Agenda

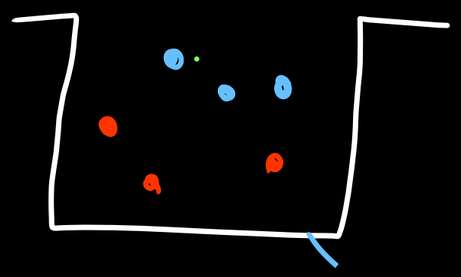
- → Gini Index
- → Underfitting & Overfitting
- → Feature Scaling
- → Feature Importance

Entropy →

$$H(y) = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

→ Gini Impurity

$$GI(y) = 1 - \sum_{i=1}^{C} p_i^2$$



$p_b = 3/6 = 1/2$ ✓

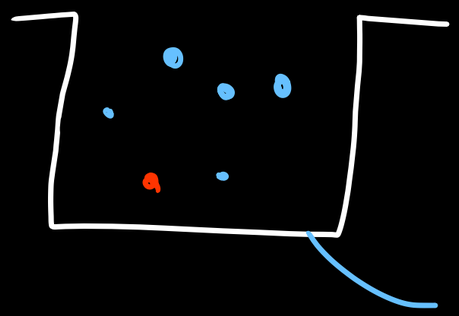
$p_n = 3/6 = 1/2$

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

$Gi = 1 - \sum p_i^2 = 1 - \left[p_b^2 + p_n^2\right]$

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

$= 0.5$ ✓



$p_b = 5/6$

$p_n = 1/6$

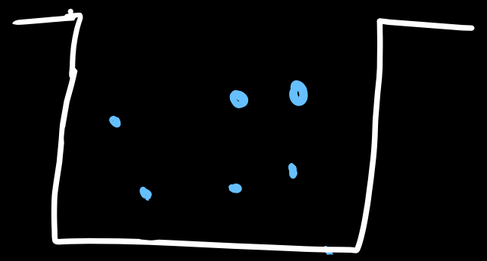
$H(y) = -\left[\frac{5}{6}\log_2(5/6) + \frac{1}{6}\log_2\left(\frac{1}{6}\right)\right]$

$= 0.65$

p_b p_n

$G(y) = 1 - \left[\left(\frac{5}{6}\right)^2 + \left(\frac{1}{6}\right)^2\right]$

$= 0.277$



$P_b = 1$

$P_\wedge = 0$

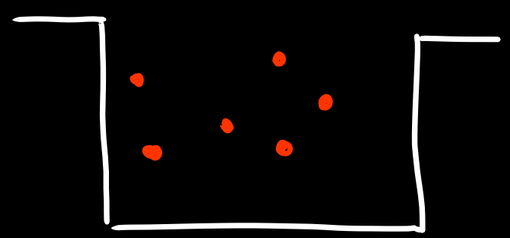
$$H(y) = -\left[1 \log_2(1) + 0 \log_2(0)\right]$$

$$= 0 \checkmark$$

$$Gn(y) = 1 - \left[(1)^2 + (0)^2\right]$$

(P_b, $P_\wedge$)

