1. (a.) Entropy $= -\sum_i p_i \log_b(p_i)$

Class 0 : $P(0) = \frac{1}{2}$

Class 1 : $P(1) = \frac{1}{2}$

Entropy $= -P(1)\log_2(P(1)) - P(0)\log_2(P(0))$

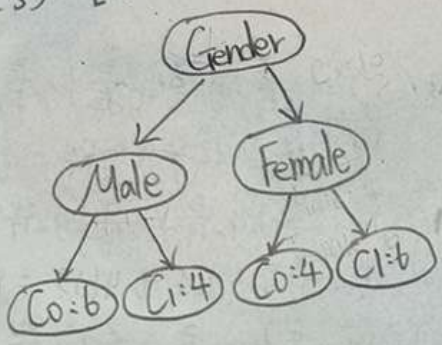
$= -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2})$

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

$= 1$ **

(b.) Information gain (Gender)

= Entropy (S) - [(Male/S) Entropy (Male) + (Female/S) Entropy (Female)]



Entropy (Male) $= -\frac{6}{10}\log_2(\frac{6}{10}) - \frac{4}{10}\log_2(\frac{4}{10})$

$= (0.6 \times -0.737) - (0.4 \times -1.321)$

$\fallingdotseq 0.4422 + 0.5284 = 0.9706$

Entropy (Female) $= -\frac{4}{10}\log_2(\frac{4}{10}) - \frac{6}{10}\log_2(\frac{6}{10})$

$\fallingdotseq 0.9706$

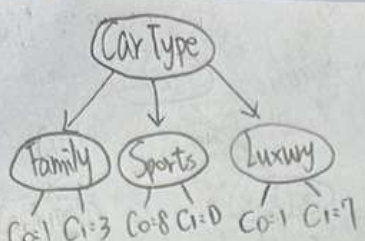
=> Entropy (S) - [(Male/S) Entropy (Male) + (Female/S) Entropy (Female)

=> $1 - [\frac{10}{20} \cdot 0.9706 + \frac{10}{20} \cdot 0.9706]$

$= 1 - 0.9706$

$= 0.0294$ ** ∴ Gender 對 information gain 提升不高
不太能有效地區分data

(C.)



$\text{Entropy}(\text{Family}) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}$
$\fallingdotseq 0.5 + 0.31125 = 0.81125$

$\text{Entropy}(\text{Sports}) = -\frac{8}{8}\log_2\frac{8}{8} - 0 = 0$

$\text{Entropy}(\text{Luxury}) = -\frac{1}{8}\log_2\frac{1}{8} - \frac{7}{8}\log_2\frac{7}{8}$
$\fallingdotseq 0.375 + 0.16852 = 0.54352$

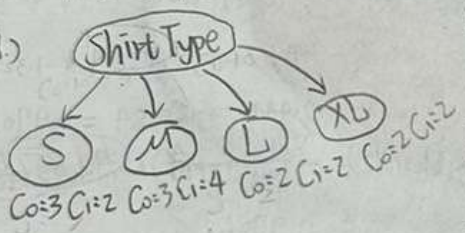
Information Gain (Car Type)

$= 1 - \left[\frac{4}{20}\cdot 0.81125 + \frac{8}{20}\cdot 0 + \frac{8}{20}\cdot 0.54352\right]$

$= 1 - [0.16225 + 0 + 0.21741]$

$= 0.62034$

(d.)



$\text{Entropy}(S) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$
$\fallingdotseq 0.4422 + 0.5284 = 0.9706$

$\text{Entropy}(M) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}$
$\fallingdotseq 0.5239 + 0.4613 = 0.9852$

$\text{Entropy}(L) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}$
$= \frac{1}{2} + \frac{1}{2} = 1$

$\text{Entropy}(XL) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$

Information Gain (Shirt Type)

$= 1 - \left[\frac{5}{20}\cdot 0.9706 + \frac{7}{20}\cdot 0.9852 + \frac{4}{20}\cdot 1 + \frac{4}{20}\cdot 1\right]$

$= 1 - [0.24265 + 0.34482 + 0.2 + 0.2] = 0.01253$

(e.)

比較各分法的 information gain 可得知

Gender = 0.0294
Car Type = 0.54352
Shirt Type = 0.01253

不太能有效區分 data

能有效區分 data

∴ Car Type is better. ※

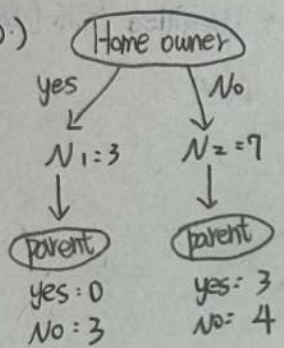
2.

Parent

No: 7
Yes: 3

(a.) Gini index = $1-(\frac{7}{10})^2-(\frac{3}{10})^2 = 0.42$ ※

(b.)



Node N_1 Gini index = $1-(\frac{0}{3})^2-(\frac{3}{3})^2 = 0$ ※

Node N_2 Gini index = $1-(\frac{3}{7})^2-(\frac{4}{7})^2 = \frac{24}{49} \fallingdotseq 0.49$ ※

Weighted Gini (N_1) = $\frac{3}{10} \cdot 0 = 0$

Weighted Gini (N_2) = $\frac{7}{10} \cdot 0.49 = 0.343$

Gini (Child) = Weighted Gini (N_1) + Weighted Gini (N_2)

= 0 + 0.343 = 0.343

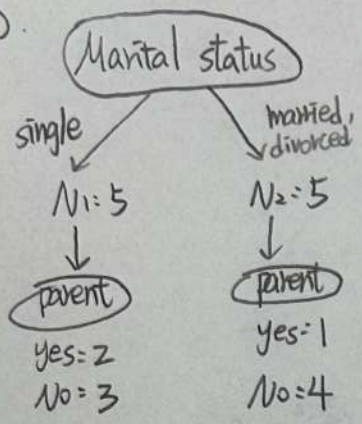
Gain in the Gini index

= Gini (Parent) − Gini (Child)

= 0.42 − 0.343 = 0.077 ※

∴ Home owner 不是很好的 attribute 方法，
因為 the gain in Gini index 不高，小於 0.2
所以沒有很好的區分 data. ※

(C.)



Node N_1 Gini index $= 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = 0.48$ ※

Node N_2 Gini index $= 1 - (\frac{1}{5})^2 - (\frac{4}{5})^2 = 0.32$ ※

Weighted Gini $N_1 = \frac{5}{10} \cdot 0.48 = 0.24$

Weighted Gini $N_2 = \frac{5}{10} \cdot 0.32 = 0.16$

Gini (Child) = Weighted Gini (N_1) + Weighted Gini (N_2)

$= 0.24 + 0.16 = 0.4$

Gain in the Gini index

= Gini (Parent) − Gini (Child)

= 0.42 − 0.4

= 0.02 ※

∴ Marital status 也不是很好的 attribute 方法，the gain in the Gini index 小於 0.2，所以沒有很好的區分 data. ※