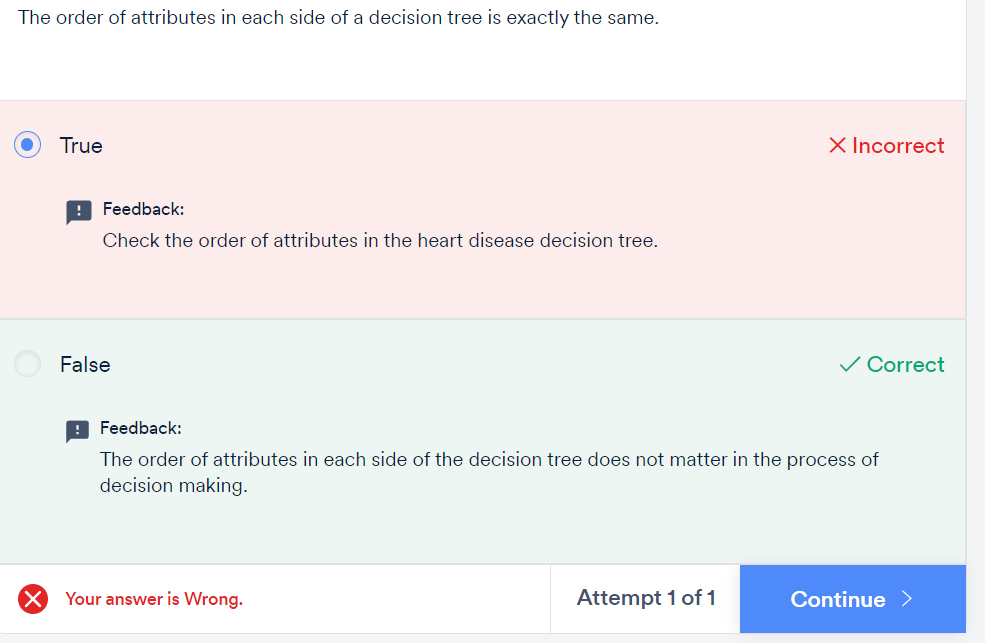
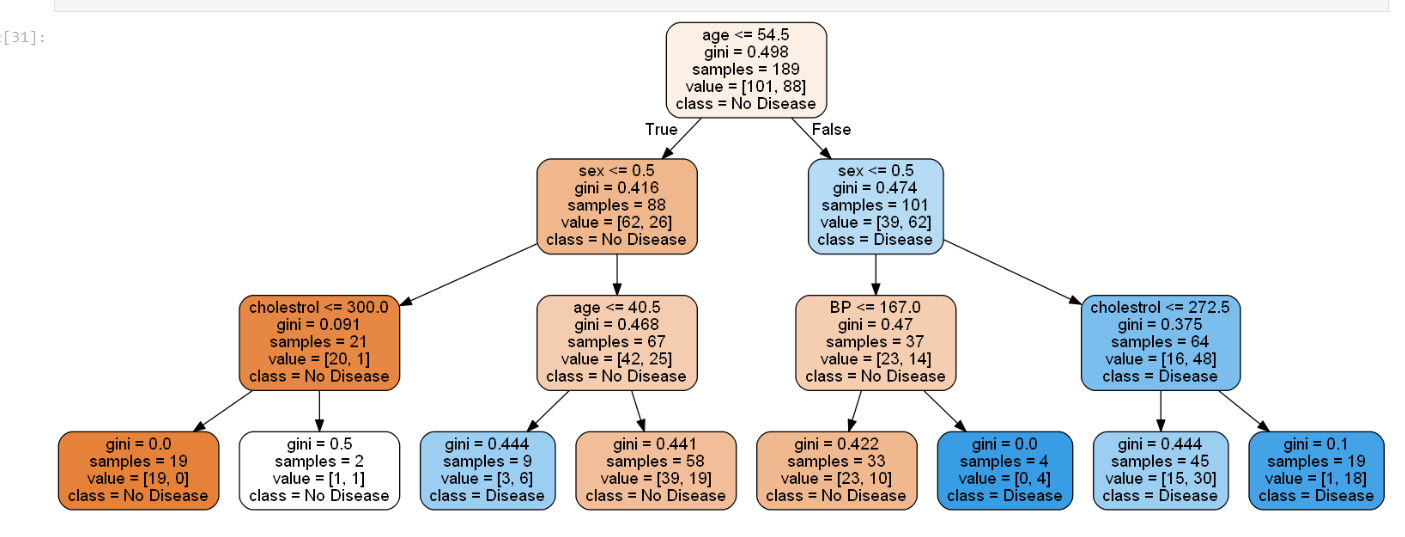
So, as you saw in the video, constructing a decision tree involves the following steps:

1. Recursive binary splitting/partitioning the data into smaller subsets
2. Selecting the best rule from a variable/ attribute for the split
3. Applying the split based on the rules obtained from the attributes
4. Repeating the process for the subsets obtained
5. Continuing the process until the stopping criterion is reached
6. Assigning the majority class/average value as the prediction

Now, the decision tree building process is a **top-down** approach. The top-down approach refers to the process of starting from the top with the whole data and gradually splitting the data into smaller subsets.

The reason we call the process **greedy** is because it does not take into account what will happen in the next two or three steps. The entire structure of the tree changes with small variations in the input data. This, in turn, changes the way you split and the final decisions altogether. This means that the process is not holistic in nature, as it only aims to gain an immediate result that is derived after splitting the data at a particular node based on a certain rule of the attribute.

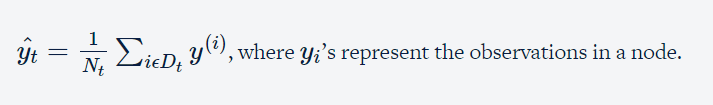


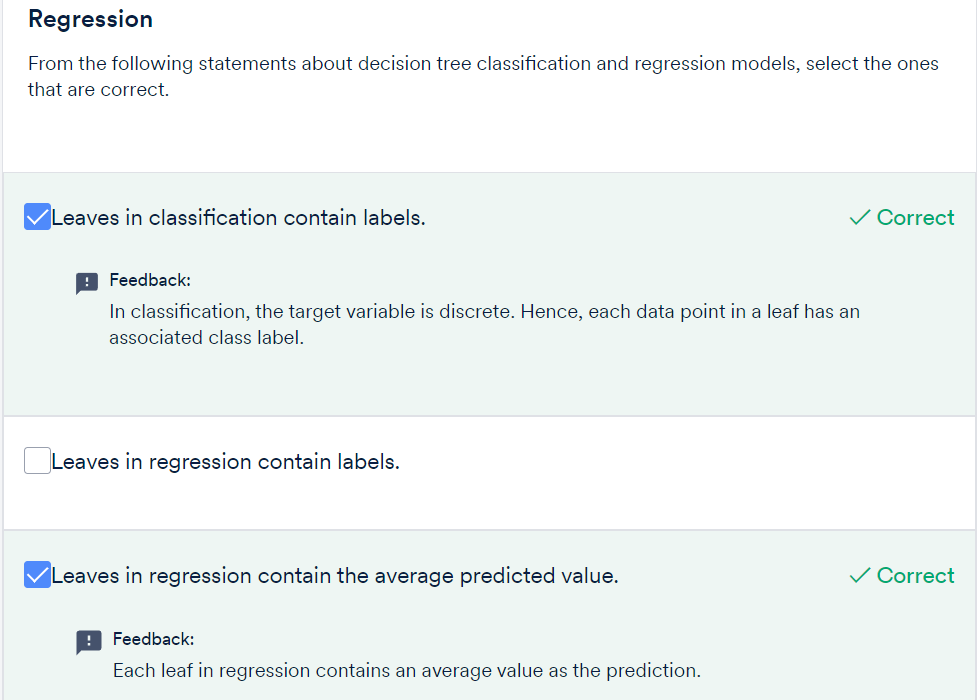


Let’s summarize the advantages of tree models one by one in the following order:

