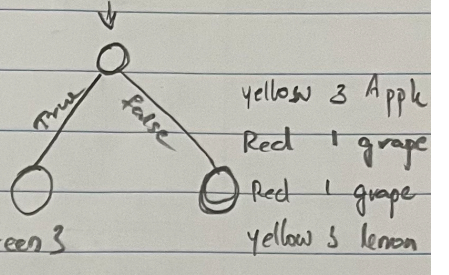
Q1)

Q1) If color is green?

Level 1: Root

$$\text{Entropy} = \sum_{i=1}^{c} -P_i \log_2 P_i$$

Color	Diam	label
green	3	Apple
yellow	3	Apple
Red	1	Grape
Red	1	Grap
yellow	3	Lemon



$$\text{Entropy} = -P(A)\log_2 P(A) - P(G)\log_2 P(G) - P(L)\log_2 P(L)$$

$$= -\frac{2}{5}\log_2\frac{2}{5} - \frac{2}{5}\log_2\frac{2}{5} - \frac{1}{5}\log_2\frac{1}{5}$$

$$= -0.4\log_2 0.4 - 0.4\log_2 0.4 - 0.2\log_2 0.2$$

$$\approx 0.53 + 0.53 + 0.46 \approx 1.52$$

Avg Imp $= \frac{5}{5} * 1.52 = 1.52$

Level 2:

left Subtree : $-P(A)\log_2 P(A) = -(1/1)\log_2(1/1) = -1\log_2 1 \approx 0$

Right Subtree : $-P(A)\log_2 P(A) - P(G)\log_2 P(G) - P(L)\log_2 P(L)$

$$= -\frac{1}{4}\log_2\frac{1}{4} - \frac{2}{4}\log_2\frac{2}{4} - \frac{1}{4}\log_2\frac{1}{4}$$

$$= -0.25\log_2 0.25 - 0.5\log_2 0.5 - 0.25\log_2 0.25$$

$$= 0.5 + 0.5 + 0.5 \approx 1.5$$

Average impurity or weighted Average $= \frac{4}{5} \times 1.5 + \frac{1}{5} \times 0 = 1.2$

Information gain $= 1.52 - 1.2 = 0.32$

If the feature color is green Information Gain from 2 methods are:

Gini $= 0.14$

Entropy $= 0.32$ Info gain is Entropy is higher than Gini

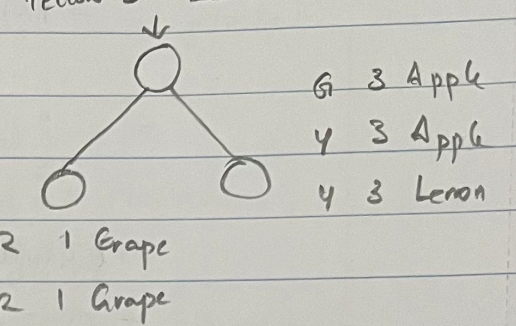
If diameter >= 3

Green	3	Apple
Yellow	3	Apple
Red	1	grape
Red	1	Grape
Yellow	3	Lemon

Level 1: Root

$$\text{Entropy} = -P(A)\log_2 P(A) - P(G)\log_2 P(G) - P(L)\log_2 P(L)$$

≈ 1.52



Avg. Imp $= \frac{5}{5} \times 1.52 = 1.52$

level 2

Entropy left Subtree : $-P(G) * \log_2 P(G) = -\frac{2}{2}\log_2\frac{2}{2} = 0$

Entropy right Subtree : $-P(A)\log_2 P(A) - P(L)\log_2 P(L)$

$$= -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}$$

$$= -0.67\log_2 0.67 - 0.33\log_2 0.33 \approx 0.38 + 0.52 \approx 0.90$$

Avg Impurity $= \frac{2}{5} \times 0 + \frac{3}{5} \times 0.9 = 0.54$

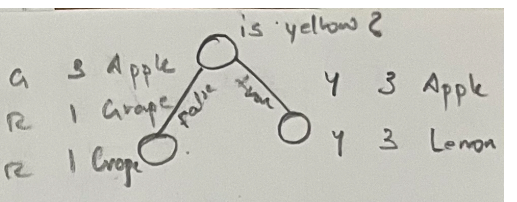
Information gain $= 1.52 - 0.54 = 0.98$

If the feature "diameter is >= 3", then information gain in 2 methods are:

Gini $= 0.37$

Entropy $= 0.98$ Information Gain in Entropy is higher than Gini

If color is Yellow ?



