Decision Tree → classification & Regression

Healthcare (Recall)

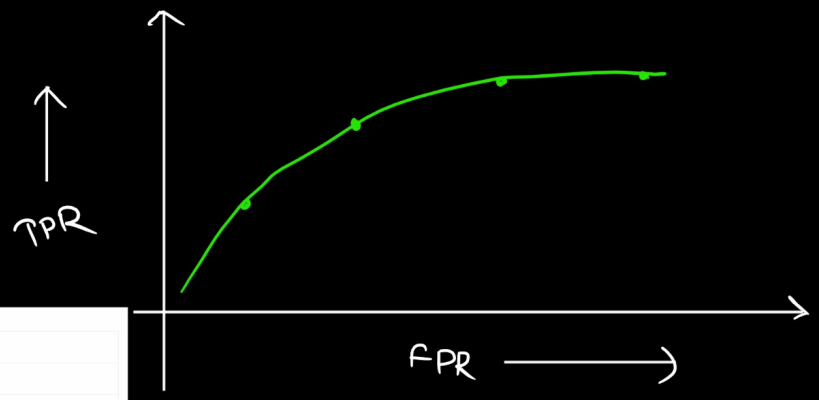
threshold → 0.2 (scenario) → FN Reduce → Roc - AUC

threshold → 0.8 (scenario) → FP Reduce → $FPR = \frac{FP}{FP+TN}$

$TPR = \frac{TP}{TP+FN}$

widely asked interview Question

	Predicted Spam (1)	Predicted Not Spam (0)
Actual Spam (1)	40 (True Positives)	10 (False Negatives)
Actual Not Spam (0)	5 (False Positives)	45 (True Negatives)



Threshold	TPR (Recall)	FPR
0.2	0.95	0.40
0.4	0.90	0.20
0.6	0.85	0.10
0.8	0.75	0.05
1.0	0.00	0.00

