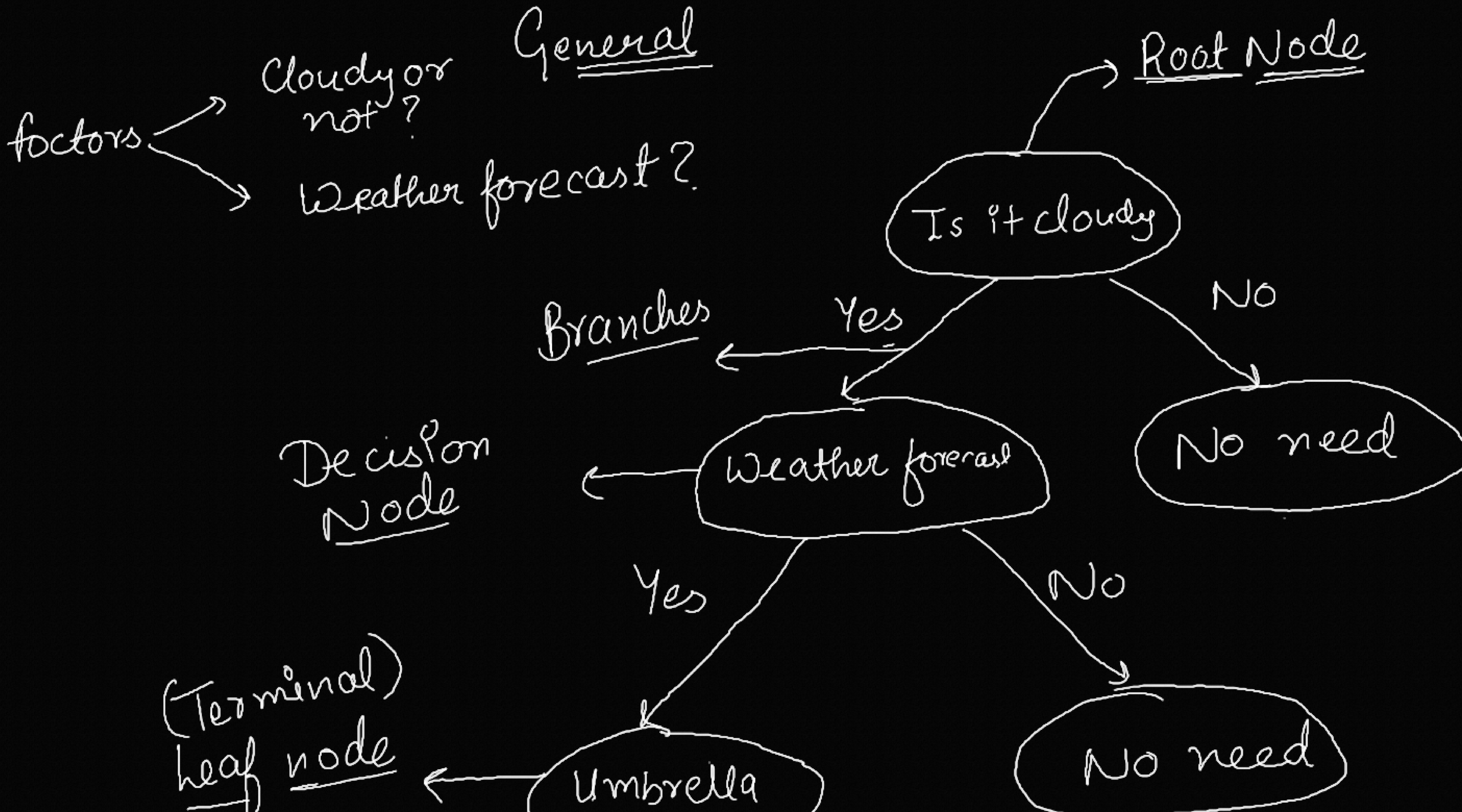
# Decision Trees

- * Classification & Regression
- * Core of Ensemble Learning
- * Looks like Nested-If else
- * Structure is like a tree
- * History