Level 1: Root

$$\text{Entropy} = -P(A) * \log_2 P(A) - P(G) * \log_2 P(G) - P(L) \log P(L)$$

$$\approx 1.52$$

$$\text{Avg Impunty} = \frac{5}{5} \times 1.52 = 1.52$$

Level 2:

Left Subtree : $-P(A)\log_2 P(A) - P(G)\log(P G) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$

$$= 0.33 \log_2 0.33 - 0.66 \log_2 0.66$$

$$= 0.52 + 0.38 \approx 0.90$$

Right Subtree : $-P(A)\log_2 P(A) - P(L)\log_2 P(L)$

$$= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} \quad = -0.5\log_2 0.5 - 0.5\log_2 0.5$$

$$= 0.5 + 0.5 \approx 1$$

$$\text{Avg Imp} = \frac{3}{5} * 0.90 + \frac{2}{5} * 1$$

$$= 0.54 + 0.4 = 0.94$$

$$\text{Information Gain} = \text{Avg Imp (Root)} - \text{Avg Imp(level2)}$$

$$= 1.52 - 0.94$$

$$= 0.54$$

If the feature color is "Yellow", then information gain by 2 methods are:

Gini = 0.17

Entropy = 0.54 Information gain by entropy is higher than Gini

Q2)

Jupyter Notebook:

https://github.com/abeednabith/CS483_AI/blob/main/assignment4/Q2_GINI_DT.ipynb

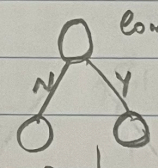
Dataset:

https://github.com/abeednabith/CS483_AI/blob/main/assignment4/Q2_GINI_DT.csv

Q2) Dataset :

Age	Competition	Type	Profit
Old	Yes	Software	Down
Old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	Software	Down
Mid	Yes	Hardware	Down
Mid	No	Hardware	Up
Mid	No	Software	Up
New	Yes	Software	Up
New	No	Hardware	Up
New	No	Software	Up
Mid	No	Hardware	?

If Competition is "Yes" ?



Old	N	Sw	Dn
Old	N	Hw	Dn
Mid	N	Hw	Up
Mid	N	Sw	Up
New	N	Sw	Up
New	N	Sw	Up

Old	Y	Sw	Dn
Mid	Y	Sw	Dn
Mid	Y	Hw	Dn
New	Y	Sw	Up

Level 1: Root

$$\text{Imp} = P(D) * (1 - P(D)) + P(U) * (1 - P(U))$$

$$= \frac{5}{10} \times \left(1 - \frac{5}{10}\right) + \frac{5}{10} \times \left(1 - \frac{5}{10}\right)$$

$$= \frac{5}{10} \times \frac{5}{10} + \frac{5}{10} \times \frac{5}{10}$$

$$= 0.5 \times 0.5 + 0.5 \times 0.5 \approx 0.5$$

$$\text{Avg Imp} = \frac{10}{10} \times 0.5 = 0.5$$

level 2:

Left Subtree

$$\text{Impurity} = P(D) * (1 - P(D)) + P(U) * (1 - P(U)) = \frac{2}{6} \times \left(1 - \frac{2}{6}\right) + \frac{4}{6} \times \left(1 - \frac{4}{6}\right)$$

$$= \frac{2}{6} \times \frac{4}{6} + \frac{4}{6} \times \frac{2}{6} = 0.33 \times 0.66 + 0.66 \times 0.33$$

