Data Mining
Assignment A4

Sindhuja Yerramalla
U00839259.

Question 1

(a) Entropy of Target Attribute "Buy Bitcoin"

$$Info(T) = -P_+ \log P_+ - P_- \log P_-$$

$$= -\frac{4}{8}\log\frac{4}{8} - \frac{4}{8}\log\frac{4}{8}$$

$$\boxed{Info(T) = 1}$$

for attribute CS-Major

Entropy of 'yes' given Target attribute (T)

$$Info(T_{yes}) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}$$

$$\boxed{Info(T_{yes}) = 1}$$

Entropy of "no" given Target attribute (T)

$$Info(T_{no}) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}$$

$$\boxed{Info(T_{no}) = 1}$$

Entropy of CS-Major given target as attribute (T)

$$\Rightarrow Info(CS\text{-}Major, T) = P_{yes} \times (Info(T_{yes})) + P_{no}(Info(T_{no}))$$

$$= \frac{4}{8}(1) + \frac{4}{8}(1) = 1$$

$$\boxed{Info(CS\text{-}Major, T) = 1}$$

Info Gain of 'cs-major' given Target attribute (T)

$$Gain(cs\text{-}major, T) = \frac{(Info(T) - Info(cs\text{-}major, T))}{Split\ Info(cs\text{-}major)}$$

$$Split\ Info(cs\text{-}major) = -\frac{4}{8}\log\frac{4}{8} - \frac{4}{8}\log\frac{4}{8}$$

$$\boxed{Split\ Info(cs\text{-}major) = 1}$$

$$Gain(cs\text{-}major, T) = \frac{1-1}{1} = 0 \text{ //}$$

For attribute Age

Entropy of "young" given Target attribute (T)

$$Info(T_{young}) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}$$

$$\boxed{Info(T_{young}) = 1}$$

Entropy of "old" given Target attribute (T)

$$Info(T_{old}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}$$

$$\boxed{Info(T_{old}) = 1}$$

Entropy of "middle" given Target attribute (T)

$$Info(T_{middle}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}$$

$$\boxed{Info(T_{middle}) = 1}$$

Entropy of "Age" given Target attribute (T).

$$\text{Info}(T_{age}) = P_{old}(\text{Info}(T_{old})) + P_{middle}(\text{Info}(T_{middle})) + P_{young}(\text{Info}(T_{young}))$$

$$= \frac{2}{8} + \frac{2}{8} + \frac{4}{8}$$

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

$\therefore$ Info $(T_{age}) = 1$

$$\text{Split Info}(T_{age}) = -\frac{2}{8}\log\frac{2}{8} - \frac{2}{8}\log\frac{2}{8} - \frac{4}{8}\log\frac{4}{8}$$

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

$$\boxed{\text{Split Info}(T_{age}) = 1.5}$$

$$\therefore \text{Gain}(Age, T) = \frac{(\text{Info}(T) - \text{Info}(Age, T))}{\text{Split Info}(Age)}$$

$$= \frac{1-1}{1.5} = 0$$

$$\boxed{\text{Gain}(Age, T) = 0}$$

For attribute "Income"

Entropy of "high", given target attribute (T).

$$\text{Info}(T_{high}) = -\frac{1}{1}\log\frac{1}{1} = 0$$

$$\boxed{\text{Info}(T_{high}) = 0}$$

Entropy of "fair" given target Attribute (T)

$$\text{Info}(T_{fair}) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}$$

$$= 0.311 + 0.5$$

$$\boxed{\text{Info}(T_{fair}) = 0.811}$$

Entropy of "low" given target attribute (T)

$$\text{Info}(T_{low}) = -\frac{3}{3}\log\frac{3}{3} = 0$$

$$\boxed{\text{Info}(T_{low}) = 0}$$

∴ Entropy of "Income" given target attribute (T)

$$\text{Info}(\text{Income}, T) = P_{fair}(\text{Info}(T_{fair})) + P_{high}(\text{Info}(T_{high})) + P_{low}(\text{Info}(T_{low}))$$