$$= 0 \checkmark$$



$P_b = 0$

$P_\wedge = 1$

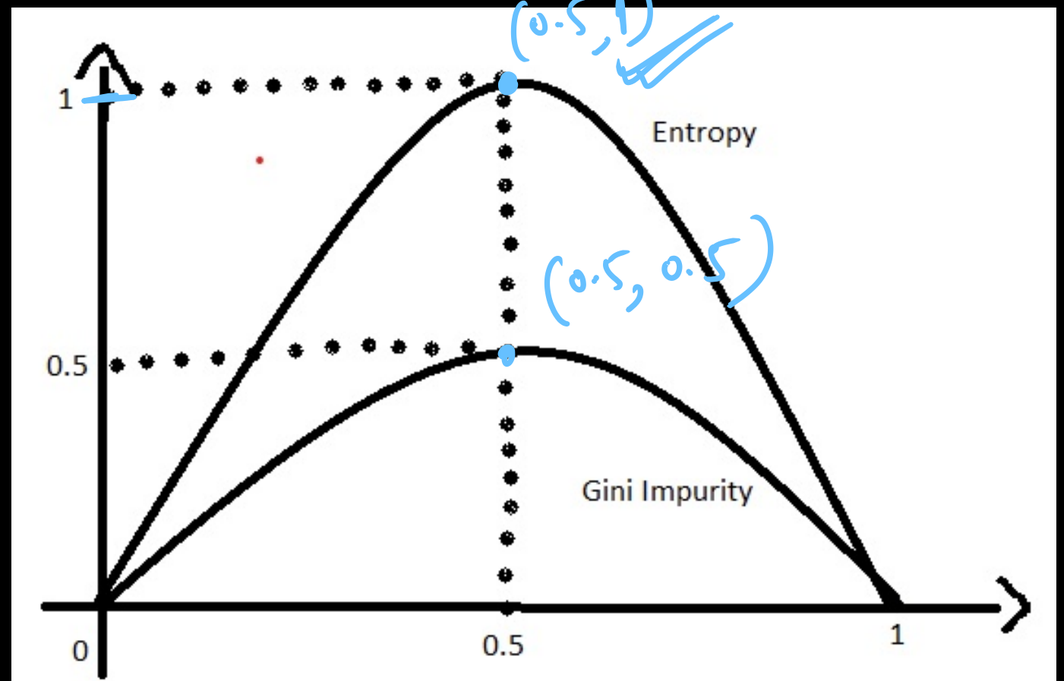
$$H(y) = -\left[0 \log_2(0) + 1 \log_2(1)\right]$$

$$= 0 \checkmark$$

$$Gn(y) = 1 - \left[(0)^2 + (1)^2\right]$$

(P_b, $P_\wedge$)

$$= 0 \checkmark$$



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

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

→ Range of Entropy
[0, 1]

→ Range of Gini
[0, 0.5]

