# Reading Assignment

## Question 1

(a) $P_+ = \frac{120}{200} = \frac{3}{5}$

$P_- = \frac{80}{200} = \frac{2}{5}$

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

$= 1 - \frac{9}{25} - \frac{4}{25}$

$= \frac{12}{25} = \boxed{0.48}$

(b) Left (50 positive, 10 negative) $\rightarrow$ 60 samples

Right (70 positive, 70 negative) $\rightarrow$ 140 samples

$\text{Gini}_{left} = 1 - \frac{25}{36} - \frac{1}{36} = \frac{10}{36} = 0.2778$

$\text{Gini}_{right} = 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2} = 0.5$

$\text{Gini}_{weighted} = \frac{60 \times \frac{10}{36} + 140 \times \frac{1}{2}}{200} = \frac{13}{30} = 0.4333$

Since Gini index decreased after the split, purity is improved

## Question 2

(a) Possible splits : 1.5, 2.5, 3.5, 4.5, 6.5, 7.5

| Split Point | Left set | Right set | SSE |
|---|---|---|---|
| 1.5 | 10 | 12, 15, 18, ..., 30 | 0 + 271.45 = 271.45 |
| 2.5 | 10, 12 | 15, 18, ..., 30 | 2 + 170.83 = 172.83 |
| 3.5 | 10, 12, 15 | 18, 21, 25, 28, 30 | 12.67 + 97.2 = 109.87 |
| 4.5 | 10, 12, 15, 18 | 21, 25, 28, 30 | 36.75 + 46 = 82.75 |
| 5.5 | 10, 12, 15, 18, 21 | 25, 28, 30 | 78.8 + 12.67 = 91.47 |
| 6.5 | 10, 12, 15, 18, 21, 25 | 28, 30 | 158.83 + 2 = 160.83 |
| 7.5 | 10, 12, 15, 18, 21, 25, 28 | 30 | 265.94 |

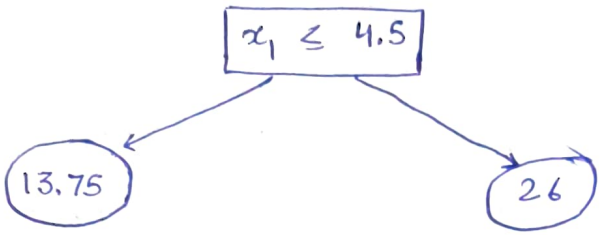
Best split : $x_1 \leq 4.5 \Rightarrow$ SSE = 82.75

(b) Left node : $x_1 \leq 4.5$ Mean, $\bar{y}_{LHS} = \frac{10+12+15+18}{4} = 13.75$

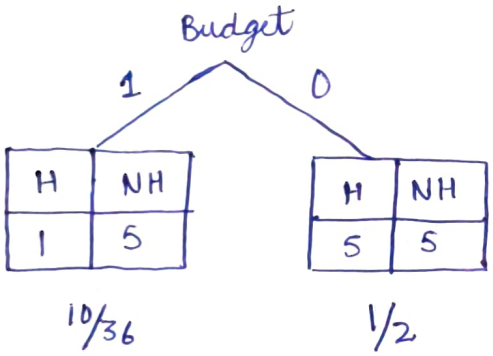
$$SSE = 36.75$$

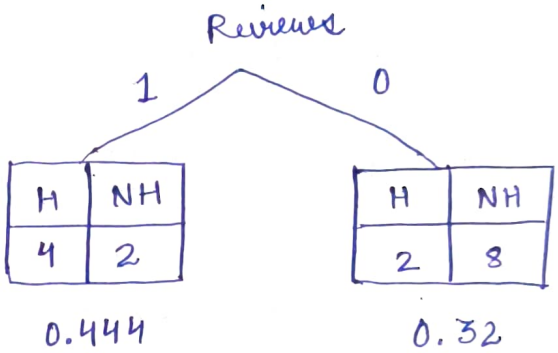
Right node : $x_1 > 4.5$ Mean, $\bar{y}_{RHS} = \frac{21+25+28+30}{4} = 26$

