ASSIGNMENT 1:

Hrudhay Reddy Garisa (577833)

Joshua Meyer (577355)

Edwin le Cointre (577328)

Bophelo Maroga (577651)

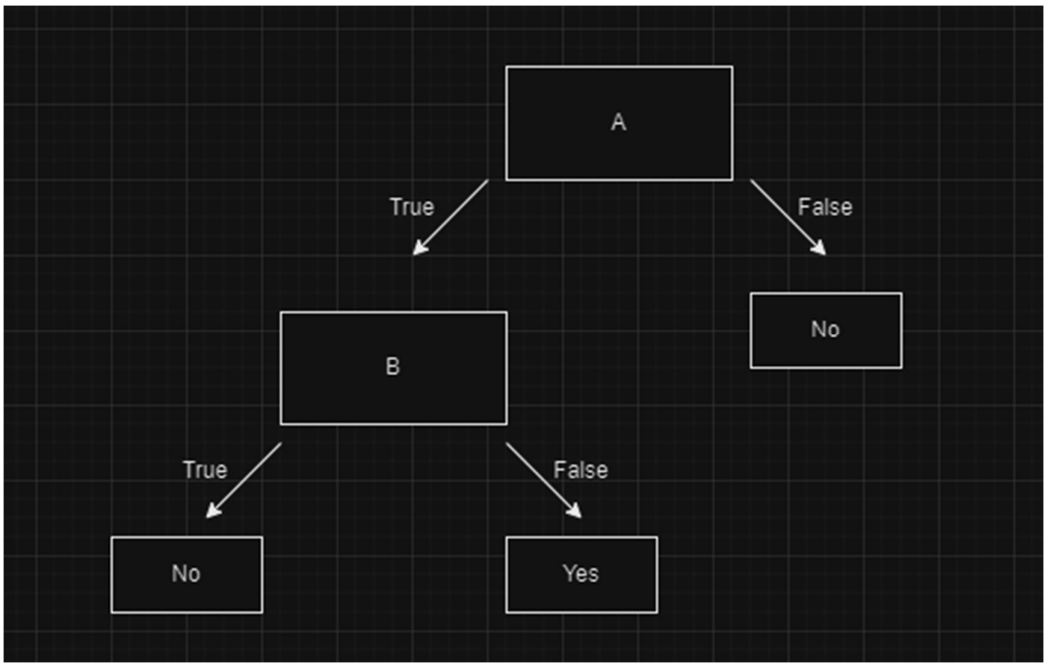
MLG381

Contents

Question 1:

a)

A ∧ ¬B



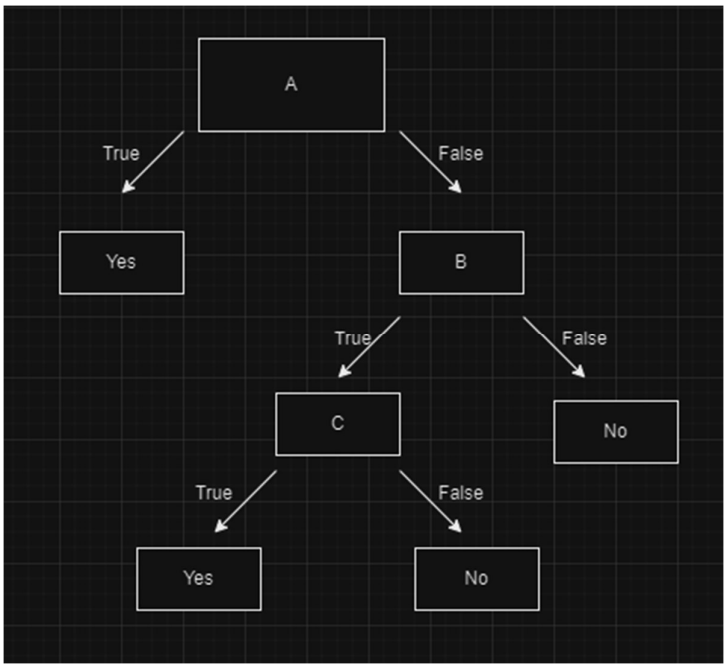
(Huddar, 2021)

