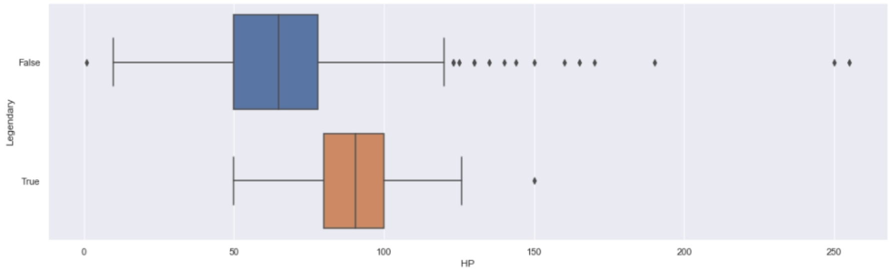
Quizzes : Module 4

Data-Driven Classification

In the LAMS Sequence, you have learned the theory behind this module. It is also expected that you have attempted the quizzes embedded within the LAMS Sequence, and have used the “unlimited attempts” opportunity to score 100%. Here are the quiz questions, consolidated with their answers and corresponding feedback. This is for your after-LAMS revision.

Question 1

Suppose that the boxplot of the HP (Hit Points) of Pokemons with respect to the categorical variable "Legendary" is as in the picture. Which of the following can you infer from this boxplot? Multiple options may be correct. Choose all options that seem right.



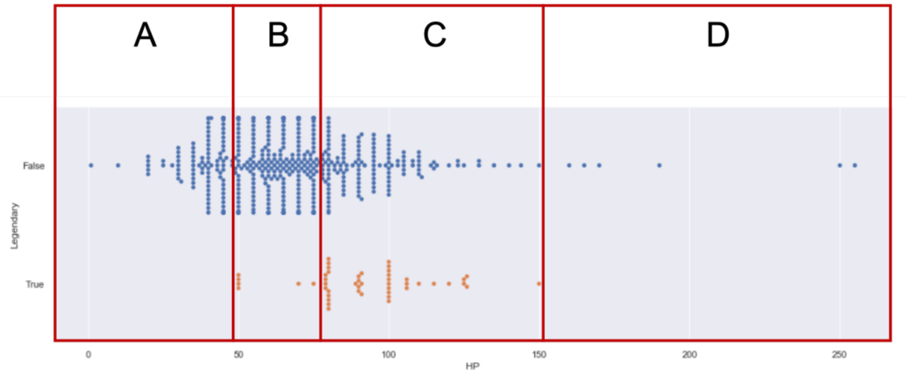
Answer Choice	Verdict	Explanation
HP (Hit Points) seems to be an important variable in predicting "Legendary".	Correct	True. The box plots for HP (Hit Points) are distinctly different for "Legendary" = True and "Legendary" = False. This indicates that HP may be an important differentiator between the two levels of the categorical variable "Legendary".
There is a strong relationship between the variables HP (Hit Points) and Legendary.	Correct	True. The box plots for HP (Hit Points) are distinctly different for "Legendary" = True and "Legendary" = False. If there was NO relation between the two, then the boxplots for HP across the various levels of "Legendary" would look quite similar.
There is a strong linear relationship between the variables HP (Hit Points) and Legendary.	Wrong	Nope. The boxplot does not suggest a linear relationship. In fact, a linear relationship is not even well-defined for a categorical variable against a continuous variable. Similarly, there is no notion of correlation between these two variables.

Reference Module 4 Topic 1 : Binary Classification Slide 5 and Slide 6

General Comment : Significant difference in the Box Plots of a continuous variable against multiple levels of a categorical variable tells us that the continuous variable is "important" in differentiating between the levels of the categorical variable. Hence, it is important for prediction.

Question 2

Suppose that the swarmplot of HP (Hit Points) against "Legendary" is as follows, with a specific partition made on the data. Which of the following statements are correct? Multiple answers may be correct. Choose all options that seem right.