* Predictions made by a decision tree are easily **interpretable**.
* A decision tree is **versatile** in nature. It does not assume anything specific about the nature of the attributes in a data set. It can seamlessly handle all kinds of data such as numeric, categorical, strings, Boolean, etc.
* A decision tree is **scale-invariant**. It does not require normalisation, as it only has to compare the values within an attribute, and it handles multicollinearity better.
* Decision trees often give us an idea of the relative **importance** of the explanatory attributes that are used for prediction.
* They are highly **efficient** and **fast** algorithms.
* They can **identify complex relationships** and work well in certain cases where you cannot fit a single linear relationship between the target and feature variables. This is where regression with decision trees comes into the picture.

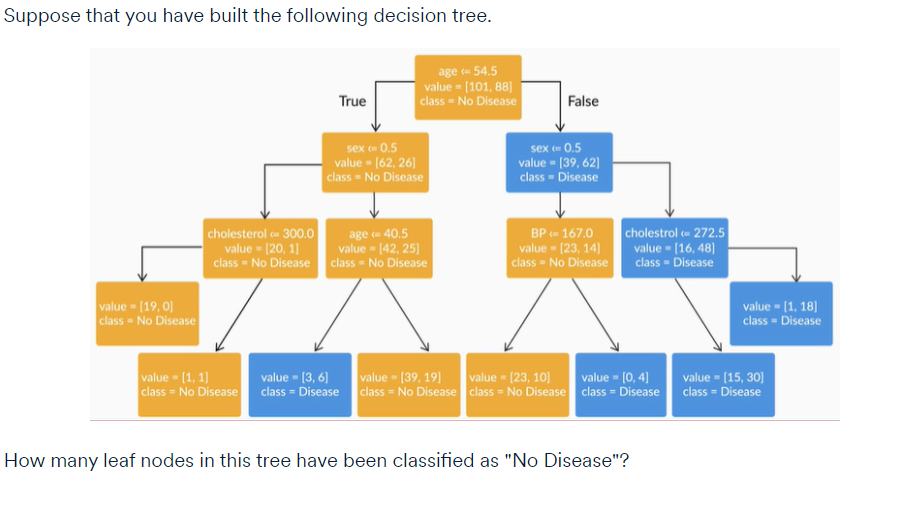
In regression problems, a decision tree splits the data into multiple subsets. The difference between decision tree classification and decision tree regression is that in **regression**, each leaf represents the **average of all the values as the prediction** as opposed to a **class label** in **classification** trees. For classification problems, the prediction is assigned to a leaf node using majority voting but for regression, it is done by taking the average value. This average is calculated using the following formula:





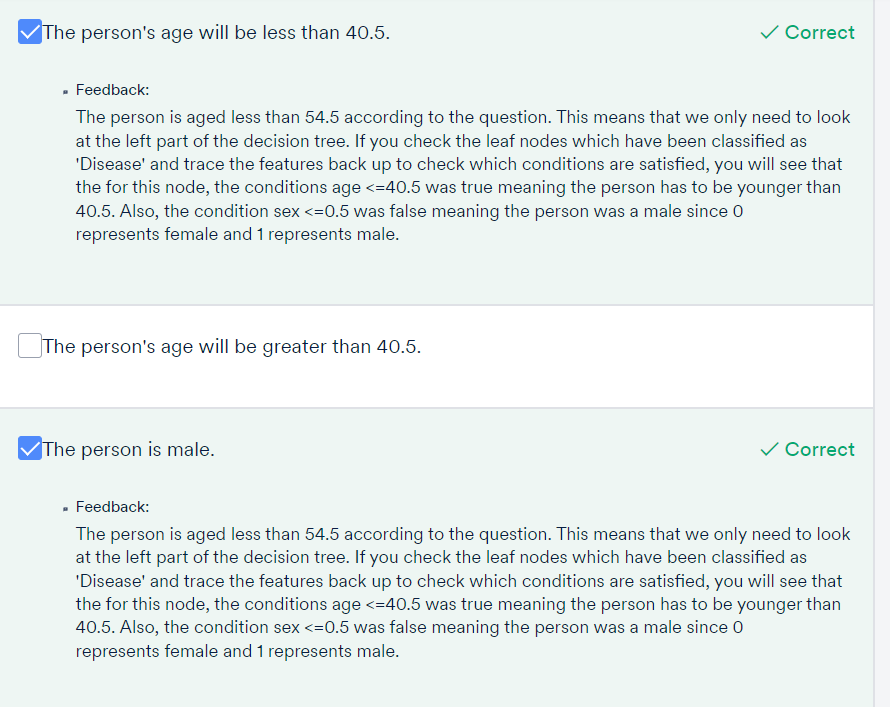


GRADED QUESTIONS:



4





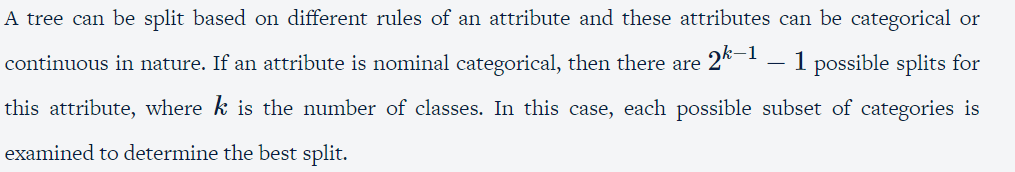
**Algorithms for Decision Tree Constructions**

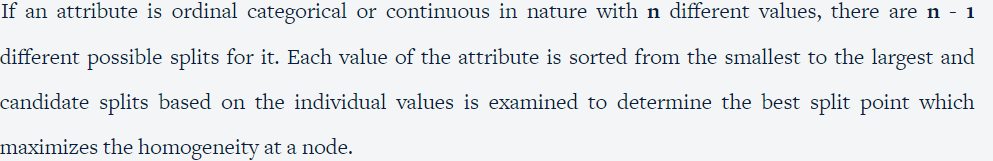
In this session:

