In Class Activity – Decision Trees (ICA 12) - Solutions

Please enter your responses at https://tinyurl.com/AIF19-ICA12

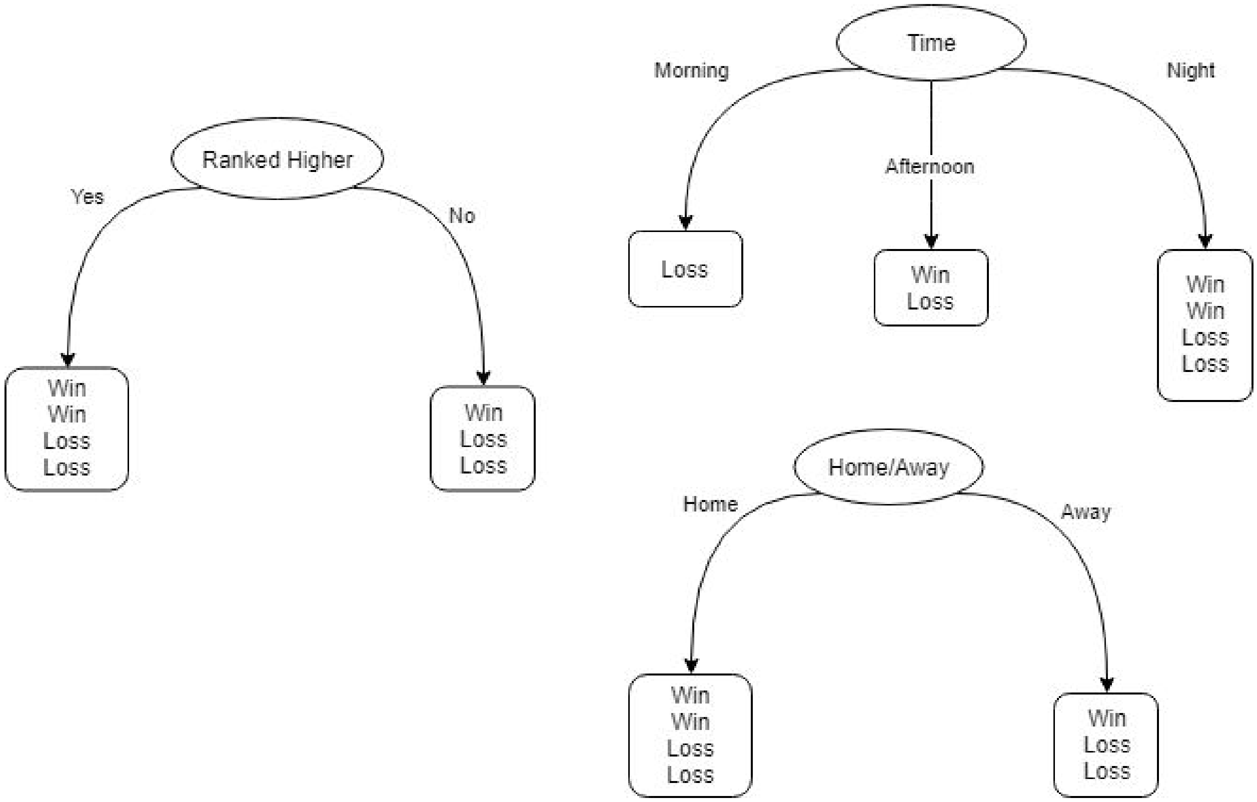
The table below shows the current results of the CU football team. The next game is with Washington State. Your goal is to predict the outcome using a Decision Tree.

CU Opponent	**Home/Away**	**Time**	**Ranked Higher**	**Win/Loss**
Colorado State	Home	Night	No	Win
Nebraska	Home	Afternoon	Yes	Win
Hollywood U	Home	Night	No	Loss
Air Force	Home	Morning	No	Loss
ASU	Away	Night	Yes	Win
Arizona	Away	Afternoon	Yes	Loss
Oregon	Away	Night	Yes	Loss
Washington St	*Away*	*Afternoon*	*No*	*?*

The entropy equation is:

$$H(S) = -\sum_{x \in X} p(x) \log_2 p(x)$$

The below diagram shows the possible decision tree first level splits for the CU football team data. It also may make the counting of data easier for the following questions.



Please enter your responses to the following questions on the Google form.