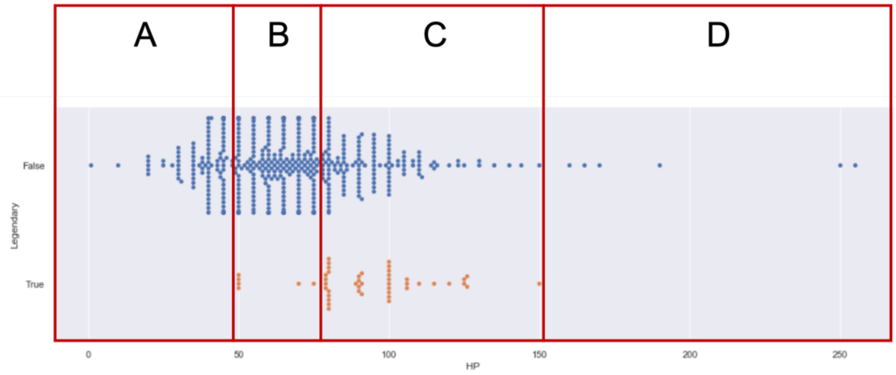
Answer Choice	Verdict	Explanation
We can be "completely confident" about Partition A -- it is "Legendary" = False.	Correct	True. There are only datapoints for "Legendary" = False in Partition A. Hence we can be "completely confident".
We can be "completely confident" about Partition D -- it is "Legendary" = False.	Correct	True. There are only datapoints for "Legendary" = False in Partition D. Hence we can be "completely confident".
We can be "completely confident" about Partition B -- it is "Legendary" = False.	Wrong	Wrong. There are datapoints for both "Legendary" = True and False in Partition B. Hence we can't be entirely confident.
We can be "completely confident" about Partition C -- it is "Legendary" = False.	Wrong	Wrong. There are datapoints for both "Legendary" = True and False in Partition C. Hence we can't be entirely confident.

Reference Module 4 Topic 1 : Binary Classification Slide 5 and Slide 6

General Comment : You are "completely confident" on the parts where all datapoints are of the same type. Otherwise, there is always a probability that the datapoints could be of one class or the other. In fact, finding "pure" partitions, with single-type datapoints, is quite rare in practice.

Question 3

Arrange the following partitions in decreasing order of their "confidence" for the variable "Legendary" to be False. That is, the partition with the maximum confidence for "Legendary" = False should be first, and decrease to the minimum confidence partition.



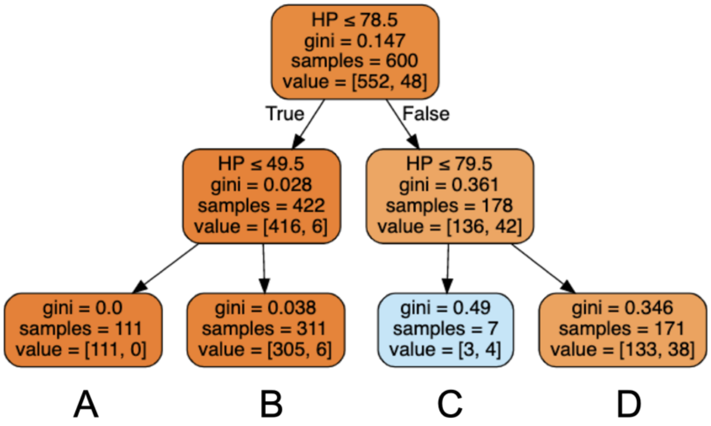
Partition A > Partition D > Partition B > Partition C

Reference Module 4 Topic 1 : Binary Classification Slide 6 and Slide 7

General Comment : The most confident partitions will have maximum datapoints of the same class, while the least confident ones will have a random mix. Partition A is higher in confidence, as it has more datapoints than Partition D.

Question 4

Arrange the leaf nodes of the following decision tree in decreasing order of their "confidence". That is, the most confident node for predicting "Legendary" should be first, and then decrease to the least confident node for predicting "Legendary".



1. Node A, with Gini = 0 and Datapoints = [111, 0]
2. Node B, with Gini = 0.038 and Datapoints = [305, 6]
3. Node D, with Gini = 0.346 and Datapoints = [133, 38]
4. Node C, with Gini = 0.49 and Datapoints = [3, 4]

Reference Module 4 Topic 1 : Binary Classification Slide 7

General Comment : The lower the Gini Index, the more confident the node is in predicting "Legendary". Gini = 0 will mean all data points are of the same "Legendary" label. Thus, Gini Index tells you about the confidence in a node.