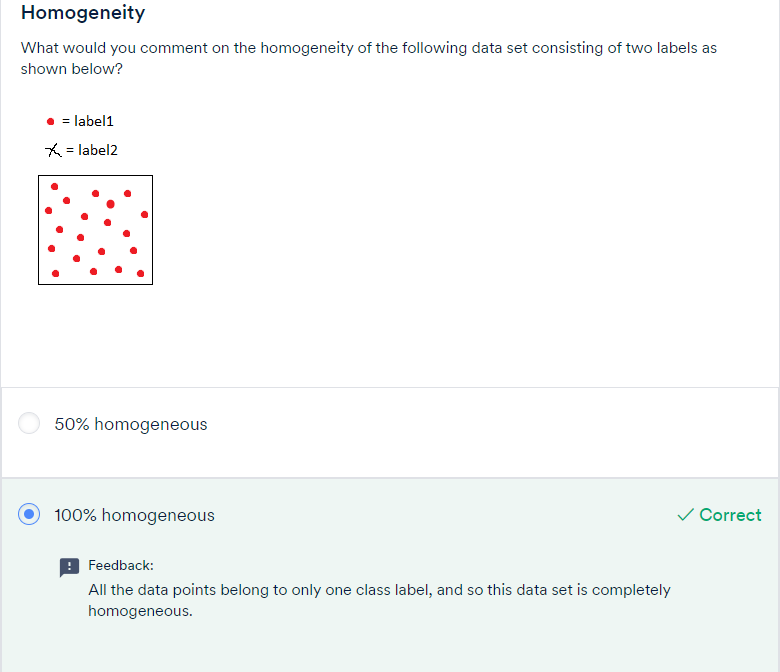
* Splitting and homogeneity
* Impurity measures
* Best split
* Regression trees

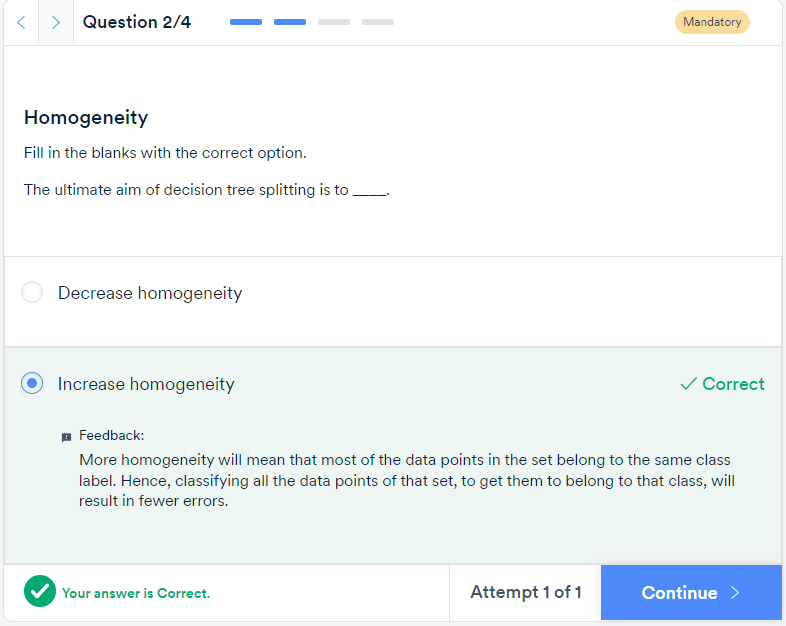
After the split the subsets **are Homogenous or pure**

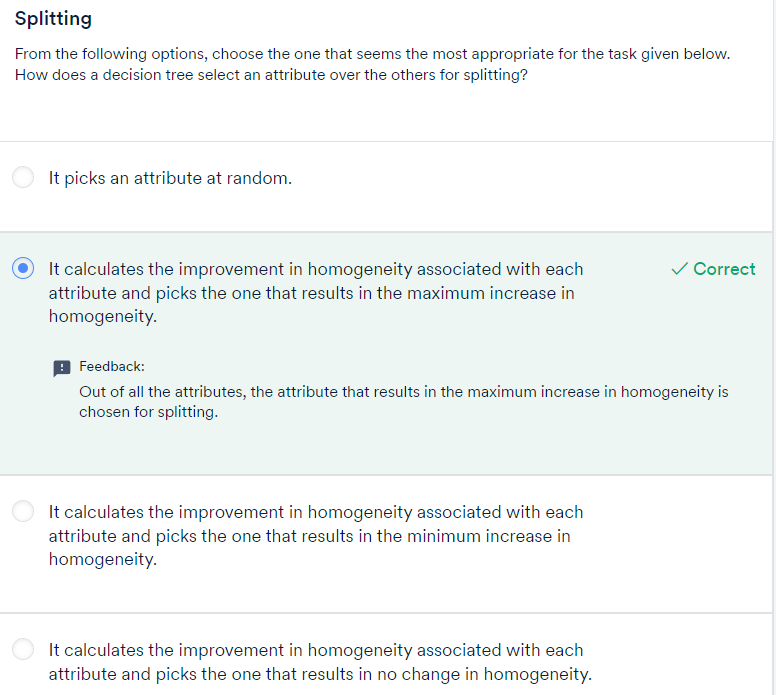
You will learn how to use specific methods to measure homogeneity, namely the Gini index, entropy, classification error (for classification), and MSE (for regression).