CM

0.2 > → 1

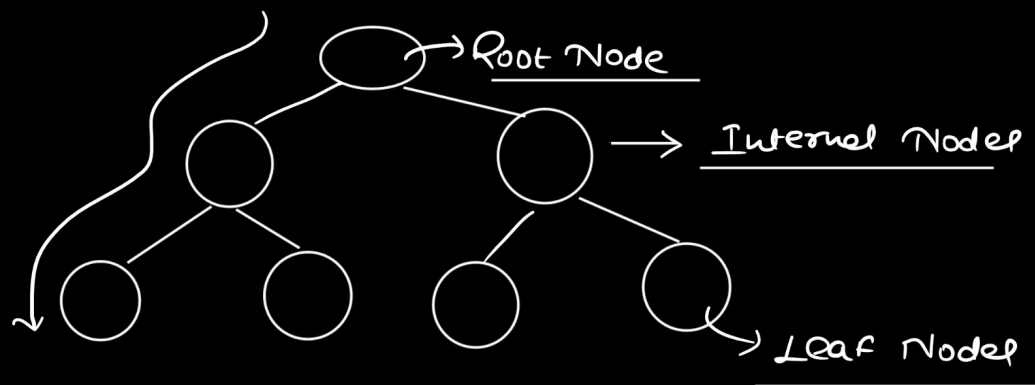
0.2 ≤ → 0

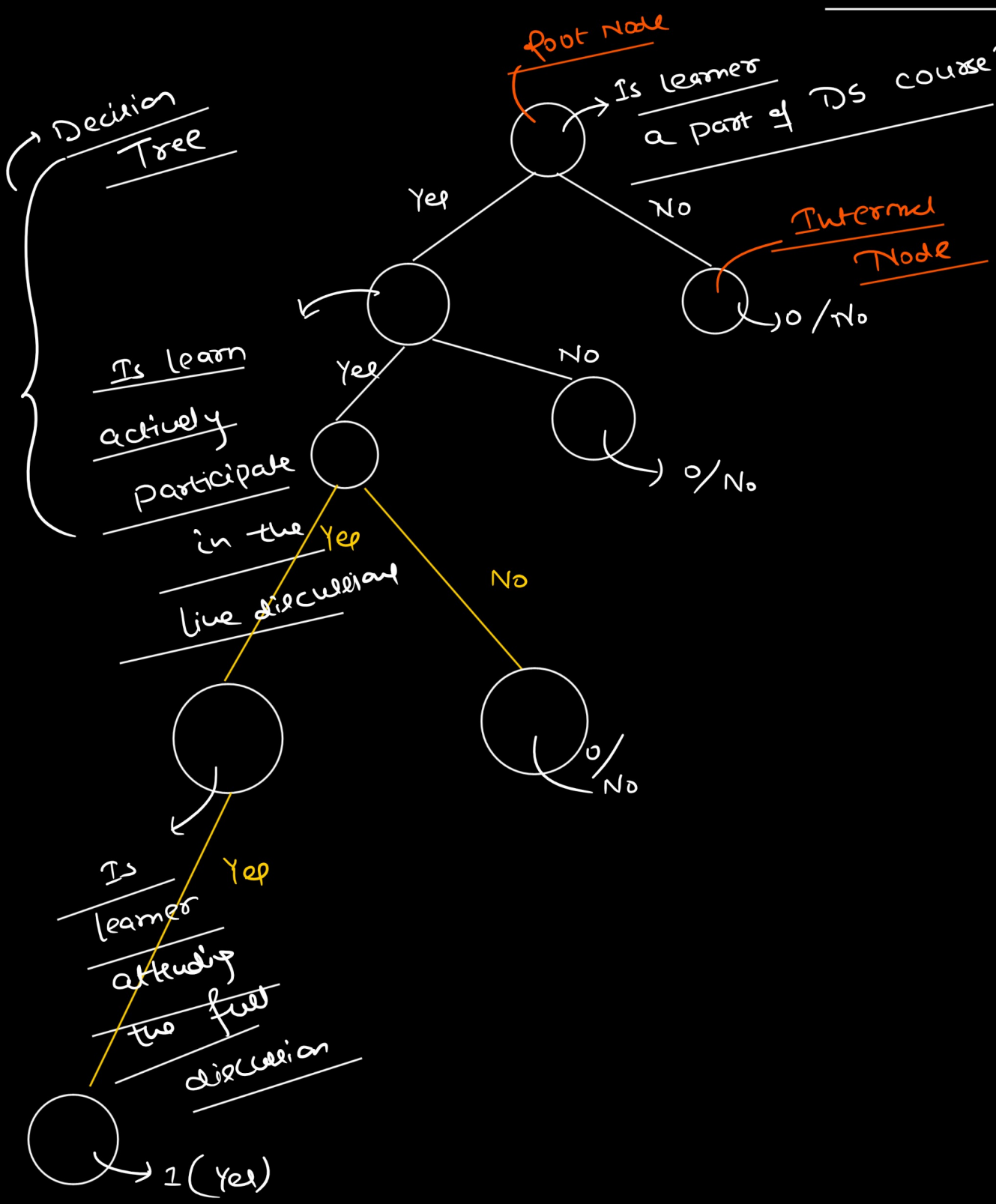
Prob

0.5 > → 1

0.5 ≤ → 0

Tree Data Structure (Non-Linear)





Gini Index & Entropy

Impurity measures

$$Gini = 1 - \sum_{i=1}^{c} p_i^2$$

f1 f2 target

Class	Gender	Stay in hostel
9	M	Yes
10	F	No
8	F	Yes
8	F	No
9	M	Yes
10	M	No
11	F	Yes
11	M	Yes
8	F	Yes
9	M	No
11	M	No
11	M	Yes
10	F	No
10	M	Yes

