SHIKHAR MOHAN
18EC10054

1. a) Entropy at root $E(S) = -0.5(\log 0.5 + \log 0.5)$

$= 1$

A1 (genders) $\longrightarrow$ [2+, 0−], [2+, 4−]

$$E(A_1) = \frac{-2}{8}(1\log 1) + \frac{-6}{8}\left(\frac{2}{6}\log\frac{2}{6} - \frac{4}{6}\log\frac{4}{6}\right)$$

$$= 0.688$$

$$Gain(S, A_1) = 0.311$$

Age group $\longrightarrow$ Attribute A2 $\longrightarrow$ [1+, 0−], [1+, 0−], [1+, 2−], [0+, 1−], [1+, 0−], [0+, 1−]

$$E(A_2) = \frac{3}{8}\left(\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right)$$

$$= 0.344$$

$$Gain(S, A2) = 1 - 0.344 = 0.656$$

A3 $\longrightarrow$ Pays by EMI $\longrightarrow$ [2+, 2−], [2+, 2−]

$$E(A3) = \frac{-2}{2}\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right)$$

$$= 1 \Rightarrow Gain(S, A3) = 0$$

Hence we split by A2 first.

$$E(A2) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}$$
$$= 0.918$$

} Entropy for node X, with age group 30-35

A1 ⟶ [0+, 0−], [1+, 2−]

$$E(A_1) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = E(A2)$$

$$\text{Gain}(X, A_1) = 0$$

A3 ⟶ [1+, 1−], [0+, 1−]

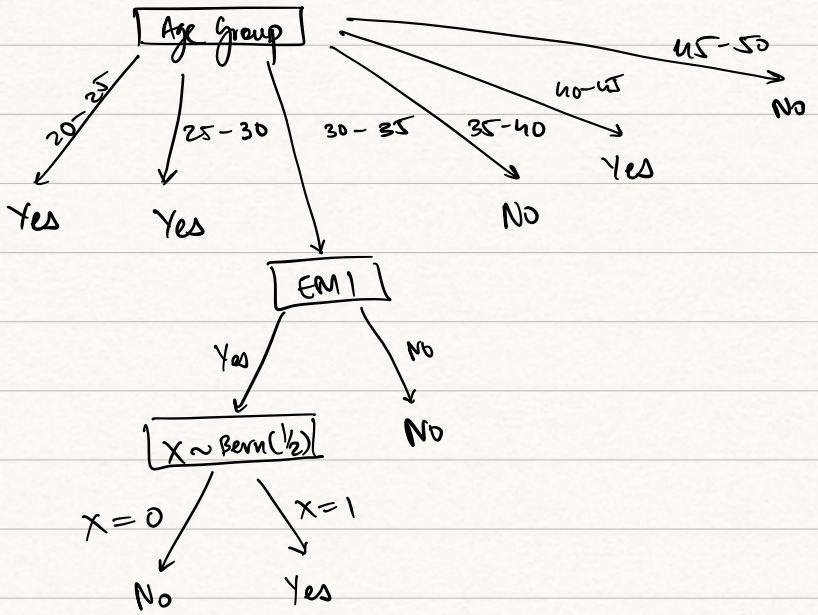
$$E(A_3) = \frac{2}{3}(1) = 0.667$$

$$\text{Gain}(X, A3) = 0.251$$

Next split is by A3

Any further splits don't improve classification, which means we assign a probability distribution to the final result. Since we find both 1+, 1− equiprobable, we assign a Bernoulli distribution with $p = 0.5$.

Final decision tree shown below



Predictions for :

i) (F, 25 – 30, Yes) ⟶ Yes

ii) (M, 35 – 40, No) ⟶ No

iii) (F, 45 – 50, Yes) ⟶ No

b. This is a bad decision tree. The degree of "Age - Group" node is too high, which results it being the first & last attribute. This makes other attributes useless since the learning now depends only on one attribute, which means due to this overdependence our model isn't robust either.

One solution might be to make larger age groups.

Secondly, the probabilistic nature of the deepest node

is a definite source of inaccuracies as well.

