MSBD5002: Data Mining and Knowledge Discovery (MicroMasters)
Exercise 5 (Suggested Solution)
Classification

Q1

(a)

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

$$Info(T_{yes}) = -1log1 - 0log0 = 0$$

$$Info(T_{no}) = -\frac{1}{5}log\frac{1}{5} - \frac{4}{5}log\frac{4}{5} = 0.7219$$

$$Info(Child, T) = \frac{3}{8}Info(T_{yes}) + \frac{5}{8}Info(T_{no}) = 0.4512$$

$$SplitInfo(Child) = -\frac{3}{8}log\frac{3}{8} - \frac{5}{8}log\frac{5}{8} = 0.9544$$

$$Gain(Child, T) = \frac{Info(T) - Info(Child, T)}{SplitInfo(Child)} = \frac{1 - 0.4512}{0.9544} = 0.5750$$

(b)

$$Info(T) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$Info(T_{yes}) = 1 - 1^2 - 0^2 = 0$$

$$Info(T_{no}) = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 = 0.32$$

$$Info(Child, T) = \frac{3}{8}Info(T_{yes}) + \frac{5}{8}Info(T_{no}) = 0.2$$

$$Gain(Child, T) = Info(T) - Info(Child, T) = 0.5 - 0.2 = 0.3$$

Q2

$$P(HD = Yes \mid E = Yes, BP = High, CP = Yes)$$

$$= \frac{P(HD = Yes,\ BP = High, CP = Yes \mid E = Yes)}{P(BP = High, CP = Yes | E = Yes)}$$

$$= \frac{P(BP = High, CP = Yes \mid HD = Yes, E = Yes) \times P(HD = Yes \mid E = Yes)}{\sum_{x \in \{Yes,No\}} P(BP = High, CP = Yes \mid HD = x, E = Yes) \times P(HD = x \mid E = Yes)}$$

$$= \frac{P(BP = High, CP = Yes \mid HD = Yes) \times P(HD = Yes \mid E = Yes)}{\sum_{x \in \{Yes,No\}} P(BP = High, CP = Yes \mid HD = x) \times P(HD = x \mid E = Yes)}$$

$$= \frac{P(BP = High \mid HD = Yes) \times P(CP = Yes \mid HD = Yes) \times P(HD = Yes \mid E = Yes)}{\sum_{x \in \{Yes,No\}} P(BP = High | HD = x) \times P(CP = Yes \mid HD = x) \times P(HD = x \mid E = Yes)}$$

$$= \frac{0.85 \times 0.8 \times 0.25}{0.85 \times 0.8 \times 0.25 + 0.2 \times 0.6 \times 0.75}$$

$$= 0.65$$

$$P(HD = No \mid E = Yes, BP = High, CP = Yes)$$

$$= 1 - 0.65$$

$$= 0.35$$

It is more likely that he has heart disease.