1986
Ross Quinlan
ID3 ⟶ 1986
CART
Leo, Jerome, Richard, Charles
Classification ⟶ Gini, Regression ⟶ MSE

⟶ 1990's
(Pruning)


General
factors
Cloudy or not ?
Weather forecast ?
Root Node
Is it cloudy
Yes
NO
Branches
Weather forecast
No need
Decision Node
Yes
No
(Terminal) leaf node
Umbrella
No need

# * Criteria to choose Nodes

## Classification

- Gini Impurity
  - Lower value is best value
- Information Gain
  - Greater value is best value

* Entropy (concept)

## Regression

- MSE
- MAE

# Mathematical (Classification)

|  | $X_1$ | $X_2$ | $Y$ |
|---|---|---|---|
| Person | Income | Age | Buy House? |
| 1 | High | Young | No |
| 2 | High | Old | Yes |
| 3 | Low | Young | No |
| 4 | Low | Old | Yes |

## ① Calculate Total Gini Impurity

$$\boxed{\text{Gini} = 1 - \sum p_i^2}$$

$P(\text{Buy House} = \text{Yes}) = \frac{2}{4} = 0.5$

$P(\text{Buy House} = \text{No}) = \frac{2}{4} = 0.5$

$\text{Gini} = 1 - \left( (0.5)^2 + (0.5)^2 \right)$

$= \underline{\underline{0.5}}$ (overall gini)

② Calculate Gini for splits

(a) Income

(i) Income = "High" $\Rightarrow$ P1 → No
P2 → Yes

$P(\text{Yes}) = \frac{1}{2} = 0.5$ $P(\text{No}) = \frac{1}{2} = 0.5$

$$\text{Gini}(\text{Income} = \text{High}) = 1 - (0.5^2 + 0.5^2)$$
$$= \underline{\underline{0.5}}$$

(ii) Income = Low $\Rightarrow$ P3 → No $\Rightarrow$ $P(\text{Yes}) = 0.5$
P4 → Yes $P(\text{No}) = 0.5$

$$\text{Gini}(\text{Income} = \text{Low}) = \underline{\underline{0.5}}$$

* Weighted Gini (Income) $= \underbrace{0.5 \times \frac{2}{4}}_{\text{High}} + \underbrace{0.5 \times \frac{2}{4}}_{\text{Low}}$

$= \underline{\underline{0.5}}$

(b) **Age**

(i) Age = Young ⇒ P1 → No

P3 → No

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

$P(\text{No}) = \frac{2}{2} = 1$

$\text{Gini}(\text{Age} = \text{Young}) = 1 - [1^2 + 0^2]$

$= \underline{\underline{0}}$

(iii) Age = Old $\Rightarrow$ P2 ⟶ Yes $\quad P(\text{Yes}) = 1$

P4 ⟶ Yes $\quad P(\text{No}) = 0$

$$\text{Gini}(\text{Age} = \text{Old}) = 1 - (1^2 + 0^2)$$

$$= \underline{\underline{0}}.$$

* $$\text{Weighted Gini}(\text{Age}) = 0 \times \frac{2}{4} + 0 \times \frac{2}{4}$$