Gini(class) = 0.404

Gini(Gender) = 0.482

→ weighted avg

$$Gini(class) = \frac{3}{14} * 4/9 + \frac{3}{14} * 4/9 + \frac{4}{14} * 3/8 + \frac{4}{14} * 3/8$$

$$\Rightarrow 0.404$$

class	Stay in hostel	P(Y)	P(N)	Gini
8	Y=2 N=1	2/3	1/3	$1-((2/3)^2+(1/3)^2) = 4/9$
9	Y=2 N=1	2/3	1/3	4/9
10	Y=1 N=3	1/4	3/4	$1-((1/4)^2+(3/4)^2) = 3/8$
11	Y=3 N=1	3/4	1/4	3/8

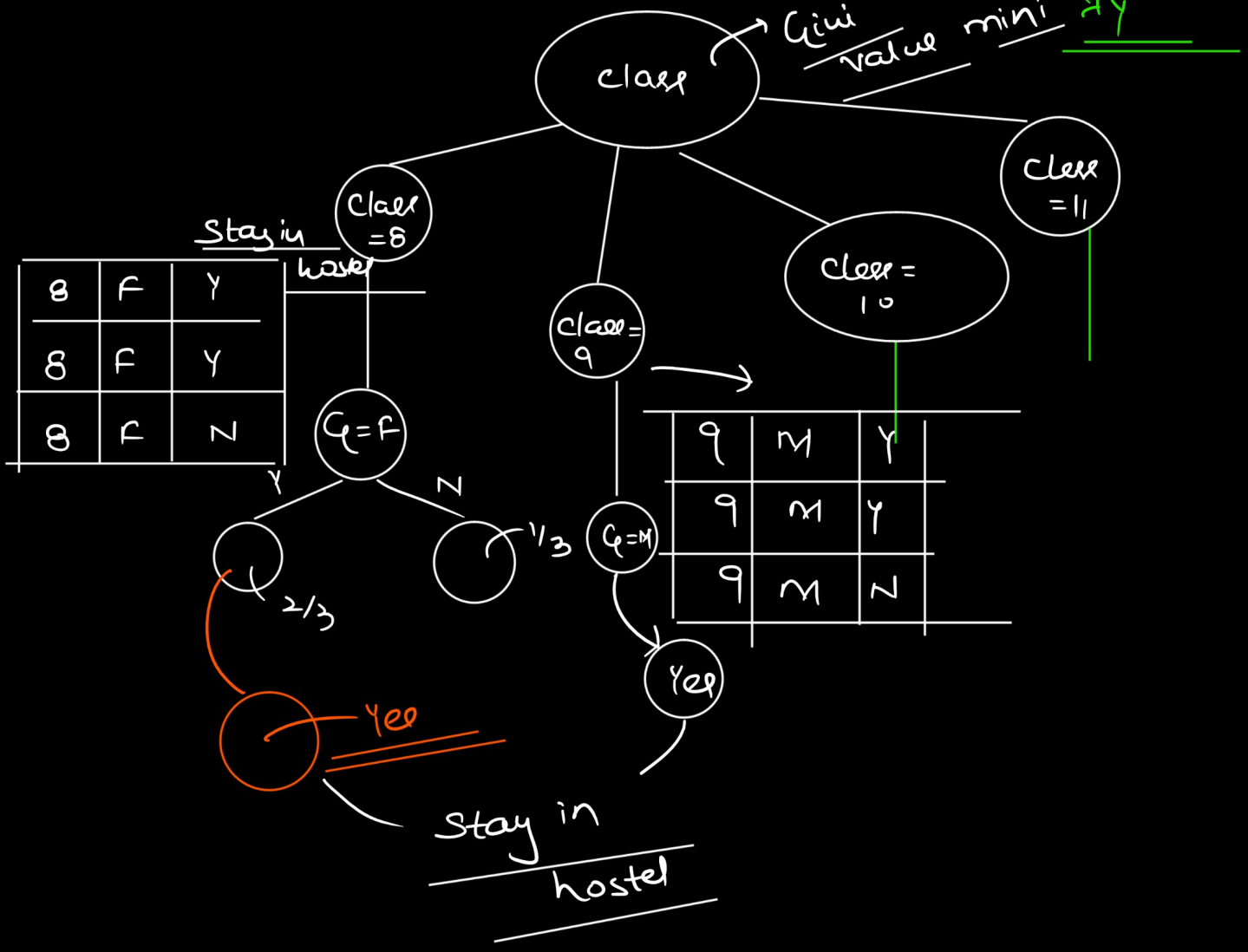
Gender	Stay in hostel		P(Y)	P(N)	Gini
M	Y = 5	N = 3	5/8	3/8	$1 - \left(\left(\frac{5}{8}\right)^2 + \left(\frac{3}{8}\right)^2\right) = 0.468$
f	Y = 3	N = 3	3/6	3/6	$1 - \left(\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right) = 0.5$