Question 5

Suppose that a specific node is a Decision Tree has the following distribution for "Legendary" : True = 300, False = 300. That is, there are 600 datapoints in that node (or partition), out of which 300 are "Legendary" = True, and 300 are "Legendary" = False. What is the Gini Index for this specific node of the tree?

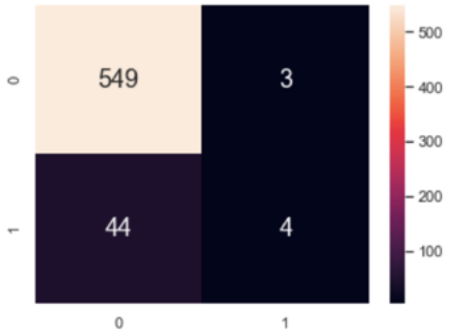
Answer	Explanation
0.5	Correct. Gini is 0.5 for a Binary Uniform distribution, with 50:50 ratio.

Reference Module 4 Topic 1 : Binary Classification Slide 7

General Comment : Note that for a perfectly uniform distribution, as in this case (300:300), the Gini Index is always 0.5. This in fact, tells you that the node has "least possible" confidence about prediction.

Question 6

Arrange the following quantities in decreasing order, as per the Confusion Matrix shown in the picture. That is, the highest quantity should come first, and then decrease to the lowest quantity coming at the end.



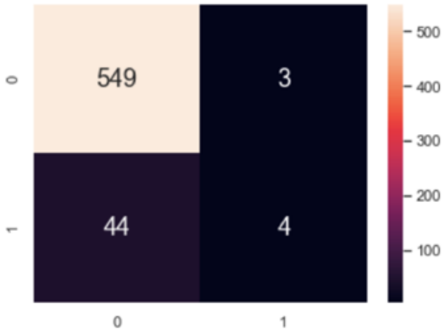
1. True Negatives -- "Legendary" = False (0) predicted as "Legendary" = False (0) 549
2. False Negatives -- "Legendary" = True (1) predicted as "Legendary" = False (0) 44
3. True Positives -- "Legendary" = True (1) predicted as "Legendary" = True (1) 4
4. False Positives -- "Legendary" = False (0) predicted as "Legendary" = True (1) 3

Reference Module 4 Topic 1 : Binary Classification Slide 9 and Slide 10

General Comment : Check definitions of True Positives, True Negatives, False Positives and False Negatives.

Question 7

Based on the confusion matrix as follows, calculate the False Positive Rate (FPR) in decimals. Submit your answer as a decimal number rounded off to four decimal places.

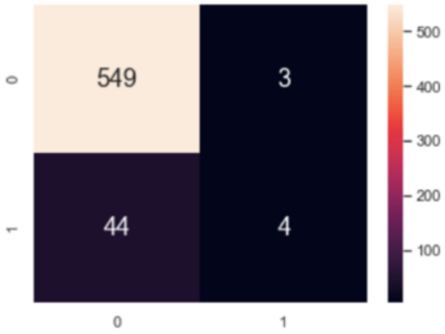


Answer	Explanation
0.0054	Correct. FPR = FP / (TN + FP) = 3 / (549 + 3) = 3 / 552 = 0.0054

Reference Module 4 Topic 1 : Binary Classification Slide 10

Question 8

Based on the confusion matrix as follows, calculate the False Negative Rate (FNR) in decimals. Submit your answer as a decimal number rounded off to four decimal places.



Answer	Explanation
0.9167	Correct. FNR = FN / (TP + FN) = 44 / (4 + 44) = 44 / 48 = 0.9167