$$= \frac{4}{8} \times 0.811 + \frac{1}{8} \times 0 + \frac{3}{8} \times 0$$

$$\boxed{\therefore \text{Info}(\text{Income}, T) = 0.406}$$

$$\text{Split Info}(\text{Income}) = -\frac{4}{8}\log\frac{4}{8} - \frac{1}{8}\log\frac{1}{8} - \frac{3}{8}\log\frac{3}{8}$$

$$= 0.5 + 0.375 + 0.531$$

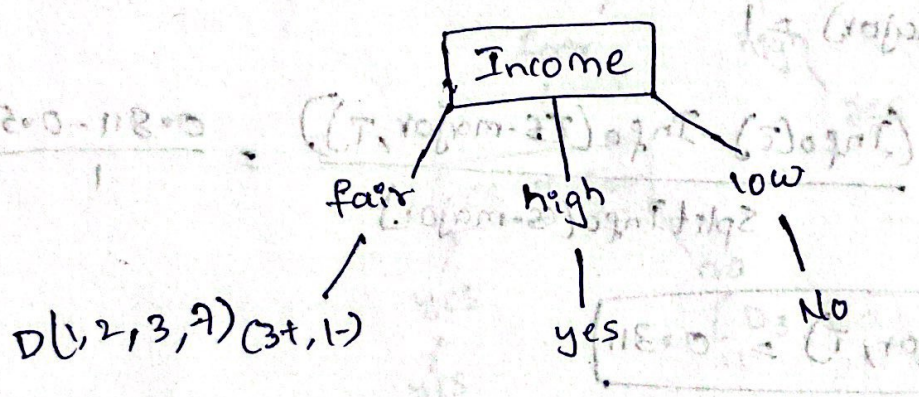
$$\boxed{\text{Split Info}(\text{Income}) = 1.406}$$

$$\text{Gain}(\text{Income}, T) = \frac{(\text{Info}(T) - \text{Info}(\text{Income}, T))}{\text{SplitInfo}(\text{Income})}$$

$$= \frac{1 - 0.406}{1.406}$$

$$\boxed{\text{Gain}(\text{Income}, T) = 0.422}$$

∴ Gain of (Income, T) is greater than other two attribute gains, so, Income will be taken as the root node of our decision Tree.



Now Consider Records D_1, D_2, D_3, D_4

$$\text{Info}(T) = -P_+ \log P_+ - P_- \log P_-$$

$$= -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}$$

$$= 0.811$$

$$\boxed{\therefore \text{Info}(T) = 0.811}$$

For attribute "cs-Major"

$$\text{Info}(T_{yes}) = -\frac{2}{2}\log\frac{2}{2} = 0$$

$$\text{Info}(T_{no}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$$

$$\text{Info}(cs\text{-major}, T) = \frac{2}{4}\times 0 + \frac{2}{4}\times 1 = 0.5$$

$$\text{SplitInfo}(cs\text{-major}) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1$$

$$\text{SplitInfo}(cs\text{-major}) = 1$$

$$\text{Gain}(cs\text{-major}, T) = \frac{(\text{Info}(T) - \text{Info}(T_{cs\text{-major}}, T))}{\text{SplitInfo}(cs\text{-major})} = \frac{0.811 - 0.5}{1}$$

$$\therefore \text{Gain}(cs\text{-major}, T) = 0.311$$

for attribute Age

Entropy of 'old' given Target attribute

$$\text{Info}(T_{old}) = -\frac{1}{1}\log\frac{1}{1} = 0$$

$$\text{Info}(T_{young}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$$

$$\text{Info}(T_{middle}) = -\frac{1}{1}\log\frac{1}{1} = 0$$

$$\Rightarrow \text{Info}(age, T) = \frac{1}{4}\times 0 + \frac{1}{4}\times 0 + \frac{2}{4}\times 1 = 0.5$$

$$\text{Split Info}(Age) = -\frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} - \frac{2}{4}\log\frac{2}{4}$$