Entropy

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

2) Homogenous $\rightarrow$ 0

3) $P_i = 0.5 \rightarrow H(y) = 1$

4) Slightly slower due $\log_2$

Gini Index

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

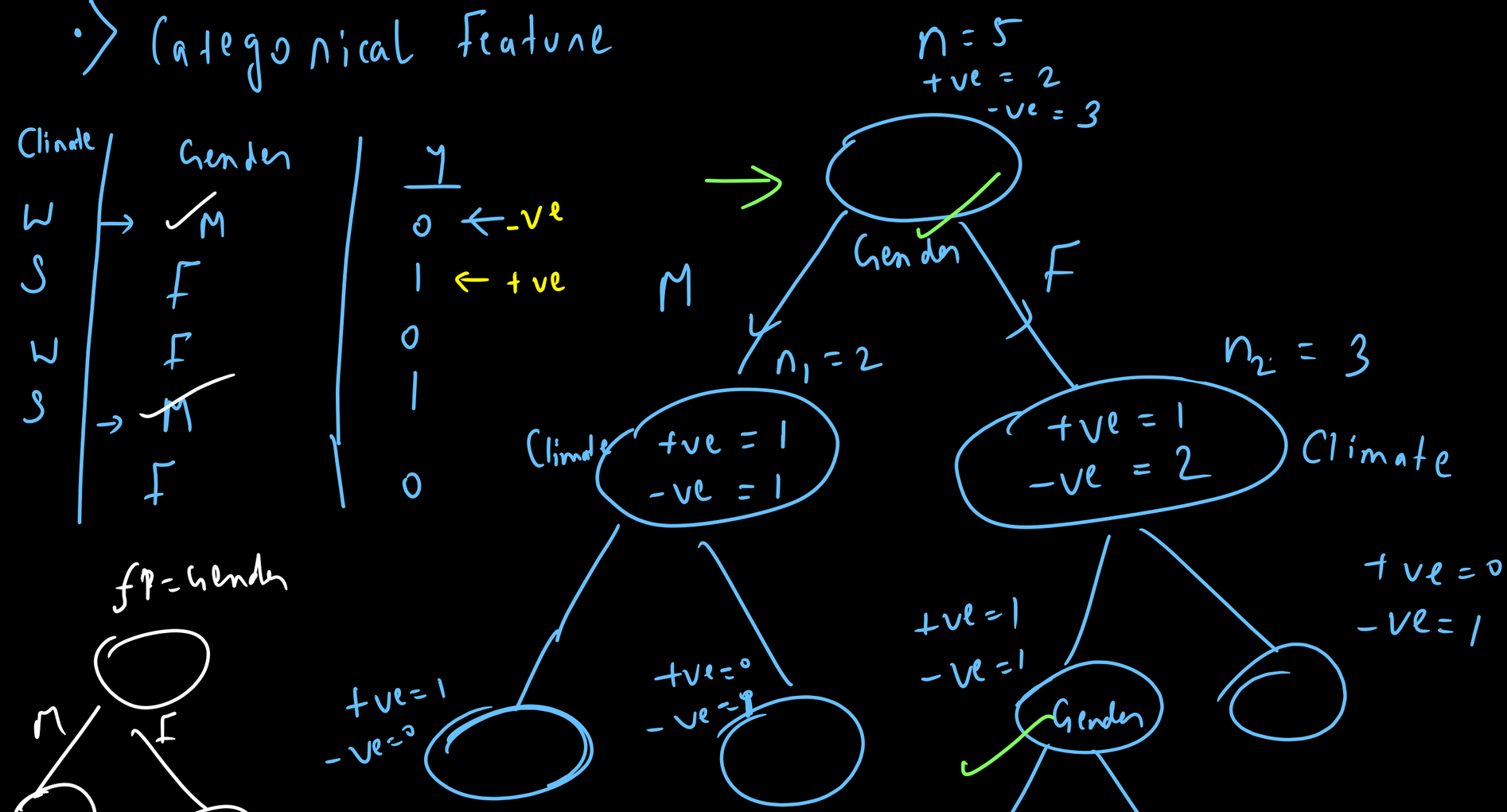
Homogenous $\rightarrow$ 0

$P_i = 0.5 \rightarrow GI(y) = 0.5$

Slightly faster

:) Categorical Feature

Climate	Gender	y
W	M	0 ← -ve
S	F	1 ← +ve
W	F	0
S	M	1
	F	0



1) Numerical feature

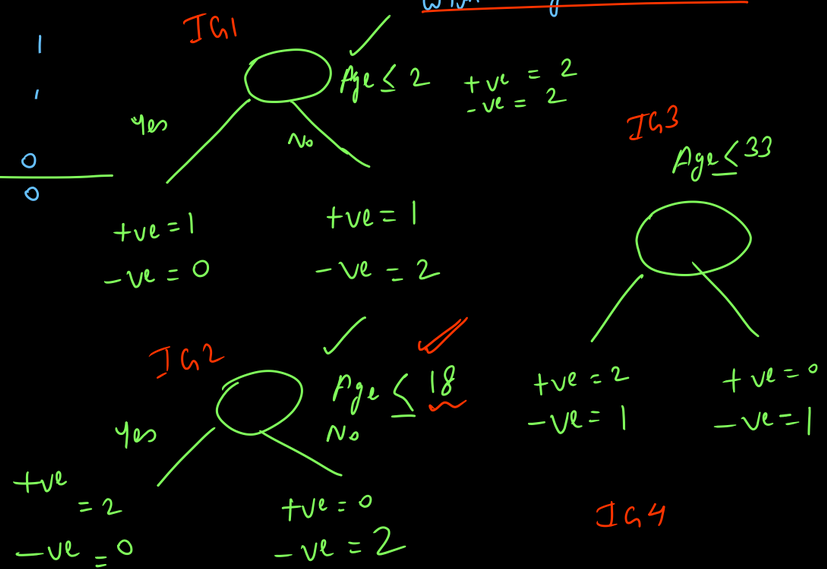
Age	y