$$= \underline{\underline{0}}.$$

## ③ Choose the best split

$\text{Gini}(\text{Income}) = 0.5$

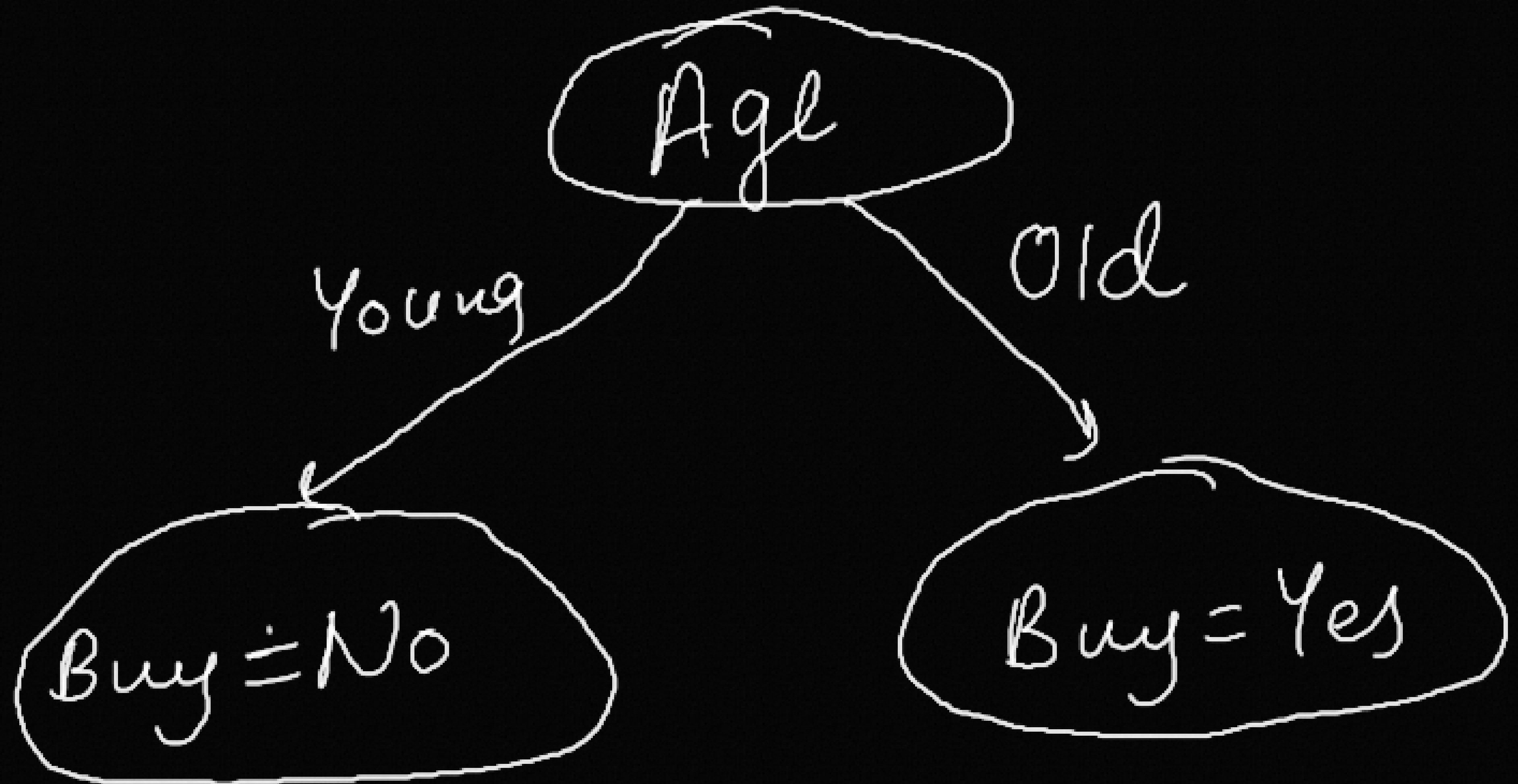
$\text{Gini}(\text{Age}) = 0$ ✓

Lower is best.

