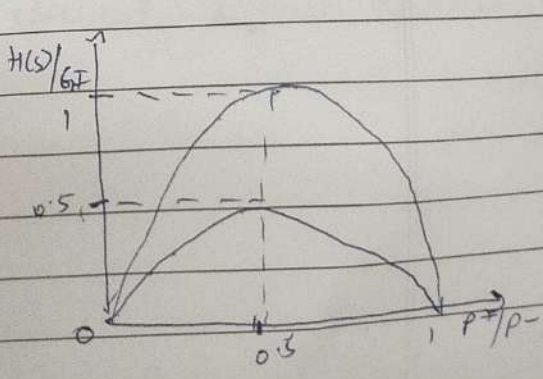
$n = 2$ output { Yes / No

$G.I = 1 - \left[(P_+)^2 + (P_-)^2\right]$

Eg:- 24/2N

$G.I \Rightarrow 1 - \left[(0.5)^2 + (0.5)^2\right]$ → Impure split

$= 0.5$



RANKA
DATE / /
PAGE

Decision Tree Regressor

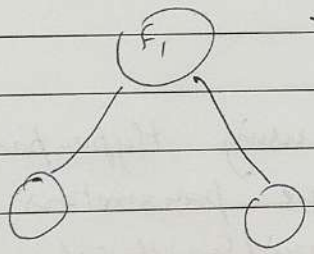
f_1	f_2	O/P
		Continuous
		-
		-
		-
		-
		-

⇒ Calculate mean of O/p. & we will not check Entropy / Gini coeff, ~~Also~~ MSE or MAE will get ~~selected~~ calculated.

↓

$$\frac{1}{2m}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

⇒ Suppose we select f_1



→ Calculate MSE or MAE
→ When we split Node certain records will go to nodes.
→ O/p of nodes will be average or mean.
→ Then again node will split, records will go & mean will be calculated.
→ Along the way MSE will get reduced & will approach towards leaf node or 0.

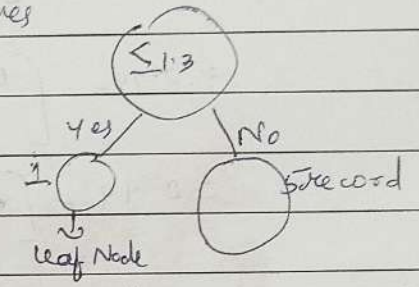
⇒ Because there is log fn in case of Entropy, Gini Impurity is fast.

for Continuous Variables

f_1	o/p
2.3	
1.3	
4	
5	
7	
3	

o/p ⇒ f_1 → Sort Values

1.3
2.3
3
4
5
7

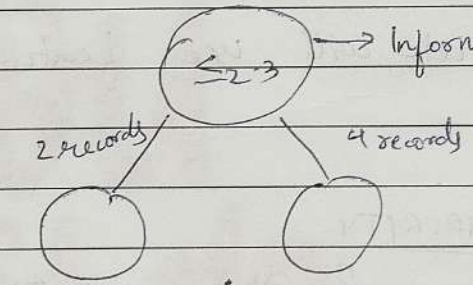


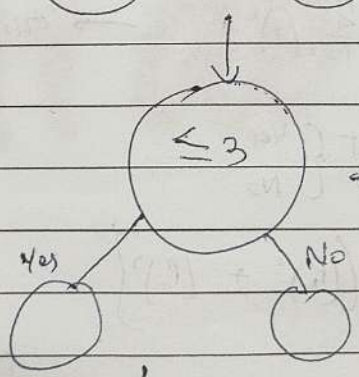
↓

Calculate Information gain of this

↓

After create D.T for 1 2 3





↓

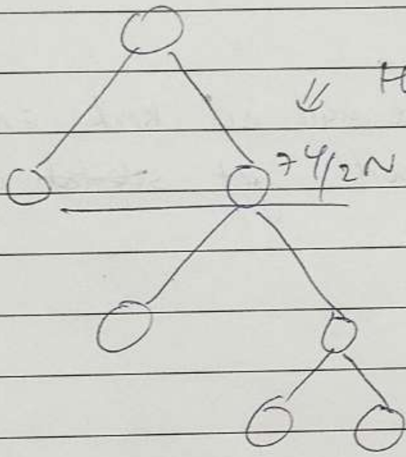
After this it will check which I.G is larger & then it will get selected.

Hyperparameters

→ Decision tree leads to overfitting
→ To prevent overfitting we will do -

1. Post Pruning
2. Pre Pruning

① Post Pruning



⇙ Here More than 80% Bos is there to get Yes. we will cut the tree.

② Pre Pruning - It is done using Hyperparameters
Max depth, max-leaf → These parameters can be ~~done~~ set by GridSearch CV.