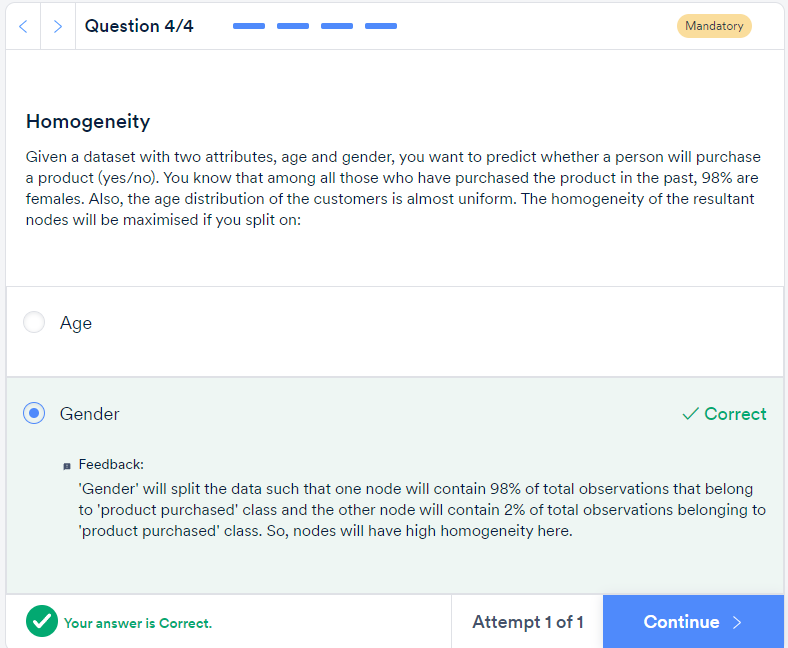


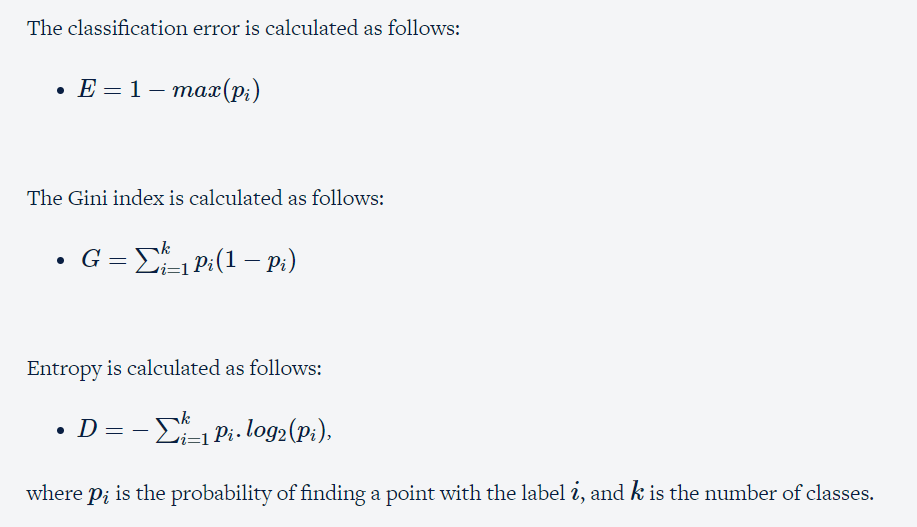










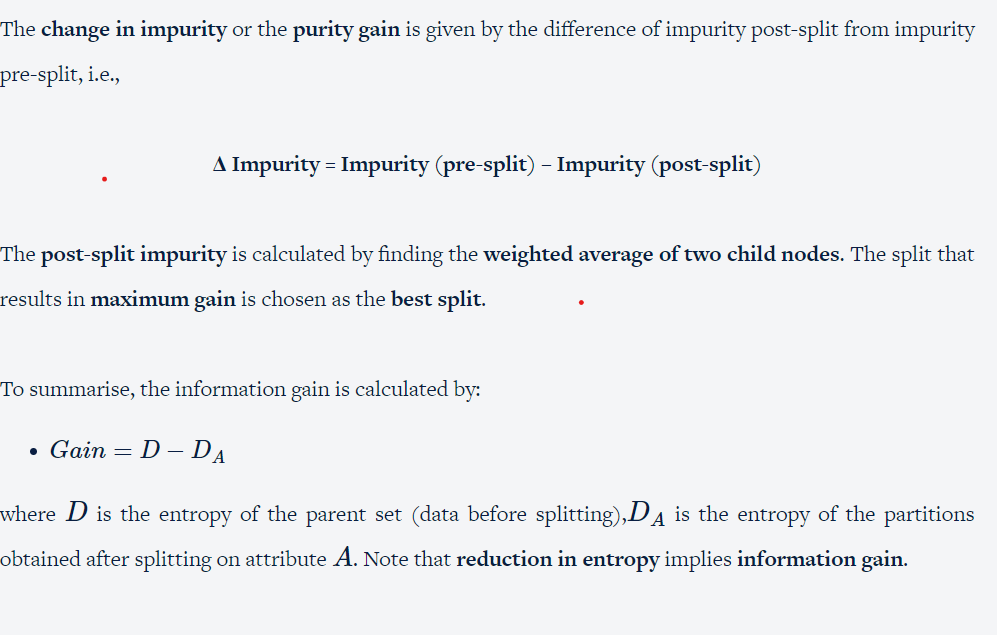


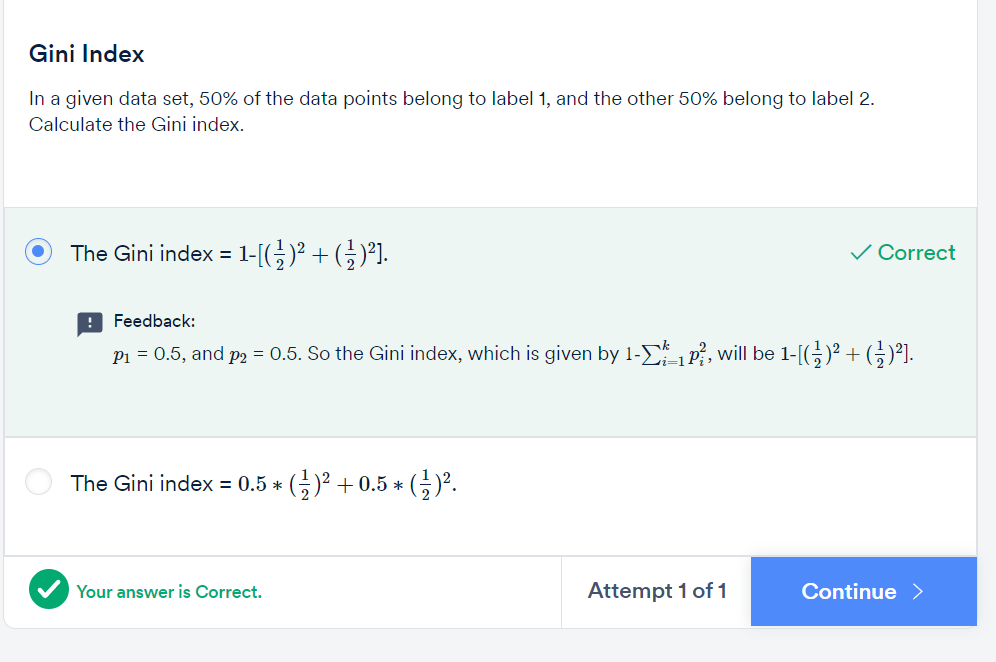


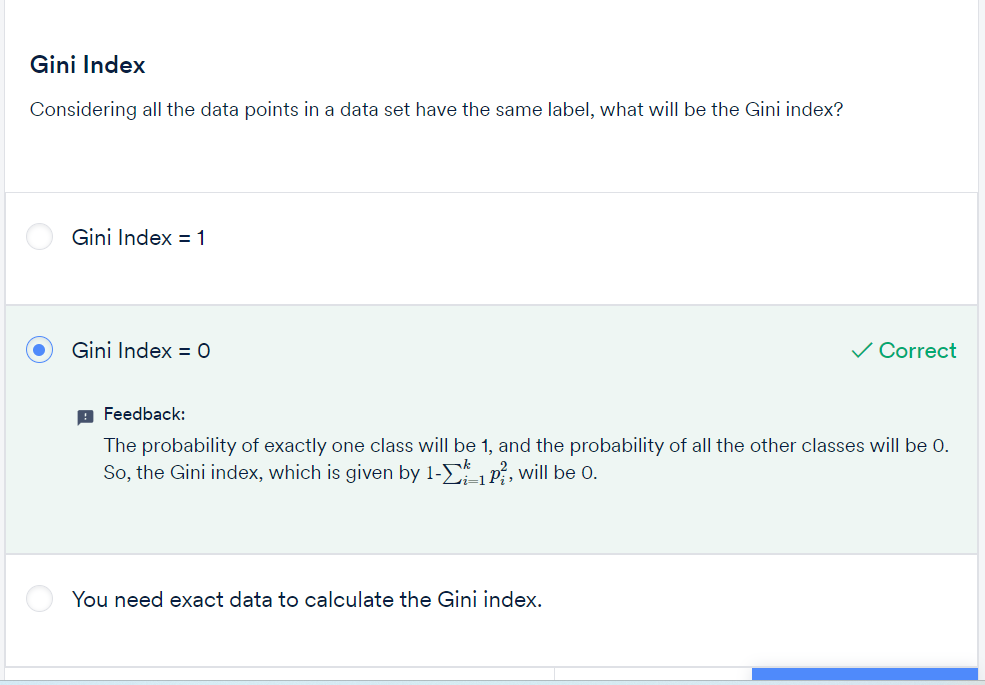
Suppose you have a data set with two class labels. If the data set is completely homogeneous, i.e., all the data points belong to label 1, then the probability of finding a data point corresponding to label 2 will be 0 and that of label 1 will be 1. So, p1 = 1 and p2 = 0. The Gini index, which is equal to 0, will be the lowest in such a case. Hence, **the higher the homogeneity, the lower the Gini index.**