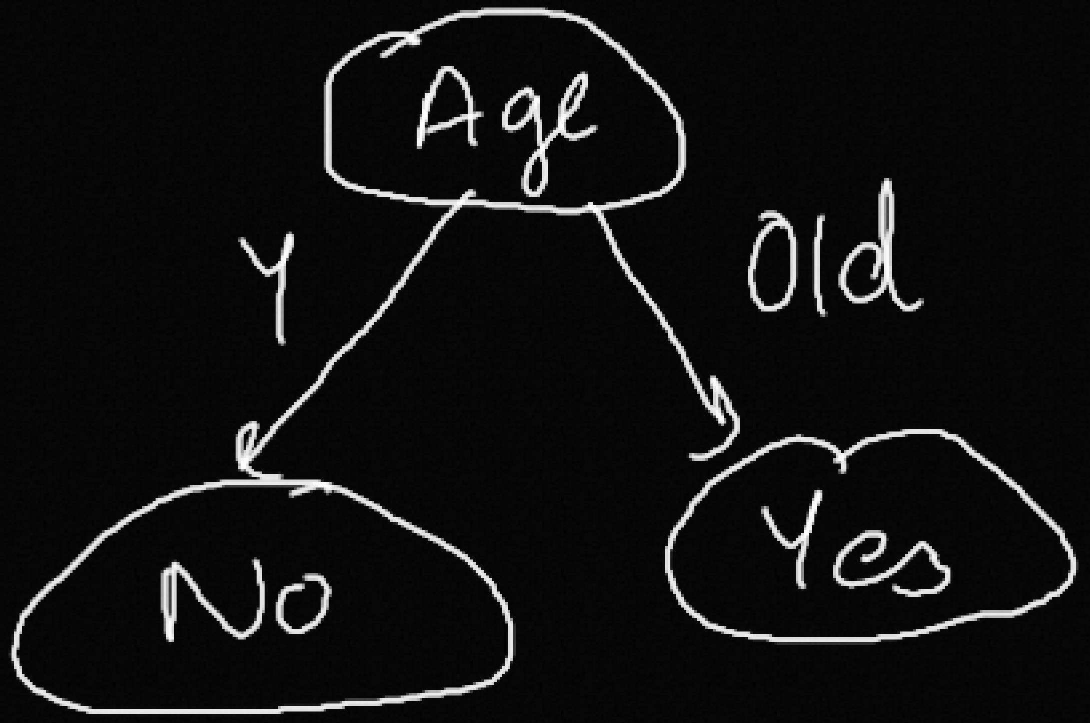
When $\text{gini} = 0$ → perfect / pure split

(DT will end when you get pure split)

final.



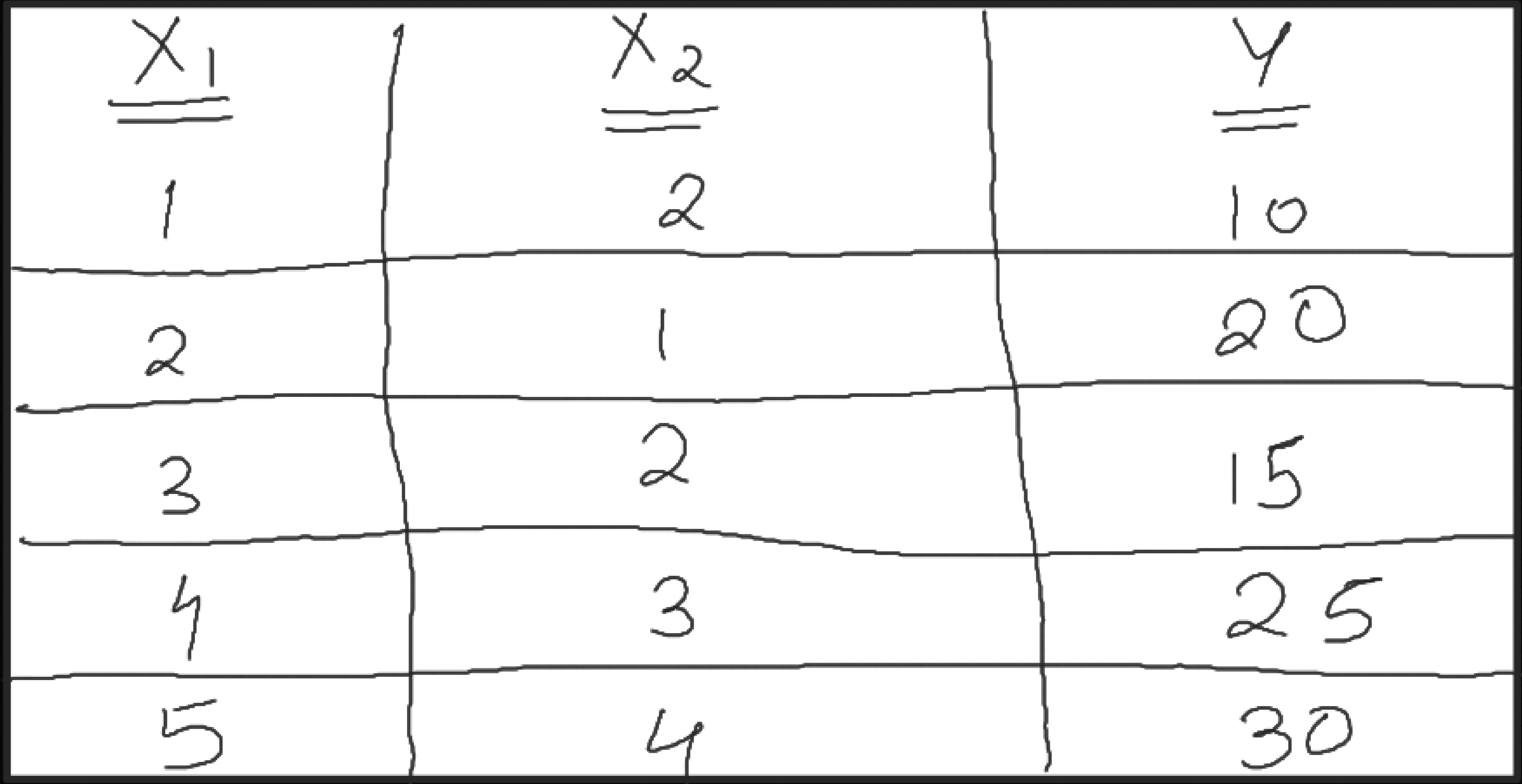
④ Repeat all the steps again till you get pure split