Explanation A

1. Start with the root node, which tests the value of A
2. If A is true, we move to the right child node. If A is false, we move to the left child node.
3. At the right child node, we check the value of B. Since we want ¬B (not B), if B is false, this branch satisfies the condition.
4. At the left child node, no matter the value of B, it doesn't satisfy the condition since we need A to be true and B to be false.

b)

A ∨ [B ∧ C]



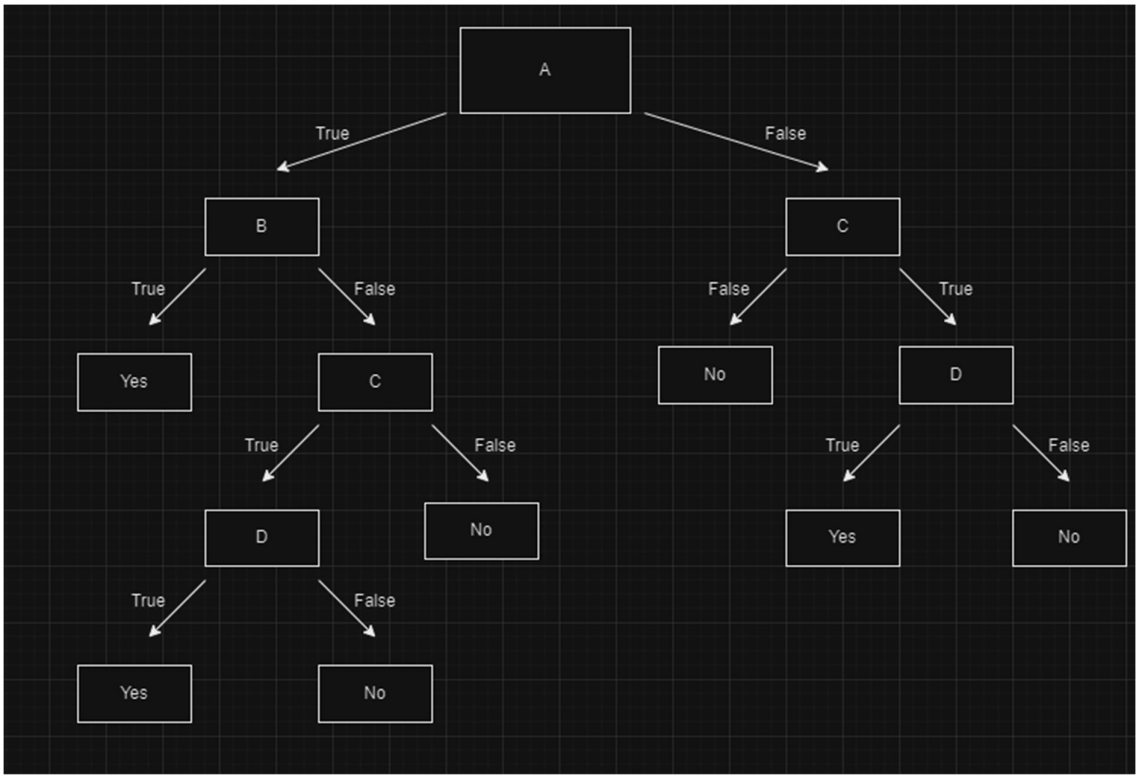
(Huddar, 2021)

Explanation B

1. Start with the root node, which tests the value of A.
2. If A is true, the condition is satisfied.
3. If A is false, we need to check the value of [B ∧ C].
4. For [B ∧ C] to be true, both B and C must be true.

C

[A ∨ B] ∧ [C ∨ D]



(Huddar, 2021)

Explanation C:

1. Start with the root node, which tests the value of [A ∨ B].
2. If [A ∨ B] is true, we move to the left child node.
3. If [A ∨ B] is false, we move to the right child node.
4. At the left child node, we test the value of [C ∨ D].
5. At the right child node, we don't need to test [C ∨ D] because [A ∨ B] is false, so the condition won't be satisfied regardless of [C ∨ D].

