Q1. (10 *points* in total)

(a)

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

For attribute "CS_Major":

$$\begin{aligned}
Info(T_{no}) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1 \\
Info(T_{yes}) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1 \\
Info(CS_Major, T) &= \frac{1}{2}Info(T_{no}) + \frac{1}{2}Info(T_{yes}) = 1 \\
SplitInfo(CS_Major) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1 \\
Gain(CS_Major, T) &= \frac{1-1}{1} = 0
\end{aligned}$$

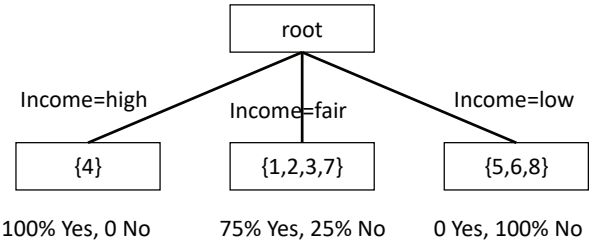
For attribute "Age":

$$\begin{aligned}
Info(T_{young}) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1 \\
Info(T_{middle}) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1 \\
Info(T_{old}) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1 \\
Info(Age, T) &= \frac{1}{2}Info(T_{young}) + \frac{1}{4}Info(T_{middle}) + \frac{1}{4}Info(T_{old}) = 1 \\
SplitInfo(Age) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} = 1.5 \\
Gain(Age, T) &= \frac{1-1}{1.5} = 0
\end{aligned}$$

For attribute "Income":

$$\begin{aligned}
Info(T_{high}) &= -1\log 1 - 0\log 0 = 0 \\
Info(T_{fair}) &= -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.8113 \\
Info(T_{low}) &= -0\log 0 - 1\log 1 = 0 \\
Info(Income, T) &= \frac{1}{8}Info(T_{high}) + \frac{1}{2}Info(T_{fair}) + \frac{3}{8}Info(T_{low}) = 0.405 \\
SplitInfo(Income) &= -\frac{1}{8}\log\frac{1}{8} - \frac{1}{2}\log\frac{1}{2} - \frac{3}{8}\log\frac{3}{8} = 1.4056 \\
Gain(Age, T) &= \frac{1-0.405}{1.4056} = 0.4233
\end{aligned}$$

Since "Income" gives the best gain, we choose it for splitting:



(2 *points* for choosing the correct attribute to split, 2 *points* for correct tree construction and node information)

Since the node for "Income=fair" does not meet the accuracy requirement, we need to further split it, we first compute the entropy of these four data record on "BuyBitcoin":

$$Info(T) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.8113$$

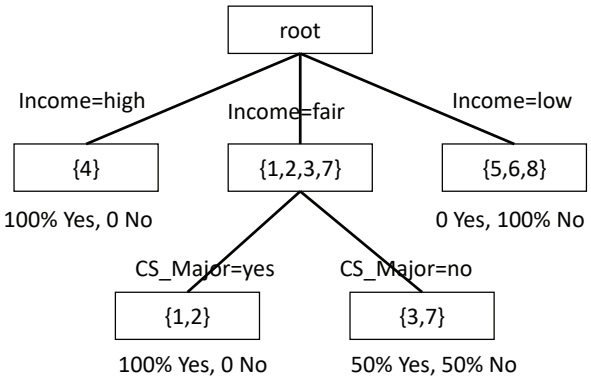
For attribute "CS_Major":

$$\begin{aligned}
Info(T_{no}) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1 \\
Info(T_{yes}) &= -1\log 1 - 0\log 0 = 0 \\
Info(CS_Major, T) &= \frac{1}{2}Info(T_{no}) + \frac{1}{2}Info(T_{yes}) = 0.5 \\
SplitInfo(CS_Major) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1 \\
Gain(CS_Major, T) &= \frac{0.8113 - 0.5}{1} = 0.3113
\end{aligned}$$

For attribute "Age":

$$\begin{aligned}
Info(T_{young}) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1 \\
Info(T_{middle}) &= -1\log 1 - 0\log 0 = 0 \\
Info(T_{old}) &= -1\log 1 - 0\log 0 = 0 \\
Info(Age, T) &= \frac{1}{2}Info(T_{young}) + \frac{1}{4}Info(T_{middle}) + \frac{1}{4}Info(T_{old}) = 0.5 \\
SplitInfo(Age) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} = 1.5 \\
Gain(Age, T) &= \frac{0.8113 - 0.5}{1.5} = 0.2075
\end{aligned}$$

Since $Gain(CS\_Major, T) > Gain(Age, T)$, we choose "CS_Major" for splitting, which results in the above decision tree that needs no further processing.



(2 *points* for choosing the correct attribute to split, 2 *points* for correct tree construction and node information)

Note *Binary logarithm is adopted for the computation above. You may have different information gain value if you use logarithm on a different base value. However, this should not affect the outcome of the decision tree.*

(b) It is very likely that this user will buy Bitcoin. (2 *points*)

Q2. (10 *points* in total)

(a)

$$P(LC = Yes) = \sum_{x \in \{Yes,No\}} \sum_{y \in \{Yes,No\}} P(LC = Yes|FH = x, S = y)P(FH = x, S = y) \text{ (2 points)}$$
$$= 0.7 \times 0.3 \times 0.6 + 0.45 \times 0.3 \times 0.4 + 0.55 \times 0.7 \times 0.6 + 0.2 \times 0.7 \times 0.4$$
$$= 0.467 \text{ (1 points)}$$

(b)

$$P(PR = Yes|FH = Yes, S = Yes)$$
$$= \sum_{x \in \{Yes,No\}} P(PR = Yes|LC = x)P(LC = x|FH = Yes, S = Yes) \text{ (2 points)}$$
$$= 0.85 \times 0.7 + 0.45 \times 0.3 = 0.73 \text{ (1 points)}$$

(c)

$$P(LC = Yes|FH = Yes, S = Yes, PR = Yes)$$
$$= \frac{P(PR = Yes|FH = Yes, S = Yes, LC = Yes)}{P(PR = Yes|FH = Yes, S = Yes)} \times P(LC = Yes|FH = Yes, S = Yes)$$
$$= \frac{0.85 \times 0.7}{0.73} = 0.8151$$
$$P(LC = No|FH = Yes, S = Yes, PR = Yes)$$
$$= 1 - P(LC = Yes|FH = Yes, S = Yes, PR = Yes) = 1 - 0.8151 = 0.1849$$

Since $P(LC = Yes|FH = Yes, S = Yes, PR = Yes)$ is greater than $P(LC = No|FH = Yes, S = Yes, PR = Yes)$, this model predicts that this person is very likely to have lung cancer. (4 *points*)

Note *If the conclusion of (c) is correct, full points should be granted. If the conclusion is wrong, showing specific steps for computing* $P(LC = Yes|FH = Yes, S = Yes, PR = Yes)$ *and* $P(LC = No|FH = Yes, S = Yes, PR = Yes)$, 2 *points can be granted.*