1. Base entropies:
   a. Calculate H(Home/Away)

Solution:

- Home, Home, Home, Home, Away, Away, Away
- Info = -4/7*log(4/7,2) -3/7*log(3/7,2) = 0.99 bits

   b. Calculate H(Time)

Solution:

- Night, Night, Night, Night, Afternoon, Afternoon, Morning
- Info = -4/7*log(4/7,2) - 2/7*log(2/7,2) - 1/7*log(1/7,2) = 1.38 bits

   c. Calculate H(Ranked Higher)

Solution:

- Yes, Yes, Yes, Yes, No, No, No
- Info = -4/7*log(4/7,2) - 3/7*log(3/7,2) = 0.99 bits

   d. Calculate H(Win/Loss)

Solution:

- Win, Win, Win, Loss, Loss, Loss, Loss
- Info = -3/7*log(3/7,2) - 4/7*log(4/7,2) = 0.99 bits

2. Calculate entropies given a column condition:
    a. Calculate H(Win/Loss | Home/Away)
        i. Home
            1. Win, win, loss, loss
            2. Info = -2/4*log(2/4,2)-2/4*log(2/4,2) = 1 bits
        ii. Away
            1. Win, loss, loss
            2. Info = -1/3*log(1/3,2)-2/3*log(2/3,2) = 0.92 bits
        iii. Avg (bits x occurrence)
            1. 1*4/7 + 0.92*3/7 = **0.97 bits**

    Splitting on Home/Away = 0.99 – 0.97 = info gain of 0.02 bits

    b. Calculate H(Win/Loss | Time)
        i. Morning
            1. Loss
            2. Info = -1/1*log(1/1,2) = 0 bits
        ii. Afternoon
            1. Win, Loss
            2. Info = -1/2*log(1/2,2)-1/2*log(1/2,2) = 1 bits
        iii. Night
            1. Win, Loss, Win, Loss
            2. Info = -2/4*log(2/4,2)-2/4*log(2/4,2) = 1 bits
        iv. Avg (bits x occurrence)
            1. 0*1/7 + 1*2/7 + 1*4/7 = **0.86 bits**

    Splitting on ToD = 0.99 – 0.86 = info gain of 0.13 bits

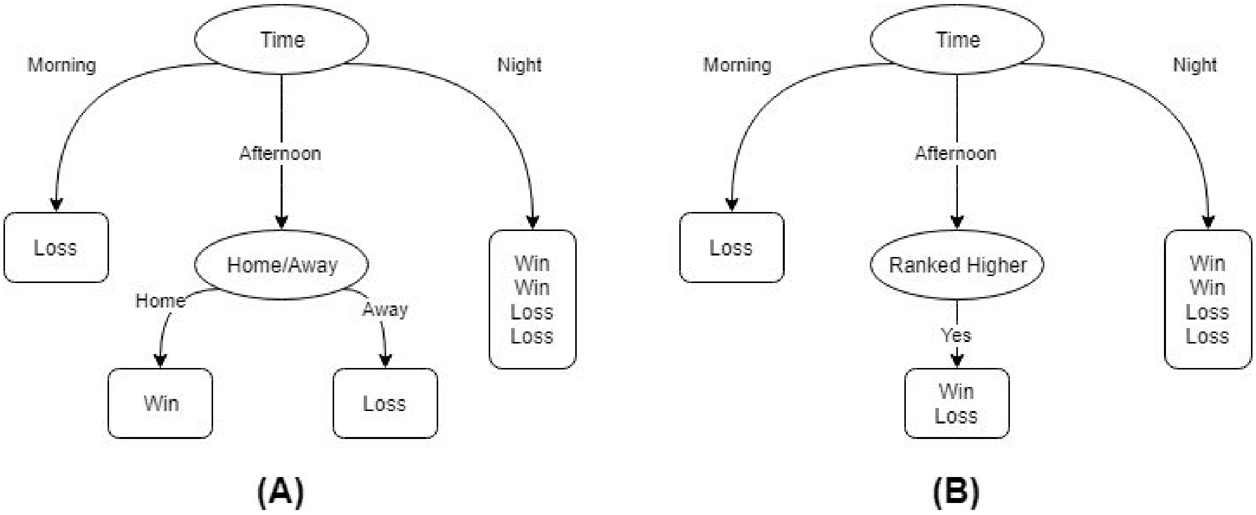
    c. Calculate H(Win/Loss | Ranked Higher)
        i. Yes
            1. Win, Win, Loss, Loss
            2. Info = -2/4*log(2/4,2)-2/4*log(2/4,2) = 1 bits
        ii. No
            1. Win, Loss, Loss
            2. Info = -1/3*log(1/3,2)-2/3*log(2/3,2) = 0.92 bits
        iii. Avg (bits x occurrence)
            1. 1*4/7 + 0.92*3/7 = **0.97 bits**

    Splitting on Ranking = 0.99 – 0.97 = info gain of 0.02 bits

3. Which node (Home/Away, Time, Ranked Higher) should you split the tree on?

    We choose to split on the node that yields the highest information gain, ie. Time of Day.

Submit your answers for the above questions on the google form to see the figure for the following two questions:
BELOW IMAGE IN SOLUTION AND GOOGLE FORM ONLY.



4. Intuitively, or with entropy and information gain calculations, should (A) or (B) be the next tree?

Solution: (A) since choosing that would lead to a deterministic, low entropy, tree. (B) would yield no new information.

5. Based on your choice in 4, how would the tree classify the outcome of the upcoming Washington State game?

Solution: It would classify CU as losing. ☹