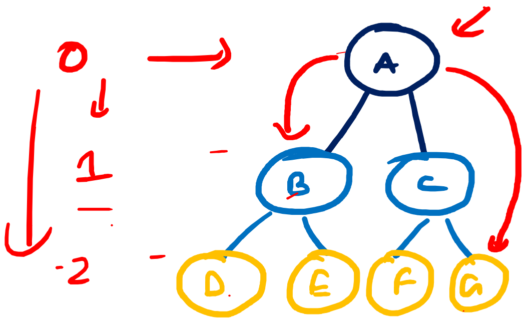
→ Decision Trees → Hierarchical Trees

→ Supervised Learning , F + T

→ Classification , Regression



Root node (A) → The starting of the tree.

Splitting → Process of dividing a node into subnodes

Decision Node → Any node that is participating in a condition or has child branches (A, B, C)

Parent Node → Any node that has a child. (A, B, C)

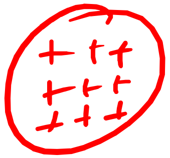
Child Node → Any sub node that has a parent (B, C, D, E, F, G)

Leaf Node → Any node that is NOT a decision node or not has a child or the tree ends there (D, E, F, G)

Depth → # of generations. Root node = 0 depth | Siblings → Nodes at same depth

→ Pure Node

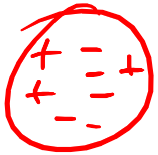
Any Node that has Rows of the same class



A

→ Impure Node

Any Node that has rows of multiple class



A

→

Algorithms	Splitting criteria
CART → Regression Also	Gini
C4.5	Entropy
ID3 → classification	Entropy
CHAID	Chi-square