(Information Gain)

| $Y_1$ | $X_2$ | Y |
|---|---|---|
| High | Young | No |
| Low | Young | No |

Same approach



| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| 1 | 2 | 10 |
| 2 | 1 | 20 |
| 3 | 2 | 15 |
| 4 | 3 | 25 |
| 5 | 4 | 30 |

## Regression (Maths)

① Calculate Total MSE

Mean= 20 ($\bar{x}$)

$$MSE = \frac{1}{n}\sum(\bar{x}-x_i)^2$$

$$MSE = \frac{1}{5}\left[(20-10)^2+(20-20)^2+(20-15)^2+(20-25)^2+(20-30)^2\right]$$

$$= \underline{\underline{50}}$$

## ② Calculate MSE at each feature

(a) $X_1$ ⇒ Threshold = 1.5

$X_1 \leq 1.5$ (Left subset)

| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| 1 | 2 | 10 |

$X_1 > 1.5$ (Right subset)

| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| 2 | 1 | 20 |
| 3 | 2 | 15 |
| 4 | 3 | 25 |
| 5 | 4 | 30 |

Left subset

Mean = 10

$MSE = 0$

Right subset

Mean = 22.5

$MSE = 31.25$

$$* \text{ Weighted } MSE(X_1) = \frac{1}{n}\left[\begin{matrix}\text{no of} \\ \text{rows} \\ \text{(left)}\end{matrix} \times MSE + \begin{matrix}\text{no of} \\ \text{row} \\ \text{(right)}\end{matrix} \times MSE\right]$$

$$= \frac{1}{5}\left[1 \times 0 + 4 \times 31.25\right]$$

$$= 25$$

(b) $X_2$ ⇒ Threshold = 2·5

$X_2 \leq 2\cdot5$
(Left)

$X_2 > 2\cdot5$
(Right)

| $X_1$ | $X_2$ | Y |
|---|---|---|
| 1 | 2 | 10 |
| 2 | 1 | 20 |
| 3 | 2 | 15 |

Mean = 15

MSE = 16·67

| $X_1$ | $X_2$ | Y |
|---|---|---|
| 4 | 3 | 25 |
| 5 | 4 | 30 |

Mean = 27·5

MSE = 6·25

$$* \text{ Weighted MSE } (X_2) = \frac{1}{5}\left[3 \times 16.67 + 2 \times 6.25\right)$$

