Question 2

Part a

1) 4.0 GPA - P, the data match the decision tree
2) 3.9 GPA - P, the data match the decision tree
3) 3.9 GPA - P, the data match the decision tree
4) 3.8 GPA - yes publications - P, the data match the decision tree
5) 3.6 GPA - no publications - rank 2 university - P, the data match the decision tree
6) 3.6 GPA - yes publications - P, the data match the decision tree
7) 3.4 GPA - no publications - rank 3 university - N, the data match the decision tree
8) GPA 3.4 - No publication - Rank 1 University - N, data match the tree
9) GPA 3.2 - N, data match the tree
10) GPA 3.1 - N, data match the tree
11) GPA 3.1 - N, data match the tree
12) GPA 3.0 - N, data match the tree

Part b

$$E(S) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1 \tag{1}$$

For GPA, the information gained is:
GPA = 4.0: 3 P, 0 N

$$E(GPA = 4.0) = -1\log_2 1 - 0\log_2 0 = 0 \tag{2}$$

GPA = 3.6: 3 P, 2 N

$$E(GPA = 3.6) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.9710 \tag{3}$$

GPA = 3.3: 0 P, 4 N

$$E(GPA = 3.3) = -0\log_2 0 - 1\log_2 1 = 0 \tag{4}$$

$$I(GPA) = \frac{1}{4} * 0 + \frac{5}{12} * 0.9710 + \frac{1}{3} * 0 = 0.4046 \tag{5}$$

$$Gain(GPA) = E(S) - I(GPA) = 1 - 0.4046 = 0.5954 \tag{6}$$

For university rank, the information gained is:
University rank = 1: 3 P, 2 N

$$E(rank = 1) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.9710 \tag{7}$$

University rank = 2: 2 P, 1 N

$$E(rank = 2) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.9183 \tag{8}$$

University rank = 3: 1 P, 3 N

$$E(rank = 3) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.8113 \tag{9}$$

$$I(rank) = \frac{5}{12} * 0.9710 + \frac{1}{4} * 0.9183 + \frac{1}{3} * 0.8113 = 0.9046 \tag{10}$$

$$Gain(rank) = E(S) - I(rank) = 1 - 0.9046 = 0.0954 \tag{11}$$

For whether the student has publications, the information gained is:
Published = Yes : 3 P, 2 N

$$E(Published = Yes) = -\frac{3}{5}\log_2\frac{3}{5}) - \frac{2}{5}\log_2\frac{2}{5} = .9710 \tag{12}$$

Published = No : 3 P, 4 N

$$E(Published = No) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = .9852 \tag{13}$$

$$I(Published) = \frac{5}{12}(.9710) + \frac{7}{12}(.9852) = 0.9792 \tag{14}$$

$$Gain(Published) = E(S) - I(Published) = 1 - 0.9792 = 0.0208 \tag{15}$$

For the quality of the student's recommendations, the information gained is:
Recommendations = Good : 5 P, 3 N

$$E(Recommendations = Good) = -\frac{5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8} = .9544 \tag{16}$$

Recommendations = Normal : 1 P, 3 N

$$E(Recommendations = Normal) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = .8113 \tag{17}$$

$$I(Recommendations) = \frac{8}{12}(.9544) + \frac{4}{12}(.8113) = 0.9067 \tag{18}$$

$$Gain(Recommendations) = E(S) - I(Recommendations) = 1 - 0.9067 = 0.0933 \tag{19}$$

For the root of the decision tree, the best attribute to use is GPA, since it has the highest information gain of all the attributes. Because a GPA of 4.0 always leads to P and a GPA of 3.3 always leads to N, we only need to investigate the next node for when GPA = 3.6.

$$E(rank = 1) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{2}{5}\log_2\frac{2}{5} = 0.9710 \tag{20}$$

$$E(rank = 2) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.9183 \tag{21}$$

$$E(rank = 3) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.8113 \tag{22}$$

$$I(rank) = \frac{5}{12} * 0.9710 + \frac{1}{4} * 0.9183 + \frac{1}{3} * 0.8113 = 0.9046 \tag{23}$$

$$Gain(rank) = E(S) - I(rank) = 1 - 0.9046 = 0.0954 \tag{24}$$