1.a) total people = $65 + 35 = 100$

$$\text{Entropy } H\left(\frac{65}{100}, \frac{35}{100}\right) = -\frac{65}{100}\log_2 \frac{65}{100} - \frac{35}{100}\log_2\frac{35}{100}$$

$$H(0.65, 0.35) = +0.40397 + 0.5301$$
$$= 0.93407$$

b) wait –

$$H\left(\frac{25}{65}, \frac{40}{65}\right) = -\frac{25}{65}\log_2\frac{25}{65} - \frac{40}{65}\log_2\frac{40}{65}$$
$$= 0.53019 + 0.43104$$
$$= 0.961229$$

Not wait –

$$H\left(\frac{20}{35}, \frac{15}{35}\right) = -\frac{20}{35}\log_2\frac{20}{35} - \frac{15}{35}\log_2\frac{15}{35}$$
$$= 0.46135 + 0.52388$$
$$= 0.98523$$

$$I_A = H - \left(\frac{65}{100}(0.96123) + \frac{35}{100}(0.98523)\right)$$

$= 0.93407\,(0.62479 \quad + 0.34483)$
$= 0.93407\,(0.96963)$
$= 0.9057$

c) Information gain will be 0 because the exact same test is being used as node A. The attribute coming into node E will already be true so there will be no information gain.

d) From node A, it goes to node B since it is a Sunday. The patron is hungry so it goes from Node B to node D. The decision tree output for this case is will wait

2. <u>A as root</u>

$A = 1, \quad X = 3, \; Y = 0$
$A = 2, \quad X = 1, \; Y = 3$
$A = 3, \quad X = 1, \; Y = 2$
$\qquad\qquad\quad \overline{5} \qquad\quad \overline{5}$

$$H_A = -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10}$$

$$= 1$$

$$H_{A1} = -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}$$

$$= 0$$

$$H_{A2} = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}$$

$$= -\frac{1}{4}\times(-2) - \frac{3}{4}(-0.415037)$$

$$= 0.5 + 0.31128$$

$$= 0.81128$$

$$H_{A3} = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$$

$$= 0.52837 + 0.38997$$

$$= 0.91834$$

Information gain -

$$I_A = 1 - \frac{3}{10}(0) - \frac{4}{10}(0.81128) - \frac{3}{10}(0.91834)$$

$$= 1 - 0.3245 - 0.2755$$

$$= 0.399998$$

B as root

$B = 1$, $X = 1$, $Y = 3$

$B = 2$, $X = 3$, $Y = 1$

$B = 3$, $X = \frac{1}{5}$, $Y = \frac{1}{5}$

$H_B = 1$

$$H_{B1} = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}$$

$$= 0.81128$$

$$H_{B2} = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\cdot\log_2\frac{1}{4}$$

$$= 0.81128$$

$$H_{B3} = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$$

$$= 1$$

Information Gain

$$I_B = 1 - \frac{4}{10}(0.81128) - \frac{4}{10}(0.81128) - \frac{2}{10}(1)$$

$$= 1 - 0.324512 - 0.324512 - 0.2$$

$$= 0.1509$$

<u>C as root</u>

$C = 1$, $X = 1$, $Y = 4$

$C = 2$, $X = 3$, $Y = 1$

$C = 3$, $X = \frac{1}{5}$, $Y = \frac{0}{5}$

$$H_c = 1$$

$$H_{c1} = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}$$

$$= 0.46439 + 0.25754$$

$$= 0.72193$$

$$H_{c2} = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$$

$$= 0.81128$$

$$H_{c3} = -\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1}$$

$$= 0$$

Information Gain

$$I_C = 1 - \frac{5}{10}(0.72193) - \frac{4}{10}(0.81128) - \frac{1}{10}(0)$$

$$= 1 - 0.360965 - 0.324512 - 0$$

$$= 0.314523$$

Since I_A is greater than I_B and I_C, A provides the highest information gain at root. So A is the root.

3.a) lowest entropy possible is 0
highest entropy possible is $\log_2 4 = 2$

b) lowest information gained 0
highest information gained $\log_2 4 = 2$

4. With more and accurate data, the accuracy will increase. Ideally this should guarantee us to get a 60% accuracy. However, to assure accuracy is over 60%, we can flip the results to if home team loses or not. This would give us 72% accuracy.