18	1
33	0
41	0
2	1

1) Sort the numerical column in ascending order

2) Take each unique value in the numerical feature as threshold and calculate IG

3) Choose the threshold with highest IG

Age	y
2	1
18	1
33	0
41	0

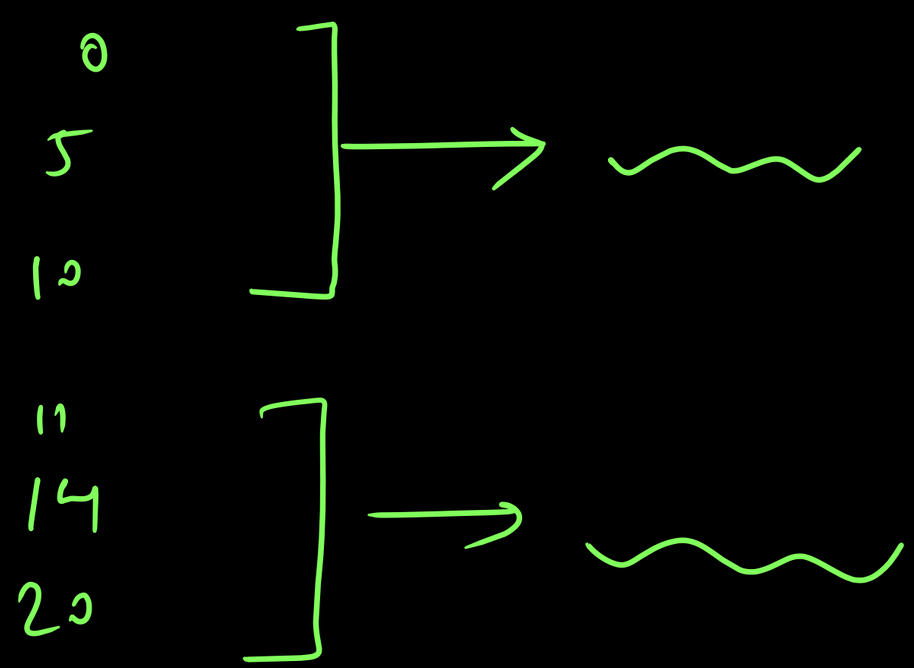


Age ≤ 18

$$IG = H(P) - \left[\frac{n_1}{n} H(C_1) + \frac{n_2}{n} H(C_2)\right]$$

$$IG = GI(P) - \left[\frac{n_1}{n} GI(C_1) + \frac{n_2}{n} GI(C_2)\right]$$

Bining



1

1

→ Overfitting & Underfitting

↓

Fits training data perfectly (noise)