Question 2:

ID3 is an algorithm that focuses on the creation of decision trees from the top to the bottom with a greedy approach. Once the entropy is calculated and the gain is found, the highest gain is then used to figure out which path to take next. The process is continued until all attribute paths have been explored.

The decision tree seeks to find the optimal conditions for playing sports. This usually starts by calculating the entropy of single attributes. I will be using the first attribute in the table below which is the sky attribute to show how the calculations are done. The general method for this is –((Yes/Total) *LOG2(Yes/Total) +(No/Total) *LOG(No/total)). This is then used to calculate each attribute gain which is calculated by (Total Entropy-(attribute1 Total/total) *a1 Entropy-(Attribute2 Total/total) *a2 Entropy.

2					
Sky	Yes	No	Total	Entropy	Gain
Sunny	3	0	3	#NUM!	0.811278
Rainy	0	1	1	#NUM!	
Total	3	1	4	0.811278124	

AirTemp	Yes	No	Total	Entropy	Gain
Warm	3	0	3	#NUM!	0.811278
Cold	0	1	1	#NUM!	
Total	3	1	4	0.811278124	

Humidity	Yes	No	Total	Entropy	Gain
High	2	1	3	0.918295834	0.122556
Normal	1	0	1	#NUM!	
Total	3	1	4	0.811278124	

Wind	Yes	No	Total	Entropy	Gain
Strong	3	1	4	0.811278124	0
Total	3	1	4	0.811278124	

Water	Yes	No	Total	Entropy	Gain
Warm	2	1	3	0.918295834	0.122556
Cool	1	0	1	#NUM!	
Total	3	1	4	0.811278124	

Forecast	Yes	No	Total	Entropy	Gain
Same	2	0	2	#NUM!	0.311278
Change	1	1	2	1	
Total	3	2	4	0.811278124	

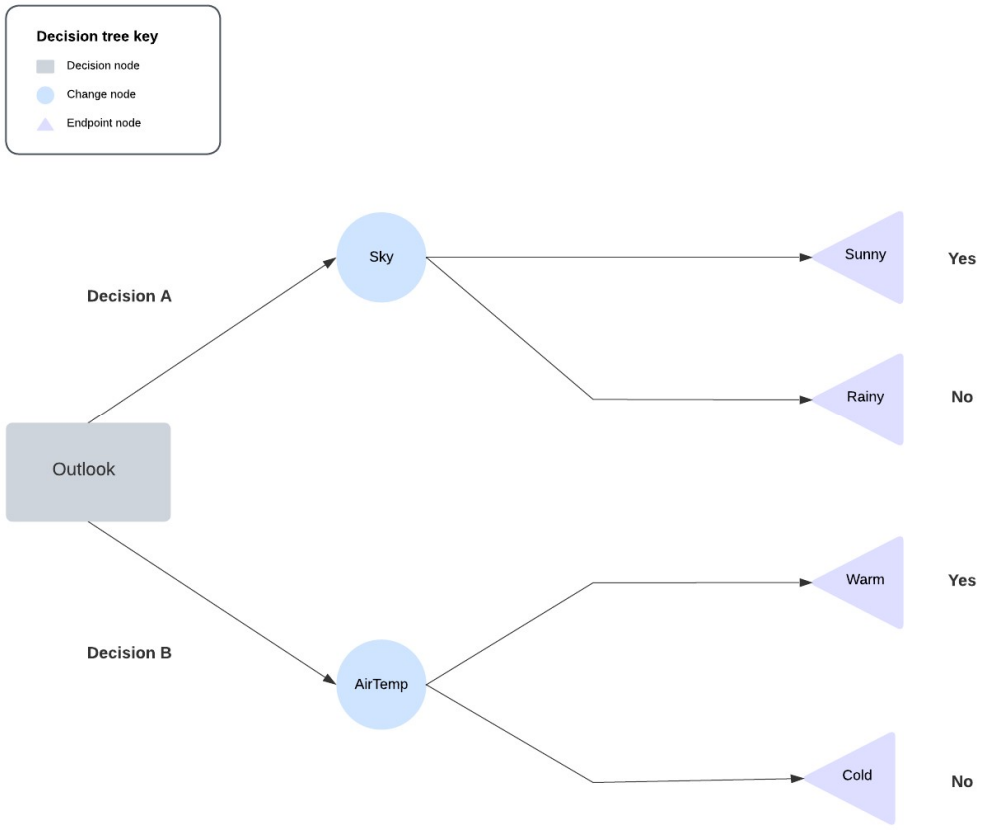
Once the gain is found the highest should be selected and deemed as the root node. In the example above there are two gain values that may qualify as the root node. The reason the highest value is chosen is that the lower the gain value the more is known on that attribute. However, the higher the gain, the more information is needed on that attribute.
After selecting the root node, you then proceed to split the root using its yes and no values if one value has all yes or all no then you can stop there as seen with Rainy Yes. In the case where the attribute has both yes and no, the attribute can then be broken down further into the next branch node.