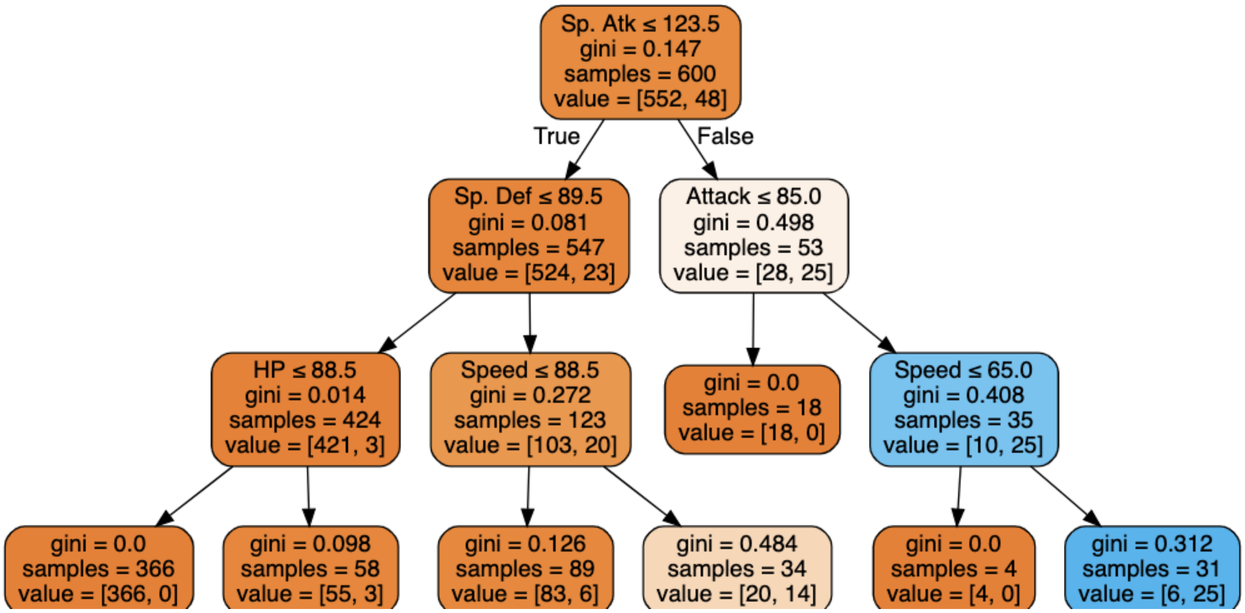
Reference Module 4 Topic 1 : Binary Classification Slide 10

General Comment : Just to recap quickly, here are the formulas : FPR = FP / (TN + FP) and FNR = FN / (TP + FN).

Question 9

Based on the Tree, arrange the variables in decreasing order of their "importance" in predicting the binary variable Legendary.

1]:



1. Sp. Atk – Special Attack of a Pokemon as it occurs in the very first split of the tree
2. Sp. Def – Special Defence of a Pokemon as it occurs right after, in the next split
3. Speed – Speed of a Pokemon as it occurs in more nodes, and improves Gini better
4. HP – Hit Points for a Pokemon as it occurs only in one node, and improves Gini slightly

Reference Module 4 Topic 1 : Binary Classification Slide 12 and Slide 13

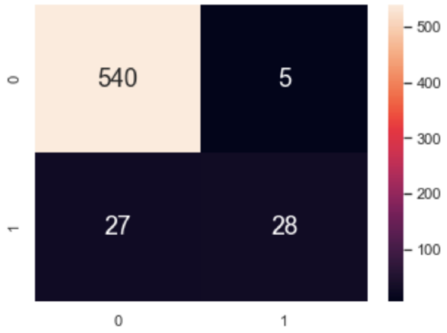
General Comment : The more the drop in Gini from a parent node to the children node, the better is the split. Now, if a variable helps in dropping the Gini from parent to children more than another variable, it is deemed more important.

The tree automatically decides the best variables by splitting the tree using them. Thus, the higher the variable occurs in a tree, the more important it is. We will use the Gini drop idea for variables that occur at the same level of a tree.

According to the logic, HP tackled an almost pure node [421, 3] and dropped to children [366, 0] and [55, 3]. This is not as significant as Speed, which tackled relatively more mixed nodes [103, 20] and [10, 25], to drop Gini much further.

Question 10

What is the classification accuracy of the Decision Tree that produces the following confusion matrix? Enter your answer as a decimal number, correct (rounded off) to four decimal places.



Answer	Explanation
0.9467	Correct. Classification Accuracy = (TP + TN) / Total = (540 + 28) / (540 + 5 + 27 + 28)

Reference Module 4 Topic 1 : Binary Classification Slide 10

General Comment : Classification Accuracy is simply the fraction of correct predictions, that is, (TP + TN) / Total.