$$= 0.43$$

Right Subtree

$$\text{Impurity} = P(D) * (1 - P(D)) + P(U) * (1 - P(U))$$

$$= \frac{3}{4} \times \left(1 - \frac{3}{4}\right) + \frac{1}{4} \times \left(1 - \frac{1}{4}\right) = \frac{3}{4} \times \frac{1}{4} + \frac{1}{3} \times \frac{3}{4}$$

$$= 0.75 \times 0.25 + 0.25 \times 0.75$$

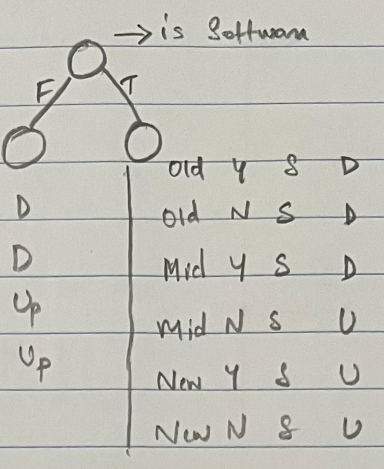
$$= 0.375$$

$$\text{Avg Imp} = \frac{6}{10} \times 0.43 + \frac{4}{10} \times 0.375$$

$$= 0.258 + 0.15$$

$$= 0.40$$

Information Gain = Avg. Imp (Root) − Avg. Imp (level 1)
2 Leads = 0.5 − 0.40
= 0.1



If "Type is Software"

Level 1: Root

Impurity = 0.5

Avg. Imp = 0.5

level 2:

Left Subtree

$$\text{Impurity} = P(D)*(1-P(D)) + P(U)\times(1-P(U)) = \frac{2}{4}\times\left(1-\frac{2}{4}\right)+\frac{2}{4}\times\left(1-\frac{2}{4}\right) = \frac{2}{4}\times\frac{2}{4}+\frac{2}{4}\times\frac{2}{4} \approx 0.5$$

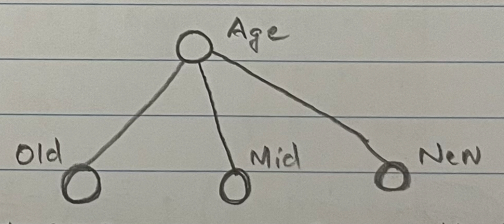
Right Subtree

$$\text{Impurity} = P(D)*(1-P(D)) + P(U)\times(1-P(U)) = \frac{3}{6}\times\left(1-\frac{3}{6}\right)+\frac{3}{6}\times\left(1-\frac{3}{6}\right) \approx 0.5$$

$$\text{Avg. Imp} = \frac{4}{10}\times 0.5 + \frac{6}{10}\times 0.5 = 0.4\times 0.5 + 0.6\times 0.5 = 0.06$$

Information Gain = 0.5 − 0.06 = 0.44

If "Age" feature



level Root

Impurity = 0.5 ; Avg Imp = 0.5

level 2

Left ⇒ $P(D)\times(1-P(D)) + P(U)\times(1-P(U))$

$= \frac{3}{3}\times\left(1-\frac{3}{3}\right)+0 = 0$

Mid ⇒ $\frac{2}{4}\times\left(1-\frac{2}{4}\right)+\frac{2}{4}\times\left(1-\frac{2}{4}\right) = 0.5$

Right ⇒ $0+\frac{3}{3}\times 1-\frac{3}{3} = 0$

Old: Y S D; N S D; N H D

Mid: Y S D; Y H D; N H Up; N S Up

New: Y S Up; N H Up; N S Up

$\text{Avg Imp} = \frac{3}{10}\times 0 + \frac{4}{10}\times 0.5 + \frac{3}{10}\times 0 = 0.2$

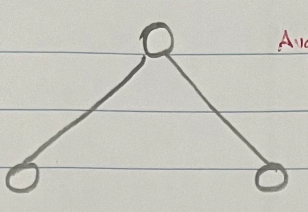
Information Gain = 0.5 − 0.2 = 0.30

With these 3 criterias Information gain is higher with the criteria "Type is Software". So we choose this decision tree for further split.