$$\text{Gini(Gender)} = \frac{8}{14} * 0.468 + \frac{6}{14} * 0.5$$

$$= 0.482$$

9 M ⟶ Yes

8 f ⟶ Y/N → Y



Gini Index

Interpretation

↳ [0, 0.5]

Pure	Impure
$1 - ((0)^2 + (1)^2)$	$1 - ((1/4)^2 + (3/4)^2)$
$\Rightarrow 0$	$\Rightarrow 0.5$
↳ Highest level of purity	↳ Highest level of impurity

Highest level of purity → 1

only diabetic patients or only non-diabetic patients

↳ 0

5 Records → Class 0 (5 samples) → $P(\text{Neg}) = 5/5 = 1$

Class 1 → 0 samples

$P(\text{Pos}) = 0$

4 Records (Impure sample)

Pos → 50% → 2 samples → 2/4

Neg → 50% → 2 Samples → 2/4

Note →

① 0 means perfectly Pure

② Closer to 1 means very impure

By default, Gini is the criterion

↳ Computationally, it's very easy

Entropy $\Rightarrow -\sum_{i=1}^{c} p_i \log_2 p_i$

↳ Information Gain

Class	Gender	Stay in hostel
9	M	Yes
10	F	No
8	F	Yes
8	F	No
9	M	Yes
10	M	No
11	F	Yes
11	M	Yes
8	F	Yes
9	M	No
11	M	No
11	M	Yes
10	F	No
10	M	Yes

Entropy (class) = 0.857 (More Purity)

Entropy (Gender) = 0.974

Entropy (target = stay in hostel) ⇒

$$-\frac{8}{14}\log\frac{8}{14} - \frac{6}{14}\log\frac{6}{14}$$

$\Rightarrow 0.985$

IG (class) ⇒ Entropy (target) − Entropy (class) ⇒ 0.985 − 0.857 = 0.128

IG (Gender) = 0.985 − 0.974 = 0.011

class	Stay in hostel		P(Y)	P(N)	Entropy
8	Y = 2	N = 1	2/3	1/3	$-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$
9	Y = 2	N = 1	2/3	1/3	0.918
10	Y = 1	N = 3	1/4	3/4	$-\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = 0.811$
11	Y = 3	N = 1	3/4	1/4	0.811

$$\text{Entropy(class)} = \frac{3}{14} * 0.918 + \frac{3}{14} * 0.918 + \frac{4}{14} * 0.811 +$$

$$\frac{4}{14} * 0.811 = \underline{\underline{0.857}}$$

Gender	Stay in hostel		P(Y)	P(N)	Entropy
M	Y = 5	N = 3	5/8	3/8	$-\frac{5}{8}\log\frac{5}{8} - \frac{3}{8}\log\frac{3}{8} = 0.954$
F	Y = 3	N = 3	3/6	3/6	$-\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6} = 1$