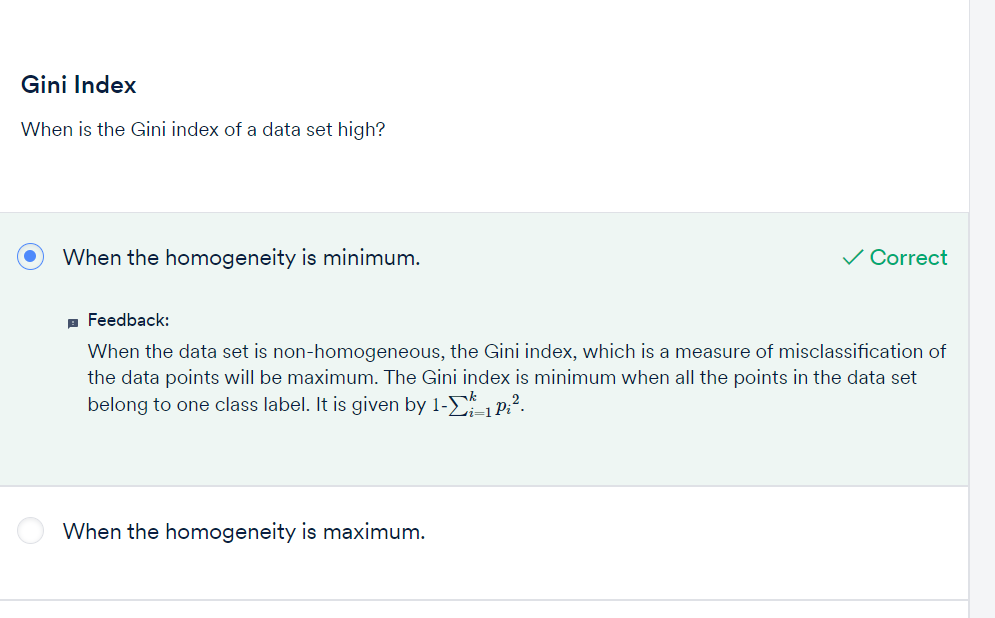
**Entropy**

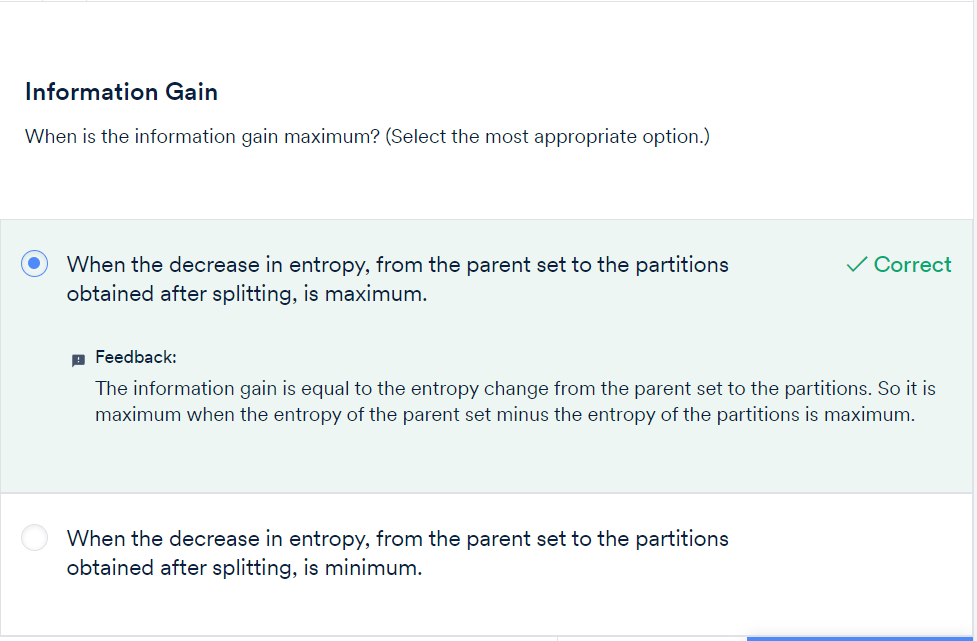
Entropy quantifies the degree of disorder in the given data, its value varies from 0 to 1. Entropy and the Gini index are similar numerically. If a data set is completely homogenous, then the entropy of such a data set will be 0, i.e., there is no disorder in the data. If a data set contains an equal distribution of both the classes, then the entropy of that data set will be 1, i.e., there is complete disorder in the data. Hence, like the Gini index,**the higher the homogeneity, the lower the entropy.**