⇓

Decision Trees

Splitting Criterias

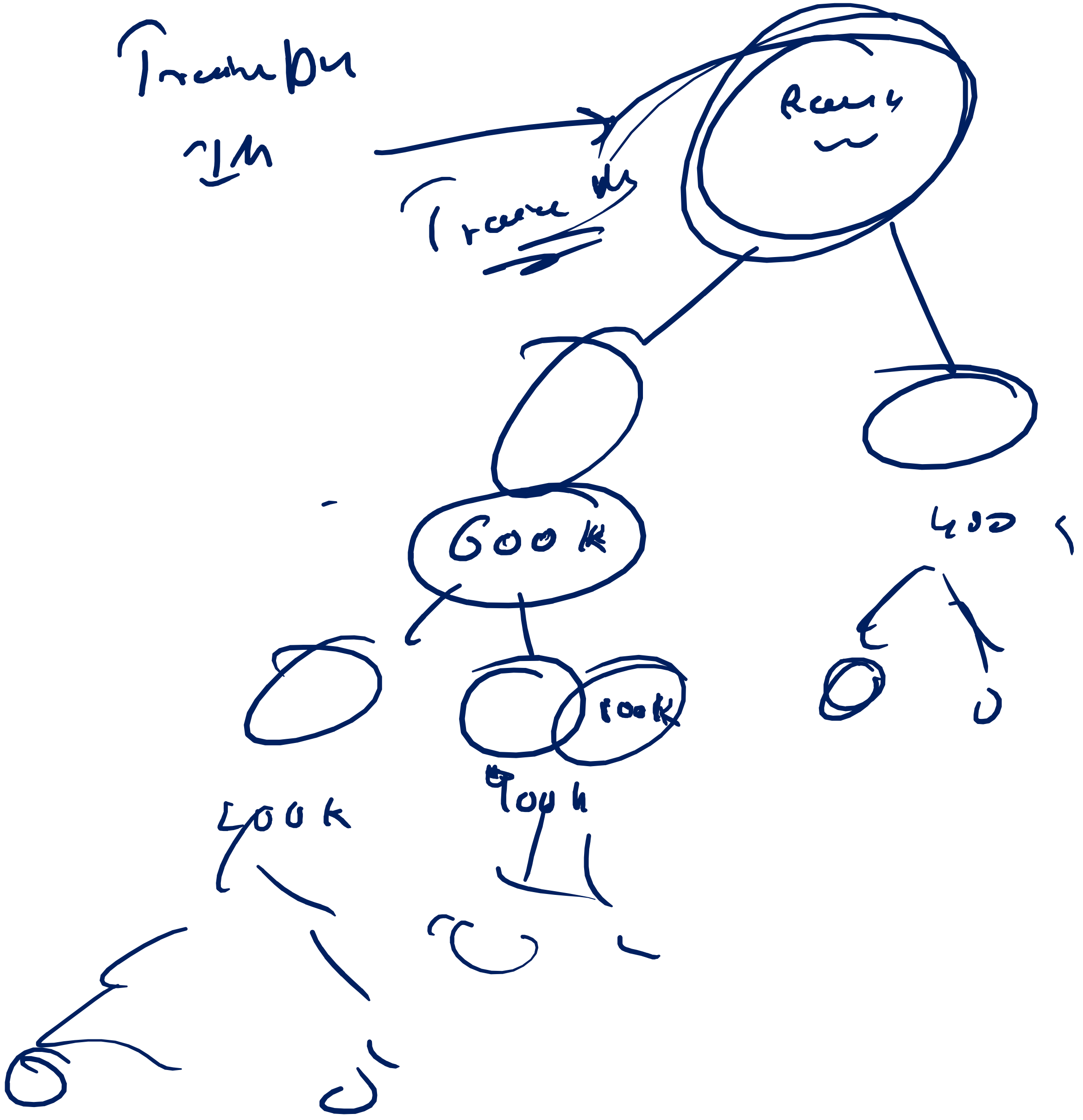
Features: Performance, Height, Class

Target: Plays Cricket

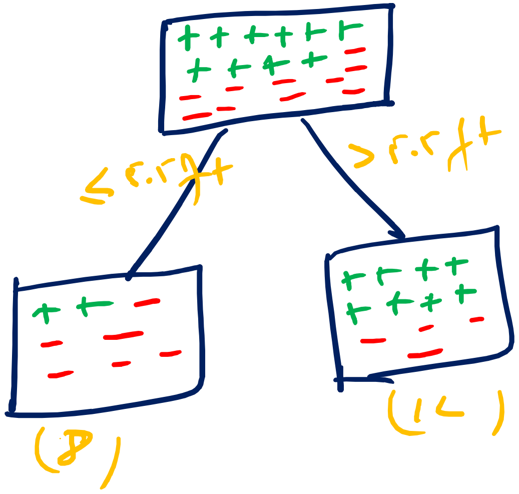
S.No.	Name	Performance	Height	Class	Plays Cricket
1	A	Above Avg.	> 5.5	IX	0
2	B	Below Avg	≤ 5.5	X	0
3	C	Below Avg	≤ 5.5	X	1
4	D	Above Avg	> 5.5	X	0
5	E	Below Avg	> 5.5	IX	1
⋮	⋮	⋮	⋮	⋮	⋮
20					

1 → +ve Class → 10/20

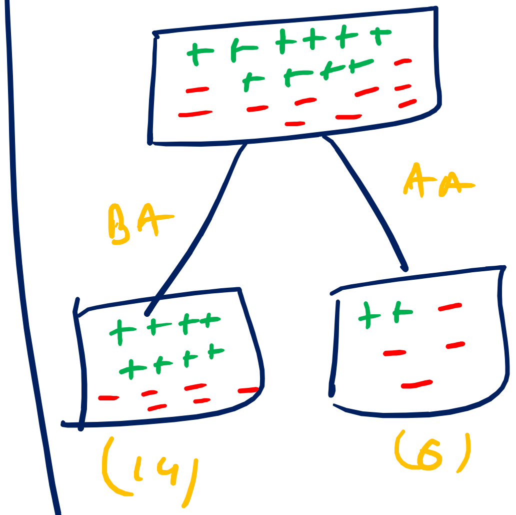
0 → -ve class → 10/20


Training Data
~1M
Training
Root
Good
400
Look
Touch
Took

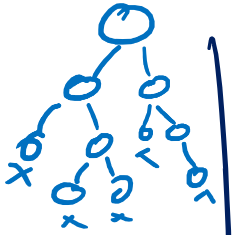
→

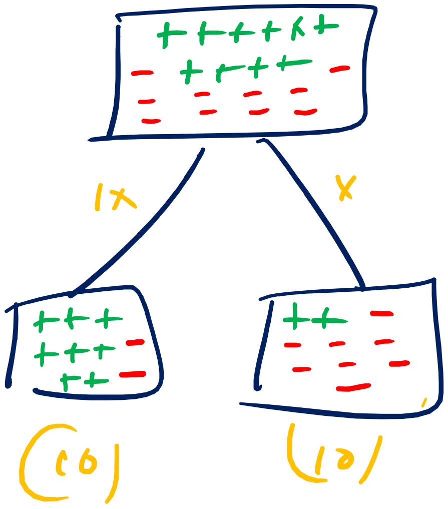


Hieght



Performance





Class

Now to decide which one feature to choose as the Best split

↳ based on splitting criteria you choose

① Gini Index & Gini Impurity (G.I.)

Gini Index: How much purity the split will provide

Gini Impurity: how much Impurity is being left after split.