&

it is unable to generalize to test data

→ High model complexity

↓

Does not fit training data properly

→ Simple model

f1

f2

f1

Shallow tree

y+ = 500

y- = 500

f2

→ d = 0

y+ = 81

y- = 81

→ d = 1

min_sample

- split

= 10

→ d = 2

n = 10

y+ = 20

y- = 20

y- = 1

y+ = 0

→ d = 3

y+ = 5

y- = 2

n = 7

d = 4

Deeper tree

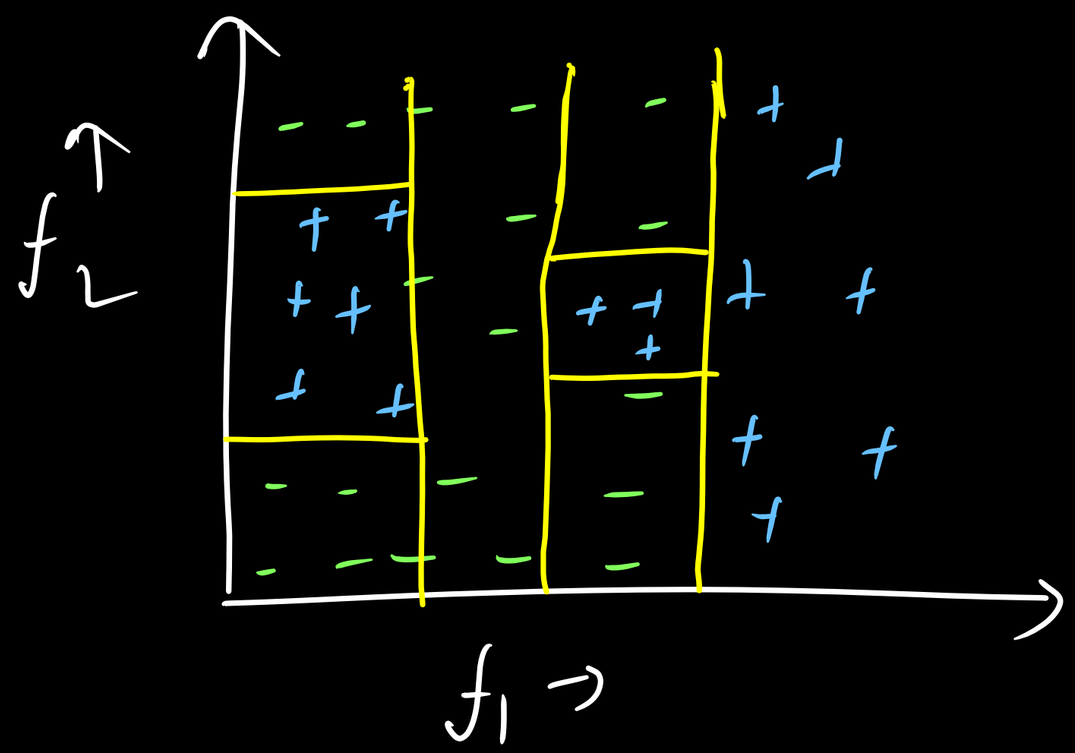
y+ = 1

y- = 0

y- = 5

y+ = 0

→ d = 5



for depth in [5, 10, 20]:

model = Decision (depth)

→ model.fit (X_train, y_train)

→ model.score (X_val, y_val)

f_1 f_2 —— f_d

→ Break until 22:24 PM

→ Feature Scaling

Age	Y
25	1
30	1
35	1
45	0
60	0

Age ≤ 30 ✓

Yes → $n_1 = 2$
+ve = 2
-ve = 0
$P_+ = 2/2 = 1$
$P_- = 0 = 0$

No → $n_2 = 3$
+ve = 1
-ve = 2
$P_+ = 1/3$
$P_- = 2/3$

Age ≤ −0.725 ✓

Age → (x−μ)/σ	Y
−1.12	1
−0.725	1
−0.322	1
0.48	0
1.69	0

+ve = 2
-ve = 0
$P_+ = 2/2 = 1$
$P_- = 0/2 = 0$

+ve = 1
-ve = 2
$P_+ = 1/3$
$P_- = 2/3$

IG → H(y) / GI(y)