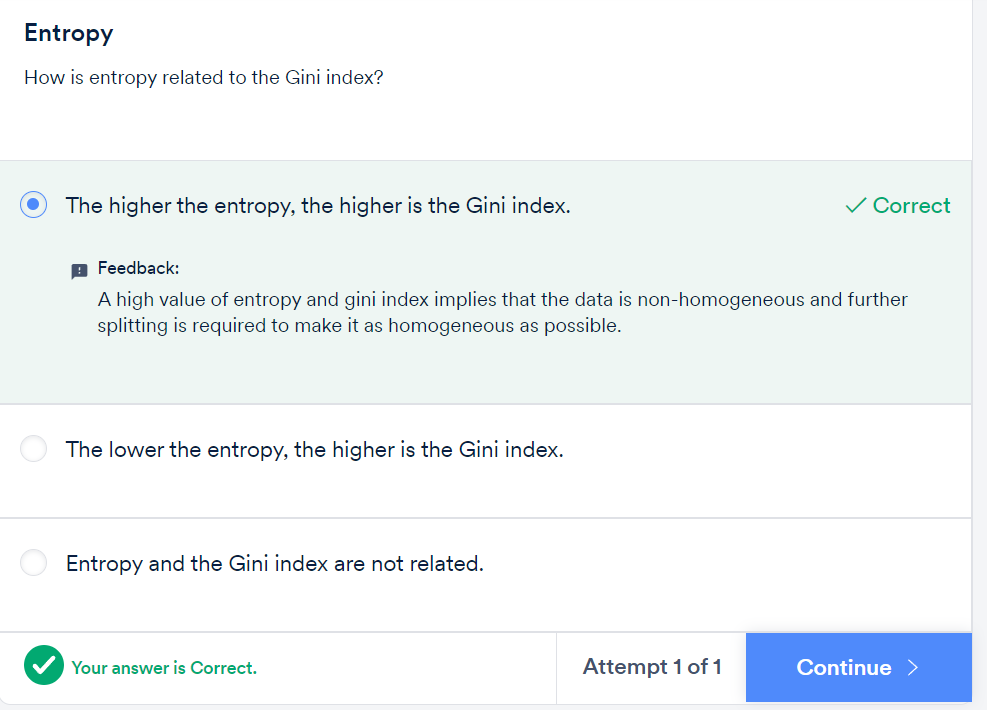


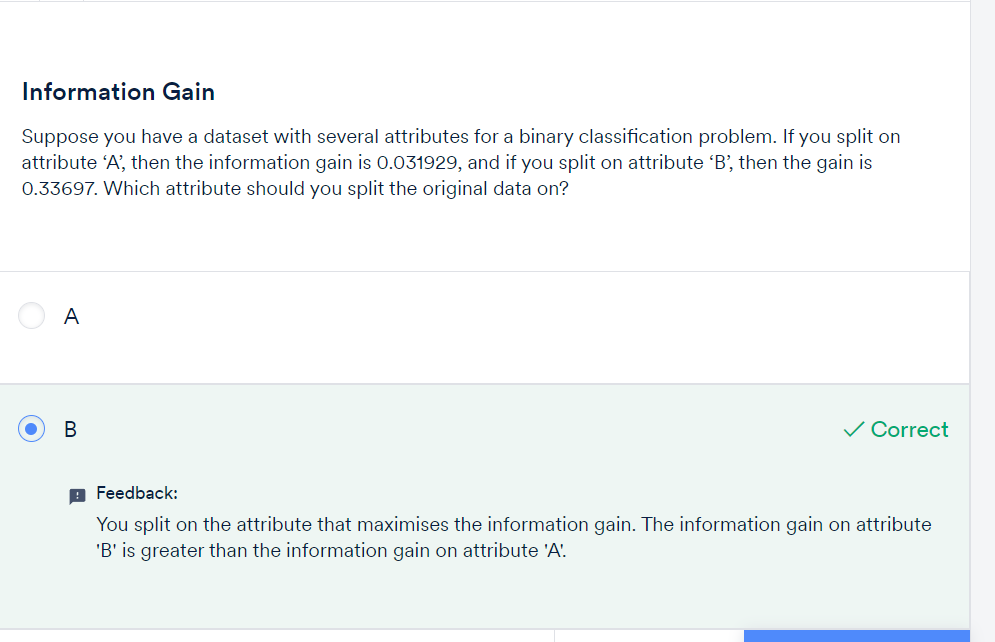




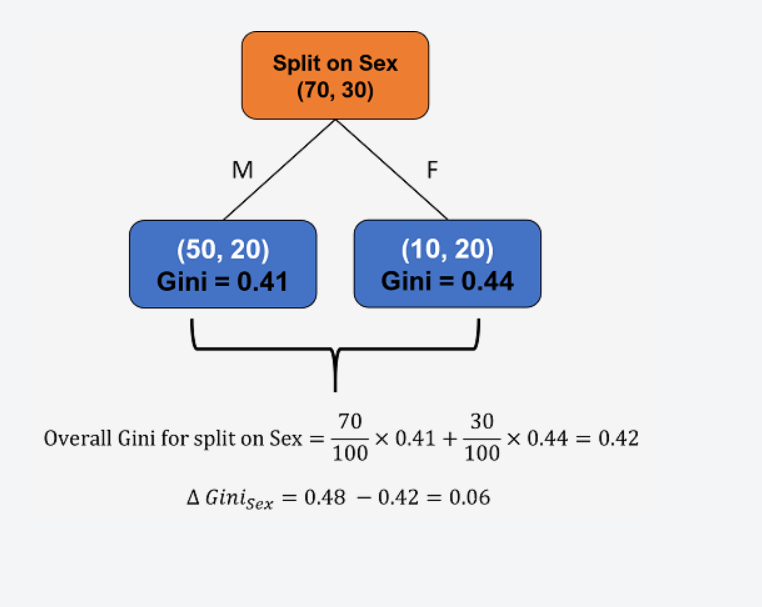


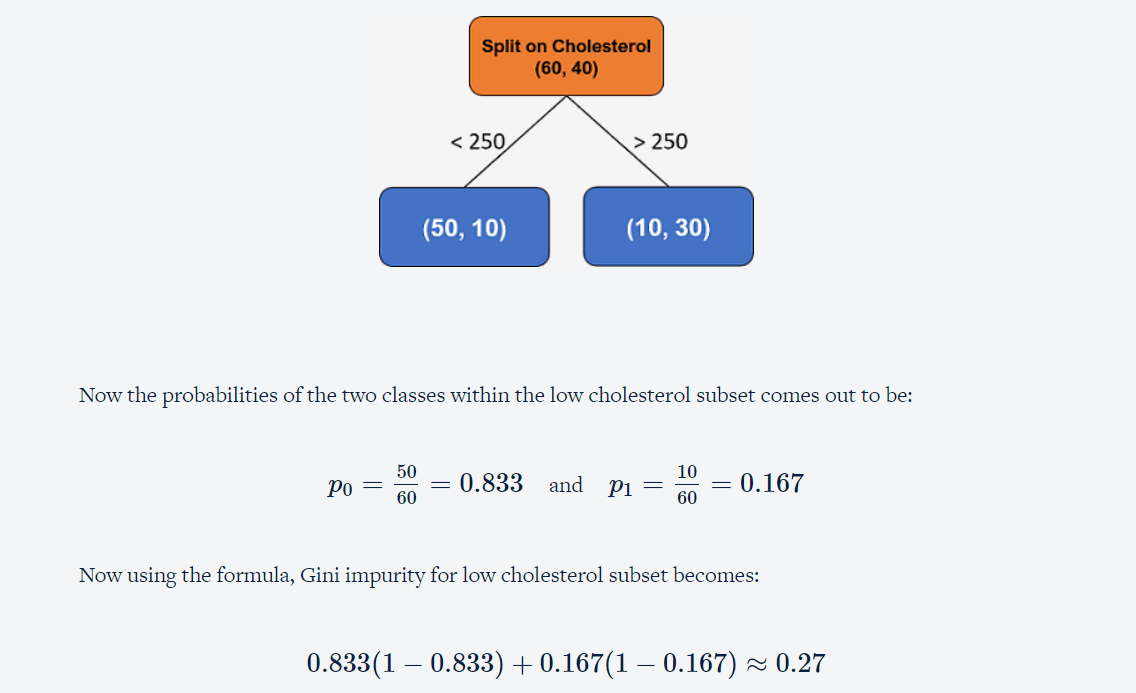