$$\text{Entropy(Gender)} = \frac{8}{14} * 0.954 + \frac{6}{14} * 1$$

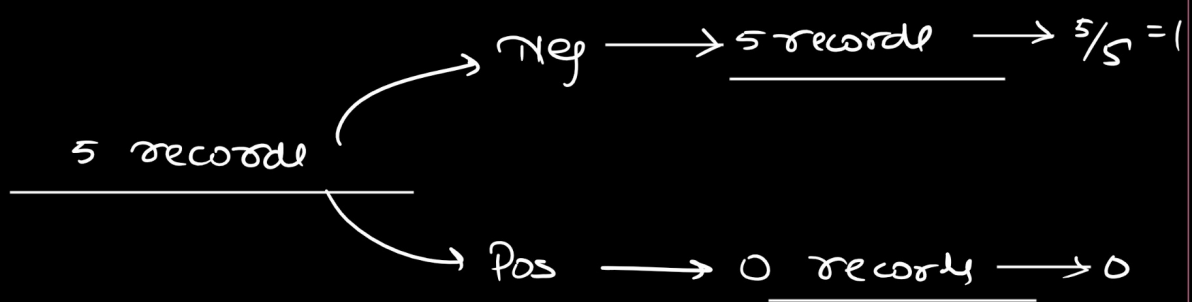
$$\Rightarrow \underline{\underline{0.974}}$$

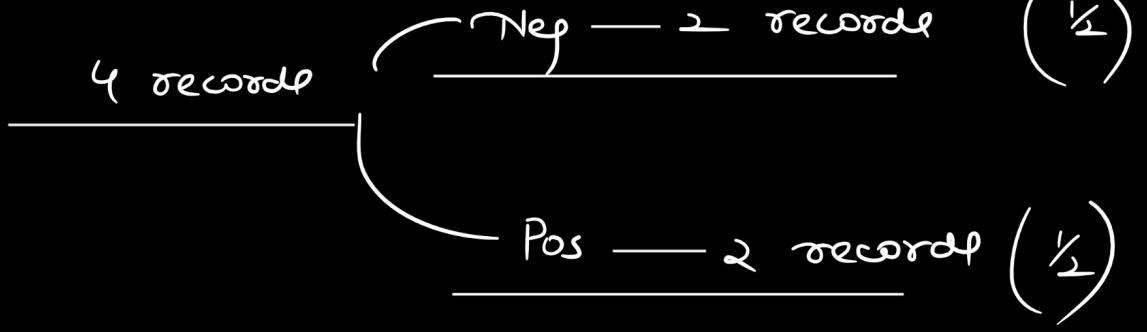
Note: Which features give the most information

↳ Information Gain

[0, 1]

Pure	Impure
Entropy → $-0 \log 0 -$	$-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$
$1 \log 1$	$\Rightarrow 1$
$\Rightarrow 0$	

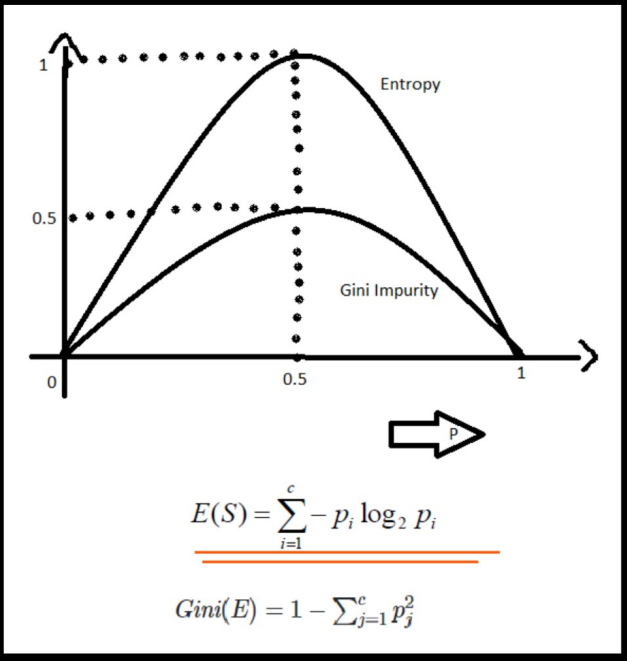




Highest Impurity

Gini → 0.5

Entropy → 1



$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$Gini(E) = 1 - \sum_{j=1}^{c} p_j^2$$

Task

use case

1. When should we go for "Entropy" as the criterion?
2. Decision Tree Regressor