Rainy = No					
Sky	Sunny				
Humidity	Yes	No	Total	Entropy	Gain
High	2	0	2	#NUM!	0
Normal	1	0	1	#NUM!	
Total	3	0	3	#NUM!	
Wind	Yes	No	Total	Entropy	Gain
Strong	3	0	3	#NUM!	0
Total	3	0	3	#NUM!	
Water	Yes	No	Total	Entropy	Gain
Warm	2	0	2	#NUM!	0
Cool	1	0	1	#NUM!	
Total	3	0	3	#NUM!	
Forecast	Yes	No	Total	Entropy	Gain
Same	2	0	2	#NUM!	0
Change	1	0	1	#NUM!	
Total	3	0	3	#NUM!	

Cold = No					
Air Temp	Warm				
Humidity	Yes	No	Total	Entropy	Gain
High	2	0	2	#NUM!	0
Normal	1	0	1	#NUM!	
Total	3	0	3	#NUM!	
Wind	Yes	No	Total	Entropy	Gain
Strong	3	0	0	0	0
Total	3	0	0	0	
Water	Yes	No	Total	Entropy	Gain

Warm	2	0	2	0	0
Cool	1	0	1	0	
Total	3	0	3	0	
Forecast	Yes	No	Total	Entropy	Gain
Same	2	0	2	0	0
Change	1	0	1	0	
Total	3	0	3	0	

After the super attributes of Sky and Airtemp were found the entropies and gains were calculated. Using the same formulas from before the entropy and gains of the branches from those super attributes were all found to be zero as seen above. Earlier it was said that the lower the gain the more information is known on the values. This means the calculations are done and the decision tree can go no further. If found to be sunny it will be an immediate yes to enjoying a sport, while if rainy it will be no. The same is found with cold and warm, with warm being yes and cold being no. The Decision tree is shown in the image below.



Question 3:

Question 3

1. Independent Variables:
    - Prevalence (Pr): Tells us the ratio the total population that are infected
    - Sensitivity (TPR): The ratio of positive tests that are accurately labelled as positive
    - Specificity (TNR): The ratio of negative tests that are accurately labelled as negative

    Dependent Variables:

    - True Positive (TP): The ratio of the total tests that are accurately labelled as positive.
    - True Negative (TN): The ratio of the total tests that are accurately labelled as negative
    - Positive Predictive Value (PPV): The probability of an infection given a positive test result. Depends on the sensitivity and specificity of the test, and the prevalence of the disease.
    - Negative Predictive Value (NPV): It measures the probability that someone who does not have the disease is given a negative test result, which is influenced by how the number of true negatives is correctly identified out of all negative results
    - Likelihood ratio (LR): The proportion of people with the Superbug with a given test result (either positive or negative) divided by the proportion of people without Superbug with that result