2. a) $SI(S, A1) = -\frac{2}{8}\log\frac{2}{8} - \frac{6}{8}\log\frac{6}{8} = 0.811$

$SI(S, A2) = 5(-\frac{1}{8}\log(\frac{1}{8})) - \frac{3}{8}\log\frac{3}{8}$

$= 2.405$

$SI(S, A3) = 1$

$GR(S, A1) = \frac{0.311}{0.811} = 0.383$

$GR(S, A2) = \frac{0.456}{2.405} = 0.278$

$GR(S, A3) = 0$

We split by A1 (gender) first: A1 $\longrightarrow$ [2+, 4-]

$E(X) = -\frac{2}{6}\log\frac{2}{3} - \frac{4}{6}\log\frac{4}{6} = 0.918$

A2 $\longrightarrow$ [0+, 0-], [1+, 0-], [1+, 2-], [0+, -], [0+, 0-], [0+, 1-]

$E(A2) = 3(0) + 2(0) + \frac{3}{6}(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3})$

$= 0.459$

$\therefore g(X, A_2) = 0.918 - 0.459$

$= 0.459$

$A_3 \longrightarrow [1+, 2-], [1+, 2-]$

$E(A_3) = 2 \times \left[\frac{3}{6}\left(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}\right)\right]$

$= 0.918$

$g(X, A_3) = 3 \Rightarrow GR(X, A_2) > GR(X, A_3)$

Splitting on age-group

We choose (F, 20–25) & (F, 40–45) as Yes since they offer perfect classification.

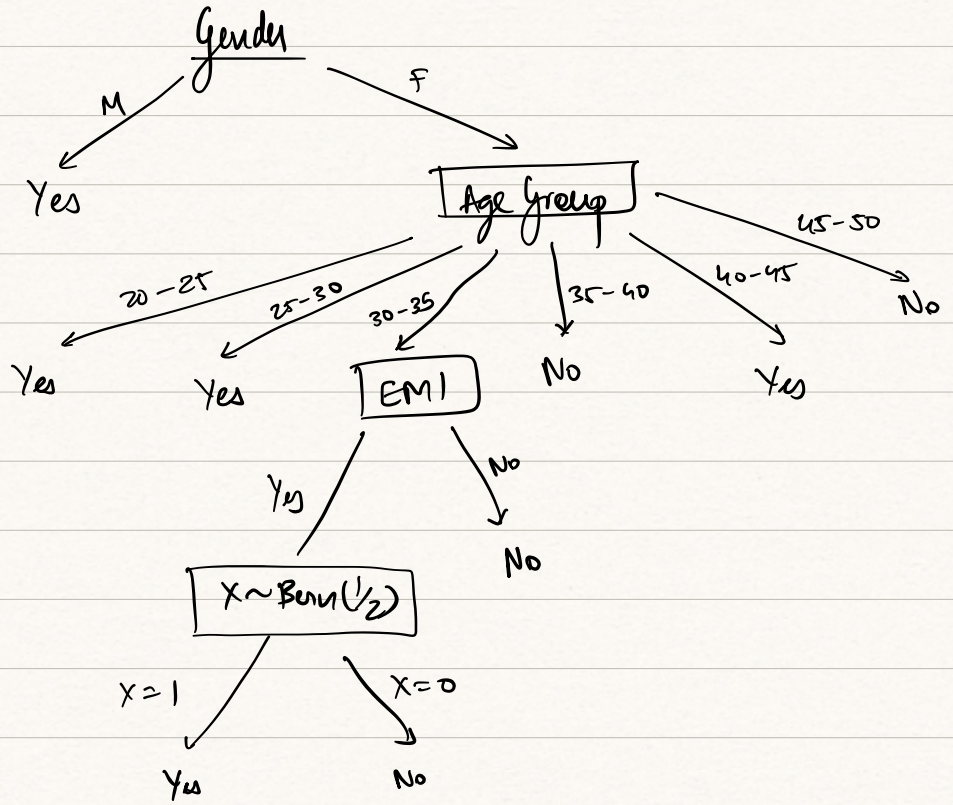
Predictions :

i) (F, 25-30, Yes) ⟶ Yes

ii) (F, 35–40, No) ⟶ Yes

iii) (F, 45-30, YES) ⟶ No

Final decision tree:



This new tree reduces the previously exaggerated importance of Age Group, leading to a more robust model. Datapoints like (M, 45-50, ?) are now classified to Yes instead of No.

b. For real data, the Gain Ratio tree will probably perform better compared to just information gain, since over-dependence on one attribute is

now reduced, making our model much more robust and generalized.

In general, information gain would be a better choice of criteria when all attribute have similar number of options. Gain ratio can be used in cases where the above isn't true (attribute space largely differing in dimensions) where the attribute with the largest dimensions (e.g. age group in the above) would dominate if Information Gain been used.