$$= \underline{\underline{12.25}}$$

③ Choose best split

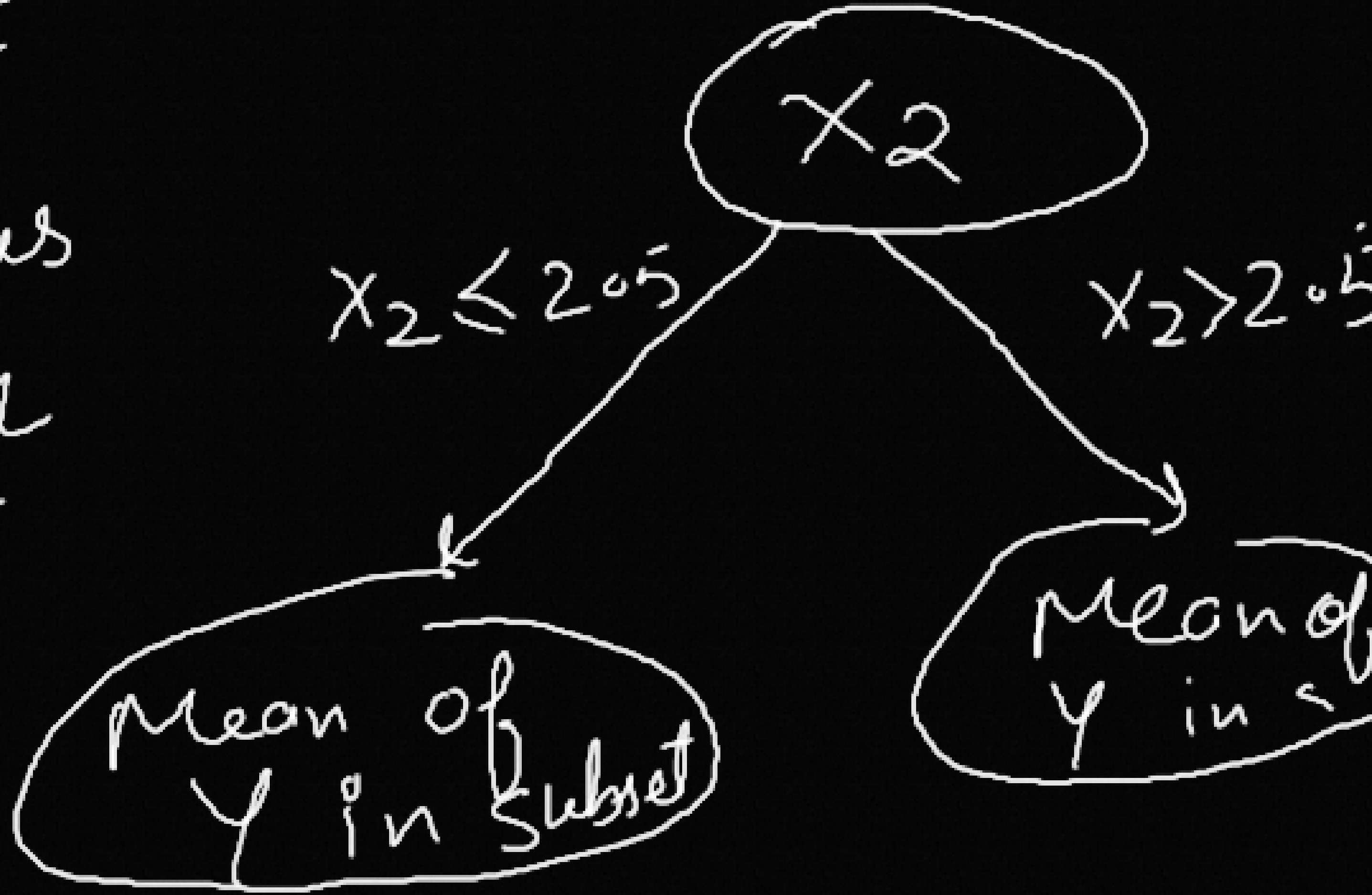
$X_1 = 25$

$X_2 = 12.25$ ✓

Lower is better

④ Repeat steps

Tree



Answer

## Pros

1. Easy to understand like if-else.
2. No feature scaling reqd.
3. Can handle Non linear data.
4. feature selection can do.

## Cons

1. Overfitting
2. Computationally extensive