Gini Index (CART)

Color	Diameter	Label
Green	3	Mango
Yellow	3	Lemon
Red	1	Grape
Yellow	3	Mango
Red	1	Grape

Step-1 → Calculate Gini Index, for the Root Node.

We have 5 fruits → 2 Mangoes, 1 Lemon, 2 Grapes.

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

$$= 1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 \right)$$

$$= 1 - \left(\frac{4}{25} + \frac{1}{25} + \frac{4}{25} \right)$$

$$= 1 - \frac{9}{25} = 1 - 0.36 = 0.64.$$

Gini Index at root node is 0.64.

Step-2 → Evaluate Split based on Gini Index.

We now evaluate potential skills on both color & diameter to see which gives the best separation of data.

Split on Color →

→ Group 1 (Green):
- [Mango] → Pure, only Mango.
- Gini for Group 1 = 0 (since it's pure).

→ Group 2 (Yellow):
- [Lemon, Mango] → 1 Lemon, 1 Mango → Mixed
- Gini for Group 2:

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

$$= 1 - \frac{1}{4} + \frac{1}{4} = 0.5.$$

→ Group 3 (Red):
- [Grape, Grape] → Pure, Only Grapes.
- Gini for Group 3 is 0.

Now, weighted Gini for this split →

$$Gini_{total} = \frac{1}{5} \times 0 + \frac{2}{5} \times 0.5 + \frac{2}{5} \times 0$$

$$= 0 + 0.2 + 0 = 0.2.$$

Split on Diameter. →

→ Group 1 (Diameter = 3)

• [Mango, Lemon, Mango] → 2M, 1L → Mixed.

• Gini → $1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right)$.

$= 1 - \left(\frac{4}{9}\right) + \left(\frac{1}{9}\right)$.

$= 1 - \frac{5}{9} = 0.44$.

→ Group-2 (Diameter = 1)

• [Grape, Grape] → Pure, only Grapes.

• Gini = 0

Compute weighted Gini for this split :-

$Gini_{total} = \frac{3}{5} \times 0.44 + \frac{2}{5} \times 0$

$= 0.264$

Step-3 → Choose the Best Split

→ Gini for Color = 0.2

→ Gini for Diameter = 0.264.

G_c is lower we choose Gini for color for split.

Step-4 → Continue Splitting.

→ Green - [Mango] - Pure (Predict Mango)
→ Yellow - [Lemon, Mango] - Mixed (Requires further)
→ Red - [Grape, Grape] - Pure (Predict Grape).

Now, focus on Yellow Group.
(Lemon, Mango)

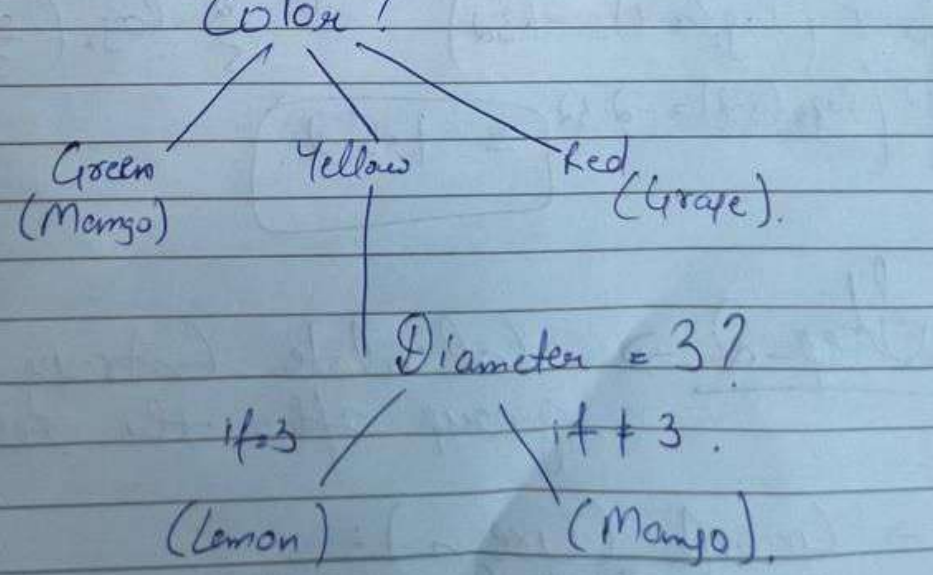
Diameter = 3

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

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

$$= 0.5$$

FINAL DECISION TREE



INFORMATION GAIN (ID3/C4.5)

Step-1 → Calculate Entropy of the Root Node.

~~Entropy is a measure~~

Entropy is a measure of the impurity or disorder in dataset. It is calculated using formula.

$$\text{Entropy} = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

Where :-

- p_i is the proportion of data points in class i
- n is the number of classes.

In root node, we have 2M, 1L, 2G.

$p(M) = 2/5$, $p(L) = 1/5$, $p(G) = 2/5$.

$$\text{Entropy}_{root} = -\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{1}{5}\log_2\left(\frac{1}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right)$$

$\frac{2}{5} = 0.4 \left(\log_2(0.4) = -1.32\right)$

$\frac{1}{5} = 0.2 \left(\log_2(0.2) = -2.32\right)$ = 1.52.

Step-2 → Calculate Entropy for each group after the split on color

→ Group 1 (Green):
[Mango] → Pure → Entropy = 0

→ Group 2 (Yellow):
[Lemon, Mango] → Mixed.

$Entropy_{Yellow} = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)$

$\log_2(0.5) = -1$

$= -(0.5 \times -1 + 0.5 \times -1)$

$-(0.5 + 0.5) = 1$

→ Grey 3 (Red):

- {Grape, Grape} → Pure → Entropy = 0.

Step-3 → Weighted Entropy.

$Entropy_{split} = \frac{1}{5} \times 0 + \frac{2}{5} \times 1 + \frac{1}{5} \times 0.$

$= \frac{2}{5} \{0.4\}.$

Step-4 → Calculate Information Gain.

Information Gain = $Entropy_{root} - Entropy_{split}$

$= 1.52 - 0.4$

$= 1.12.$