$$= 1.5$$

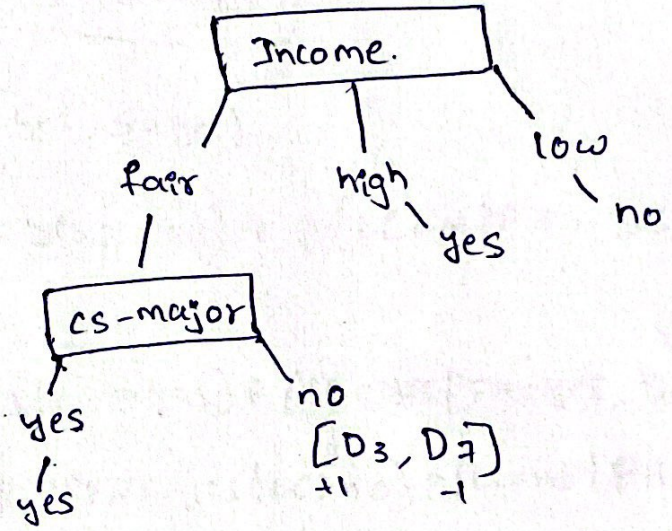
$$Gain(Age, T) = (Info(T) - Info(Age, T)) / SplitInfo(Age)$$

$$= \frac{0.811 - 0.5}{1.05} = 0.2073$$

Gain (Age, T) = 0.2073

Since. Gain (cs-major, T) is greater than Gain(age, T), cs-major will be the next node in the tree.

Decision tree :-



b) The question is asking about the user who is studying cs-major and his income is fair. As per the decisiontree we got in the previous solution, The user will buy bitcoin.

Question 2

a) $$P(LC = yes) = \sum_{x \in (yes,no)} \sum_{y \in (yes,no)} P(LC = yes \mid FH = x, S = y) * P(FH = x, S = y)$$

$$= P(LC = yes \mid FH = yes, S = yes) * P(FH = yes, S = yes) + P(LC = yes \mid FH = yes, S = no) * P(FH = yes, S = no) +$$

$P(LC=yes \mid FH=no, S=yes) * P(FH=no, S=yes) +$

$P(LC=yes \mid FH=no, S=no) * P(FH=no, S=no)$

$\therefore P(LC=yes) = 0.7 \times 0.3 \times 0.6 + 0.45 \times 0.3 \times 0.4 +$

$0.55 \times 0.7 \times 0.6 + 0.2 \times 0.7 \times 0.4$

$= 0.467$

$\therefore P(LC=yes) = 0.467$

b) $P(PR=yes \mid FH=yes, S=yes)$

$= \sum_{x \in (yes,no)} P(PR=yes \mid LC=x) * P(LC=x \mid FH=yes, S=yes)$

$= P(PR=yes \mid LC=yes) P(LC=yes \mid FH=yes, S=yes)$

$+ P(PR=yes \mid LC=no) P(LC=no \mid FH=yes, S=yes)$

$= 0.85 \times 0.7 + 0.45 \times 0.3 = 0.595 + 0.135$

$= 0.73$

$\therefore P(PR=yes \mid FH=yes, S=yes) = 0.73$

c) $P(LC=yes \mid PR=yes, FH=yes, S=yes)$

$= \frac{P(PR=yes \mid LC=yes, FH=yes, S=yes)}{P(PR=yes \mid FH=yes, S=yes)} * P(LC=yes \mid FH=yes, S=yes)$

$$= \frac{P(PR=yes|LC=yes)\,P(LC=yes|FH=yes, S=yes)}{\sum_{x\in(yes,no)} P(PR=yes|LC=x)\,P(LC=x|FH=yes, S=yes)}$$

$$= \frac{0.85 \times 0.7}{0.85 \times 0.7 + 0.45 \times 0.3}$$

$$= \frac{0.595}{0.73} = 0.815$$

⇒ The new person is more likely to have lung cancer.