2. Prevalence = Number of people diagnosed with the Superbug / Total population tested
   Pr = 100/10 000
   Pr = 0.01 or 1%

3. $Sensitivity = \frac{TP}{TP+FN}$ (GeeksforGeeks, 2023) $Specificity = \frac{TN}{TN+FP}$ (GeeksforGeeks, 2023)
   $TPR = \frac{90}{90+10}$ $TNR = \frac{40}{40+10}$
   $TPR = 90/100$ $TNR = \frac{40}{50}$
   TPR =0.9 or 90% TNR = 0.8 or 80%

4. $FPR = \frac{FN}{FN+}$ (GeeksforGeeks, 2023) $FPR = \frac{FP}{FP+TN}$ (GeeksforGeeks, 2023)
   $FNR = \frac{10}{10+90}$ $FPR = \frac{10}{10+40}$
   $FNR = \frac{10}{100}$ $FPR = \frac{10}{50}$
   FNR = 0.1 or 10% FPR = 0.2 or 20%

5. Because sensitivity measures the ratio of positive tests correctly identified as positive, and FNR measures the ratio of positives tests incorrectly identified as negative. FNR can be seen as the complement of sensitivity, this means that if we know 1 then we can calculate the other, thus their sum must equal to 1.

   FPR + TPR = 1

   $$\frac{FN}{FN+TP} + \frac{TP}{TP+FN} = 1$$

   $$\frac{FN+TP}{FN+TP} = 1$$

   FPR + 0.9 = 1

   FPR = 1 – 0.9

   FPR = 0.1 or 10%

   Because specificity measures the ratio of negatives correctly identified as positive, and FPR measures the ratio of negative tests incorrectly identified as positive. FPR can be seen as the complement of specificity, this means that decreasing specificity will increase FNR, and vice versa.

   FPR + TNR = 1

   $$\frac{FP}{FP+TN} + \frac{TN}{TN+FP} = 1$$

   $$\frac{TN+FP}{TN+FP} = 1$$

   FPR + 0.8 = 1

   FPR = 1 – 0.8

   FPR = 0.2 or 20%

6. Probability of being infected and testing positive (PPV):

   $$PPV = \frac{TP}{TP+FP}$$ (Jayaswal, 2020)

   $$PPV = \frac{90}{90+1}$$

   PPV = 0.9 or 90%

   Probability of not being infected and testing negative (NPV):

   $$NPV = \frac{90TN}{TN+FN}$$ (Jayaswal, 2020)

   $$NPV = \frac{40}{40+1}$$

   NPV = 0.8 or 80%

   Likelihood ratio (The NNT, n.d.):

LR (+) = TPR/FPR
LR (+) = 0.9 / 0.2
LR (+) = 4.5

LR (-) = FNR/TNR
LR (-) = 0.1 / 0.8
LR (-) = 0.125

Bibliography

GeeksforGeeks, 2023. *Calculate Sensitivity, Specificity and Predictive Values in CAREt*. [Online]
Available at: https://www.geeksforgeeks.org/calculate-sensitivity-specificity-and-predictive-values-in-caret/
[Accessed 15 March 2024].

Huddar, M., 2021. *How to build a decision Tree for Boolean Function | Machine Learning by Mahesh Huddar*. [Online]
Available at: https://www.youtube.com/watch?v=fs0wsU2sSPQ
[Accessed 2024].

Huddar, M., 2021. *How to build a decision Tree for Boolean Function | Machine Learning by Mahesh Huddar mp4*. [Online]
Available at: https://www.youtube.com/watch?v=gn85J4U4pbw
[Accessed 2024].

Jayaswal, V., 2020. *Performance Metics: Confusion matrix, Precision, Recall, and F1 Score.*. [Online]
Available at: https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262?gi=70c966c11492
[Accessed 15 March 2024].

The NNT, n., n.d. *Diagnostics and Likelihood Ratios, Explained*. [Online]
Available at: https://thennt.com/diagnostics-and-likelihood-ratios-explained/
[Accessed 15 March 2024].