$$SSE = 46$$




## Question 3

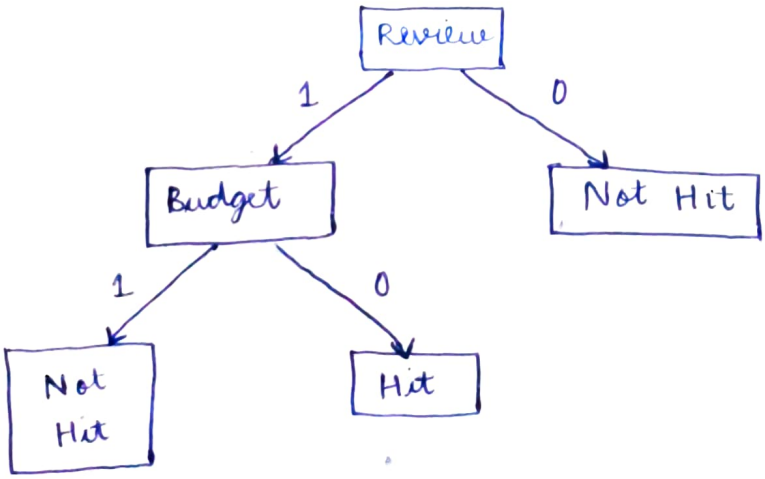



$Gini_{weighted} = 0.4687$

$Gini_{weighted} = 0.366 < Gini_{weighted}$ (Budget)

Final Tree



## Question 4

(a) For whole dataset, $H(x) = -\left(\frac{6}{10}\log_2\frac{6}{10} + \frac{4}{10}\log_2\frac{4}{10}\right) = 0.971$

(i) Attribute : Price

| Low | Med | High |
|---|---|---|
| 4 samples → 3 yes, 1 no | 4 → 1 yes, 3 no | 2 → 2 yes |
| Entropy $= -\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right)$ $= 0.811$ | Entropy $= 0.811$ | Entropy $= 0$ |

Weighted entropy, $E_{price} = \frac{4}{10} \times 0.811 + \frac{4}{10} \times 0.811 + \frac{2}{10} \times 0 = 0.6488$

Information Gain, $IG_{price} = 0.971 - 0.6488 = 0.3222$

(ii) Attribute : Maintenance

| Low | Med | High |
|---|---|---|
| 2 → 2 yes | 4 → 2 yes, 2 no | 4 → 2 yes, 2 no |
| Entropy $= 0$ | Entropy $= 1$ | Entropy $= 1$ |

$E_{maintenance} = \frac{2}{10} \times 0 + \frac{4}{10} \times 1 + \frac{4}{10} \times 1 = 0.8$

$IG_{maintenance} = 0.971 - 0.8 = 0.171$

(iii) Attribute : Capacity

| 2 | 4 | 5 |
|---|---|---|
| 2 → 2 yes | 6 → 3 yes, 3 no | 2 → 1 yes, 1 no |
| Entropy $= 0$ | Entropy $= 1$ | Entropy $= 1$ |

$E_{capacity} = \frac{2}{10} \times 0 + \frac{6}{10} \times 1 + \frac{2}{10} \times 1 = 0.8$

$IG_{capacity} = 0.971 - 0.8 = 0.171$

(iv) Attribute : Airbag

Yes

5 → 2 yes, 3 no

Entropy = 0.971

No

5 → 4 yes, 1 no

Entropy = 0.722

$E_{airbag} = \frac{5}{10} \times 0.971 + \frac{5}{10} \times 0.722 = 0.8465$

$IG_{airbag} = 0.971 - 0.8465 = 0.1245$

Best attribute to split on → Max$^m$ IG

⇒ Price is the root node

(b) To decide next split, we will

- recalculate entropy with each child node
- calculate IG with remaining attributes
- choose the one with highest IG as next split

Group 1 : Price = Low

Profitable : 3 yes, 1 no → Entropy = 0.811

Maintenance → Low : 2 yes → Entropy = 0

→ Med : 0 yes → Entropy = 0

→ High : 1 yes → Entropy = 0

IG = 0.811

Best split → Maintenance

Likewise,

Group 2 : Price = Med

Profitable : H(x) = 0.811

$IG_{airbag} = 0.311$ $IG_{maintenance} = 0.311$

Best split → Either airbag or maintenance

Group 3 : Price = High

All are profitable , Entropy = 0

No split required