G.I. Height 5 | Perf. ③

Gini Impurity = 1 − Gini

Prob of the class in that node

Gini of a Node = $P_1^2 + P_2^2 + P_3^2 + P_4^2 + \ldots + P_n^2$, n = # of categories in the target

Gini of a split = Weighted Avg (Node A, Node B) = $\dfrac{W_A G_A + W_B G_B}{2}$

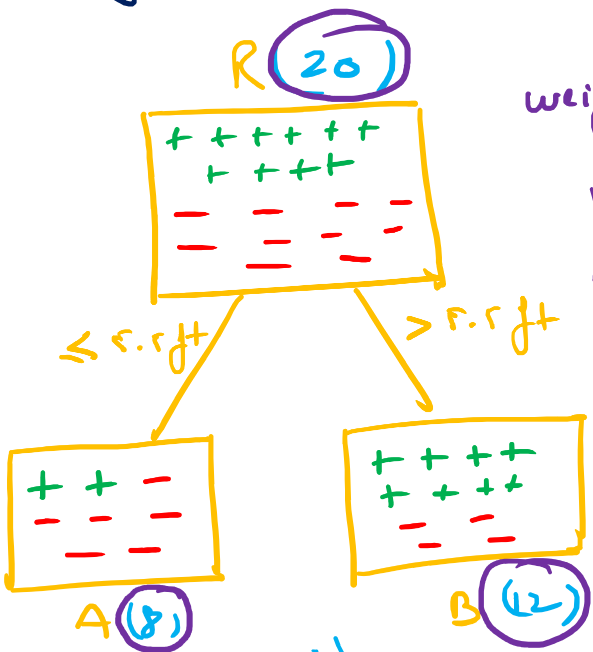
→ Split on Hieght based Gini Impurity

$$G.I._{\text{Hieght}} = W_A \, G.I_A + W_B \, G.I_B$$

$\Downarrow$

weighted G.I. $\Rightarrow \frac{8}{20} * 0.375 + \frac{12}{20} * 0.442$

$$\boxed{G.I_{\text{hieght}} \Rightarrow 0.415}$$



+ve = 2

$P(+ve) = 2/8 = 0.25$

$P(-ve) = 0.75$

$P(+ve) = 8/12 = 0.67$

$P(-ve) = 0.33$

$$G.I. = 1 - \text{Gini} = 1 - (P_1^2 + P_2^2 + P_3^2 + \ldots P_n^2)$$

$$G.I._A = 1 - [P(+ve)^2 + P(-ve)^2]$$

$$= 1 - [(0.25)^2 + (0.75)^2]$$

$$\Rightarrow 0.375$$

$$G.I._B = 1 - [P(+ve)^2 + P(-ve)^2]$$

$$= 1 - [0.67^2 + 0.33^2]$$

$$\Rightarrow 0.442$$

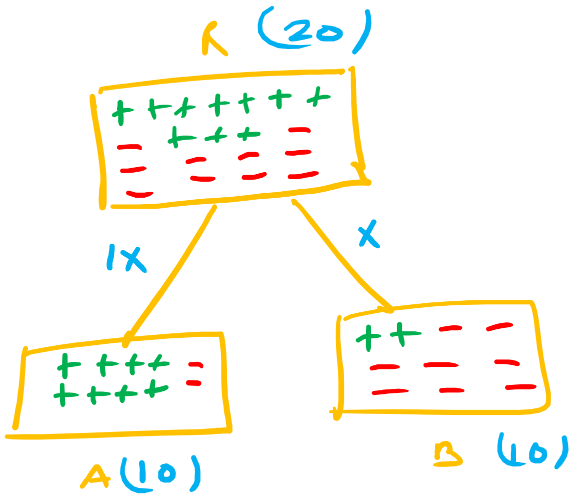
→ Split on Class based on Gini Impurity

$G.I._A = 1 - [0.8^2 + 0.2^2]$

$\Rightarrow 0.32$

$G.I._B = 1 - [0.2^2 + 0.8^2]$

$\Rightarrow 0.32$



$P(+ve) = 0.8$

$P(-ve) = 0.2$

$P(+ve) = 0.2$

$P(-ve) = 0.8$

$G.I._{Class} \Rightarrow w_A GI_A + w_B GI_B$

$\Rightarrow \frac{10}{20} * 0.32 + \frac{10}{20} * 0.32 = 0.32 \cancel{\left(\frac{1}{2} + \frac{1}{2}\right)}$

$\boxed{G.I._{Class} \Rightarrow 0.32}$

Feature	G.I
Height	0.415
✓ Class	0.32
Perf ✓	0.16