Level 1

Avg Imp = 0.5

Level 2



Old N H D
Mid Y H D
Mid N H Up
New N H Up

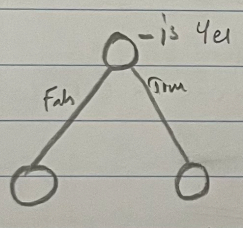
Old Y S D
Old N S D
Mid Y S D
Mid N S U
New Y S U
New N S U

Avg Imp = 0.06

Level 3

For level 3, If "Competion is Yes"

Left subtree



Old N H D
Mid N H Up
New N H Up

Mid Y H D

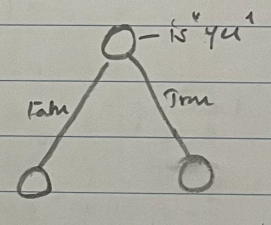
Left Subtree ⇒ $\text{Imp} = \frac{1}{3}\times\left(1-\frac{1}{3}\right)+\frac{2}{3}\times\left(1-\frac{2}{3}\right)$

$= 0.33\times 0.66 + 0.66\times 0.33 \approx 0.43$

Right Subtree ⇒ $\text{Imp} = \frac{1}{1}\times 1-\frac{1}{1} = 0$

$\text{Avg Imp} = \frac{3}{10}\times 0.43 = 0.129$

Right subtree.



Old N S D
Mid N S Up
New N S Up

Old Y S D
Mid Y S D
New Y S Up

Left Subtree ⇒ $\frac{1}{3}\times\left(1-\frac{1}{3}\right)+\frac{2}{3}\times\left(1-\frac{2}{3}\right) \approx 0.43$

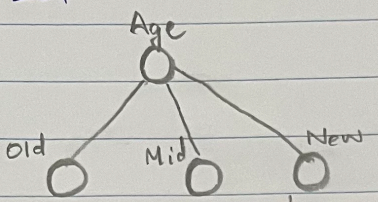
Right Subtree ⇒ $\frac{2}{3}\times\left(1-\frac{2}{3}\right)+\frac{1}{3}\times\left(1-\frac{1}{3}\right) \approx 0.43$

$\text{Avg Imp} = \frac{3}{10}\times 0.43 + \frac{3}{10}\times 0.43 = 0.258$

Information gain for 3 levels = 0.5 − 0.06 − 0.387 = 0.053

level 3, if "Age" feature

LEFT TREE Level 2



Level 3

Old N H D | Mid Y H D | New N H Up

Mid N H Up

Old $\Rightarrow P(D) \times |1 - P(D)| = \frac{1}{1} \times \left(1 - \frac{1}{1}\right) = 0$

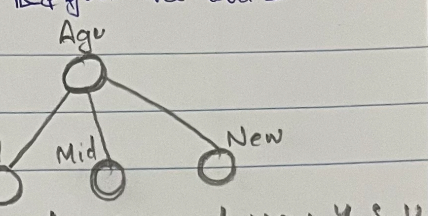
Mid $\Rightarrow P(D) \times (1 - P(D)) + P(U) \times (1 - P(U))$

$= 0$

New $\Rightarrow P(D) \times (1 - P(U)) = \frac{1}{1} \times 1 - \frac{1}{1} = 0$

Avg Imp = 0 + 0 + 0 = 0

Right tree level 2



Level 3

Old Y S D | Mid Y S D | New Y S U

Old N S D | Mid N S U | New N S U

Old $\Rightarrow \frac{2}{2} \times \left(1 - \frac{2}{2}\right) = 0$

Mid $\Rightarrow \frac{1}{2} \times \left(1 - \frac{1}{2}\right) = 0.5 \times 0.5 = 0.25$

New $\Rightarrow \frac{2}{2} \times \left(1 - \frac{2}{2}\right) = 0$

Avg Imp = 0 + 0.25 + 0 = 0.25

Information gain for 3 levels = 0.5 − 0.06 − 0.25

= 0.19

From this calculation the following decisions are best suitable for tree:

Level 1 Root + level 2 = Type is "Software"

level 3 = Use feature "Age"; ie Old, Mid, New

So, the profit value for : Mid No Software Up