Question 2

A)

|  |  |  |
| --- | --- | --- |
| No. | Table | Tree |
| 1 | P | P |
| 2 | P | P |
| 3 | P | P |
| 4 | P | P |
| 5 | P | P |
| 6 | P | P |
| 7 | N | N |
| 8 | N | N |
| 9 | N | N |
| 10 | N | N |
| 11 | N | N |
| 12 | N | N |

Tree correctly categorizes the given data.

B)

E(S) = -1/2 log2(1/2) - 1/2 log2(1/2) = 1

GPA

E(GPA >= 3.9) = -1 log2(1) – 0 log2(0) = 0

E(3.2 < GPA < 3.9) = -3/5 log2(3/5) – 2/5 log2(2/5) = 0.9710

E(GPA <= 3.2) = -0 log2(0) – 1 log2(1) = 0

I(GPA) = 1/4 \* 0 + 5/12 \* 0.9710 + 1/3 \* 0 = 0.4046

Gain(GPA) = E(S) – I(GPA) = 1 – 0.4046 = 0.5954

Published

E(Published = Yes) = -3/5 log2(3/5) – 2/5 log2(2/5) = 0.9710

E(Published = No) = -3/7 log2(3/7) – 4/7 log2(4/7) = 0.9852

I(Published) = 5/12 \* 0.9710 + 7/12 \* 0.9792 = 0.9792

Gain(Published) = E(S) – I(Published) = 1 – 0.9792 = 0.0954

Rank

E(rank = 1) = -3/5 log2(3/5) – 2/5 log2(2/5) = 0.9710

E(rank = 2) = -2/3 log2(2/3) – 1/3 log2(1/3) = 0.9183

E(rank = 3) = -1/4 log2(1/4) – 3/4 log2(3/4) = 0.8113

I(rank) = 5/12 \* 0.9710 + 1/4 \* 0.9183 + 1/3 \* 0.8113 = 0.9046

Gain(rank) = E(S) – I(rank) = 1 – 0.9046 = 0.0954

Recommendation

E(Recommendation = Good) = -5/8 log2(5/8) – 3/8 log2(3/8) = 0.9544

E(Recommendation = Normal) = -1/4 log2(1/4) – 3/4 log2(3/4) = 0.8113

I(Recommendation) = 8/12 \* 0.9544 + 4/12 \* 0.8113 = 0.9067

Gain(Recommendation) = E(S) – I(Recommendation) = 1 – 0. 9067 = 0.09033

After selecting GPA as the first attribute, we will see that (GPA <= 3.2) and (GPA >= 3.9) reaches to consistent answers of P and N in that order. For the middle group we find that having published prior research reaches to a P answer with 0.42 gain. Lastly, ranking will identify the remainder data with N,P,N as in the order of the ranks.

C)

Decision tree generated in part b is identical to the given decision tree. Since we used the most optimized way to set the tree in part b, we get the same result as the given example.