GI → P_i ✓

Data Imbalance

y
NC
NC
C

← 99 NC, 1 C

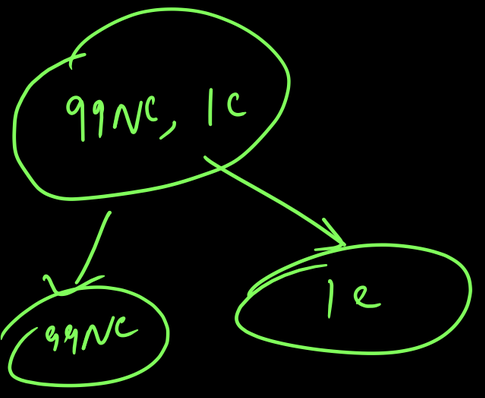
$P_C = 0.01$

$P_{NC} = 0.99$

$GI = 1 - \sum P_i^2$

$= 1 - [(0.01)^2 + (0.99)^2]$

$= 0.0198 \approx 0$



→ For homogenous → $GI = 0$

→ Class weight
→ SMOTE
→ Under/Over

→ Class weight

4 Nc
→ 1 C

$P_{Nc} = 4/5 = 0.8$

$P_c = 1/5 = 0.2$

$GI = 1 - [0.8^2 + 0.2^2] = 0.32$

weight $_C = 4$

weight $_{NC} = 1$

Weight NC count = $4 \times 1 = 4$

Weight C count = $1 \times 4 = 4$

$P_{NC} = \frac{4}{4+4} = \frac{1}{2}$

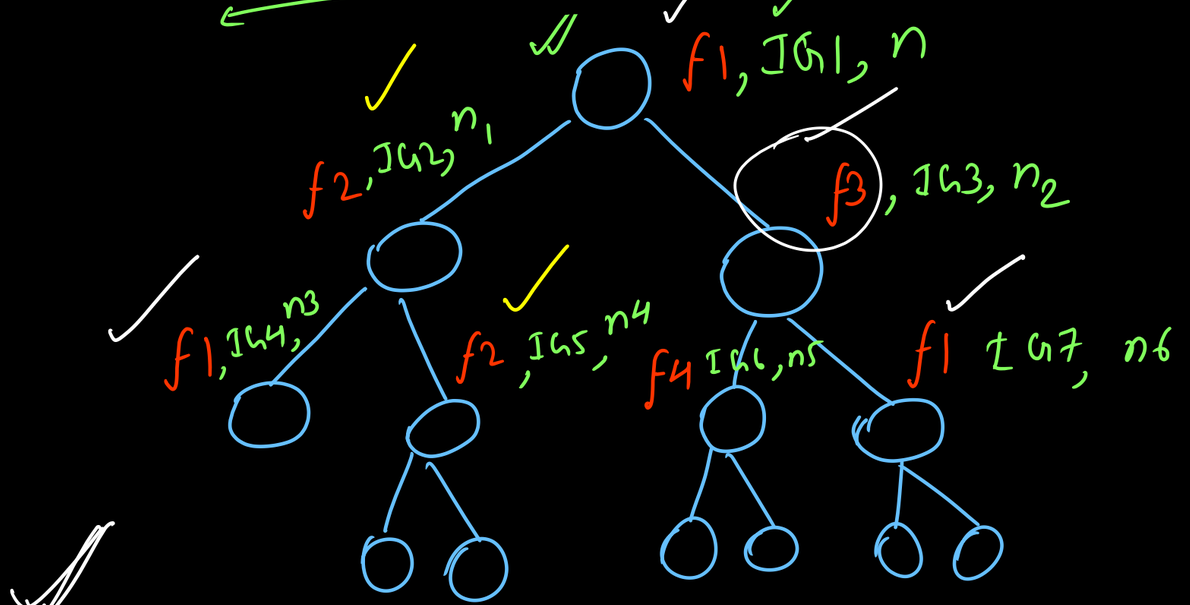
$P_c = \frac{4}{4+4} = \frac{1}{2}$

R = 0.6
G = 0.3
B = 0.1

$GI = 1 - [0.5^2 + 0.5^2] = 0.5$

$H(P) - \left[\frac{n_1}{n} H(C_1) + \frac{n_2}{n} H(C_2)\right]$ ← Feature Importance

f1	f2	f3	f4	y



Feature Importance $f1 = \frac{n3}{n} IG4 + \frac{n6}{n} IG7 + \frac{n}{n} IG1$

↳ Weight. Avg of IG's where f1 is used for split

Feature Importance $f2 = \frac{n4}{n} IG5 + \frac{n1}{n} IG2$

Feature Importance $f3 = \frac{n2}{n} IG3$

→ 0.6
→ 0.3
→ 0.1