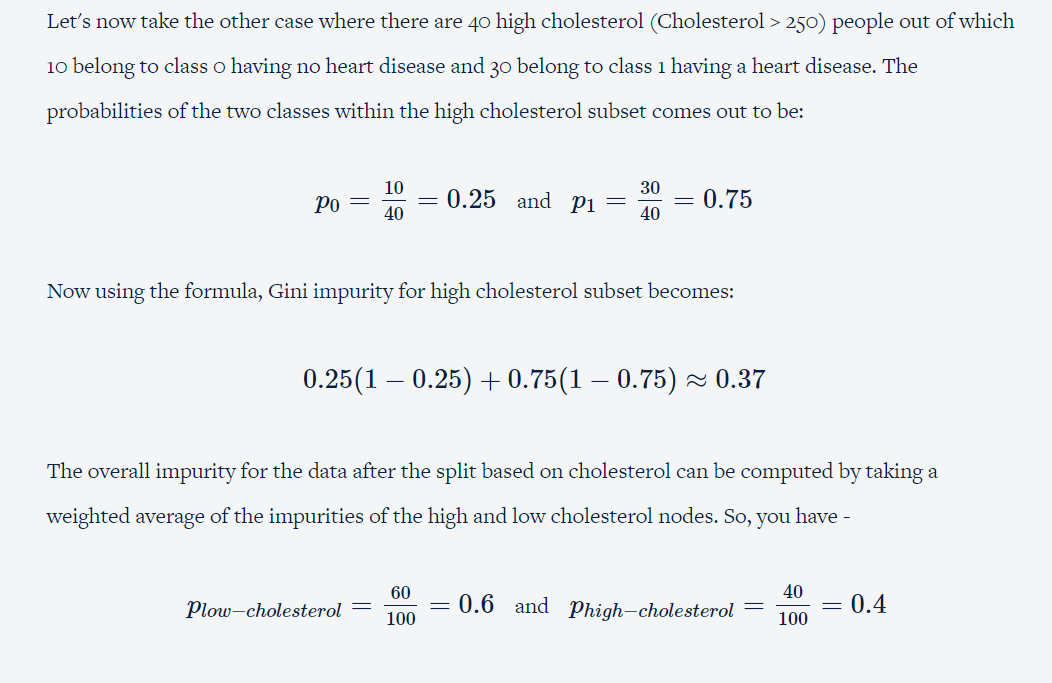


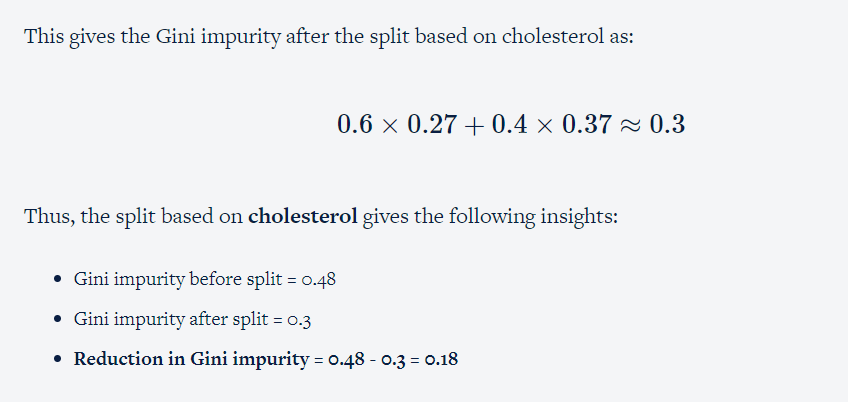


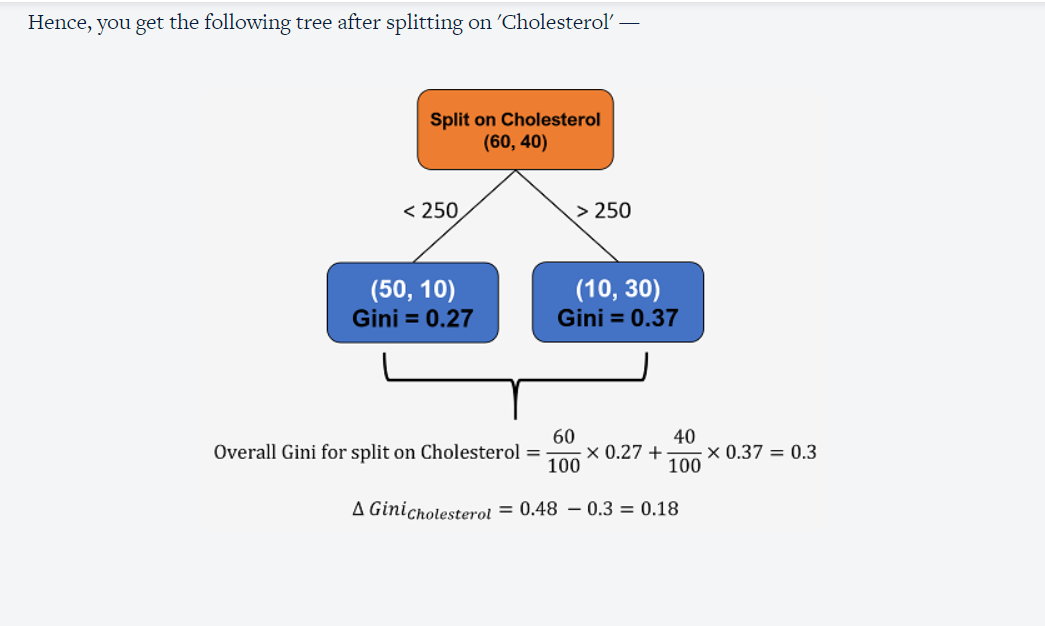
GINI INDEX:





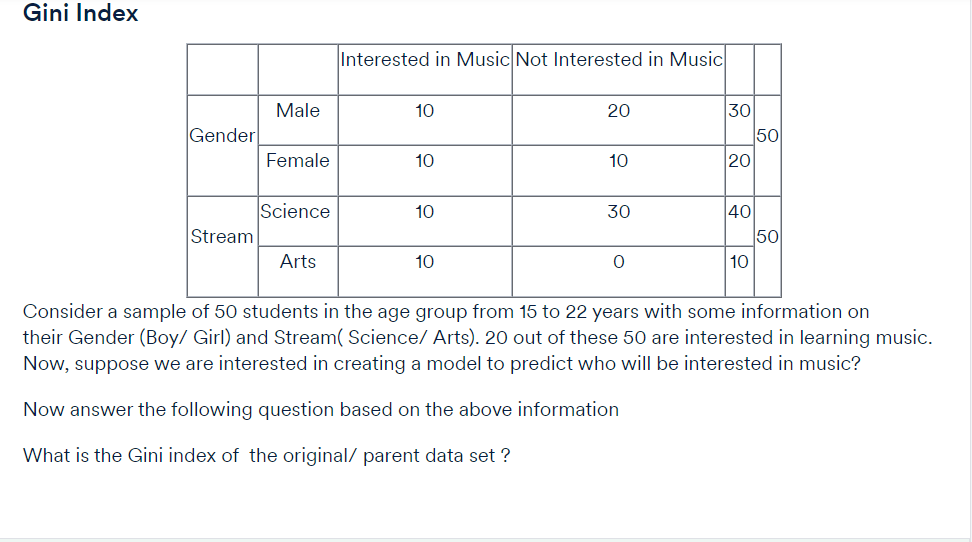


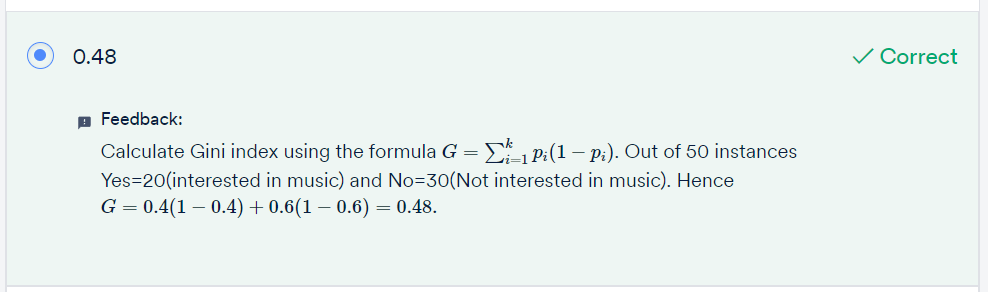


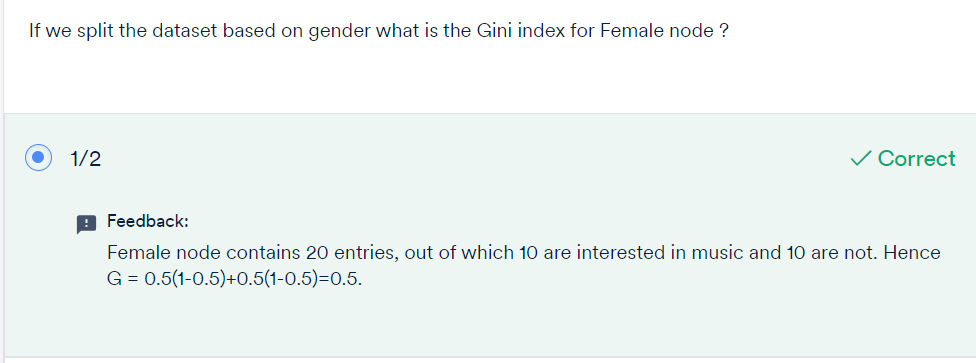


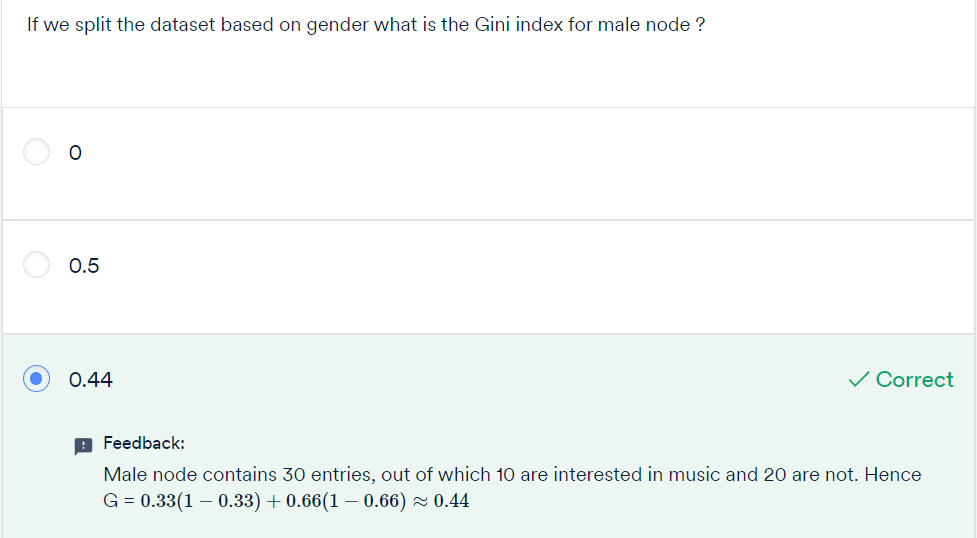
Let's summarise all the steps you performed.

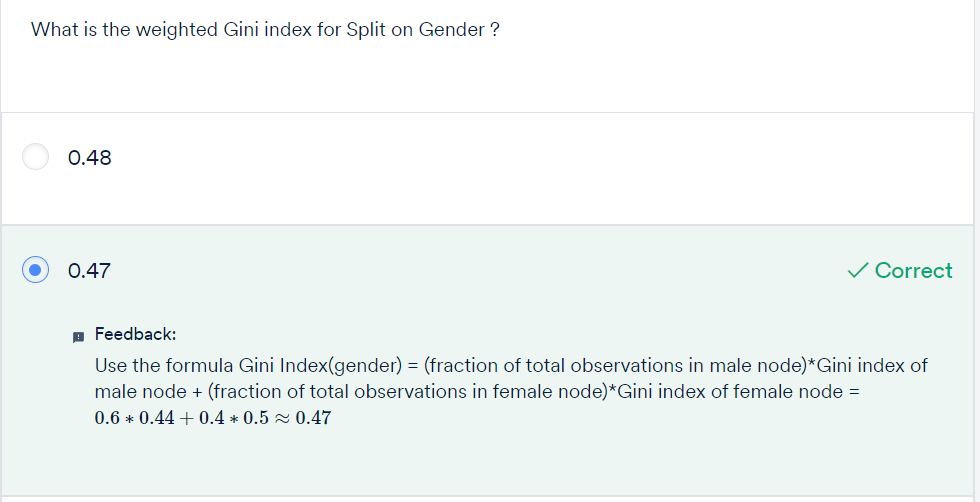
1. Calculate the Gini impurity before any split on the whole dataset.
2. Consider any one of the available attributes.
3. Calculate the Gini impurity after splitting on this attribute for each of the levels of the attribute. In the example above, we considered the attribute 'Sex' and then calculated the Gini impurity for both males and females separately.
4. Combine the Gini impurities of all the levels to get the Gini impurity of the overall attribute.
5. Repeat steps 2-5 with another attribute till you have exhausted all of them.
6. Compare the decrease in Gini impurity across all attributes and select the one which offers maximum reduction.

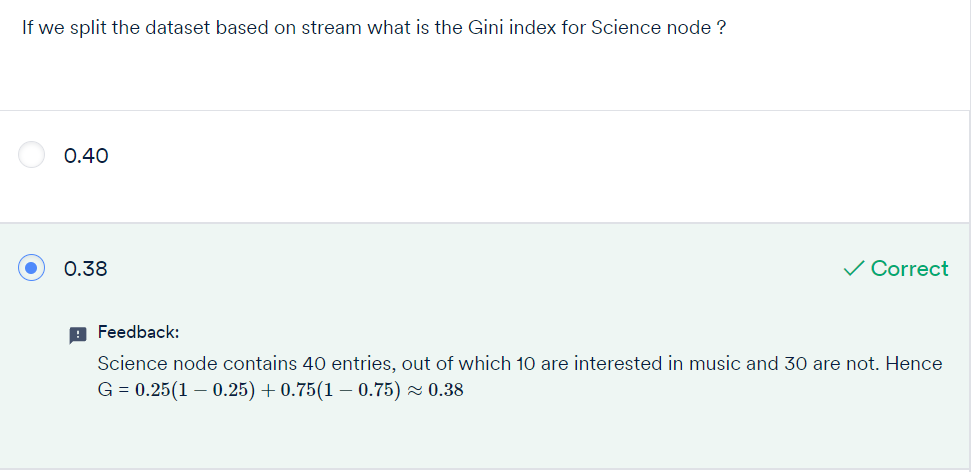


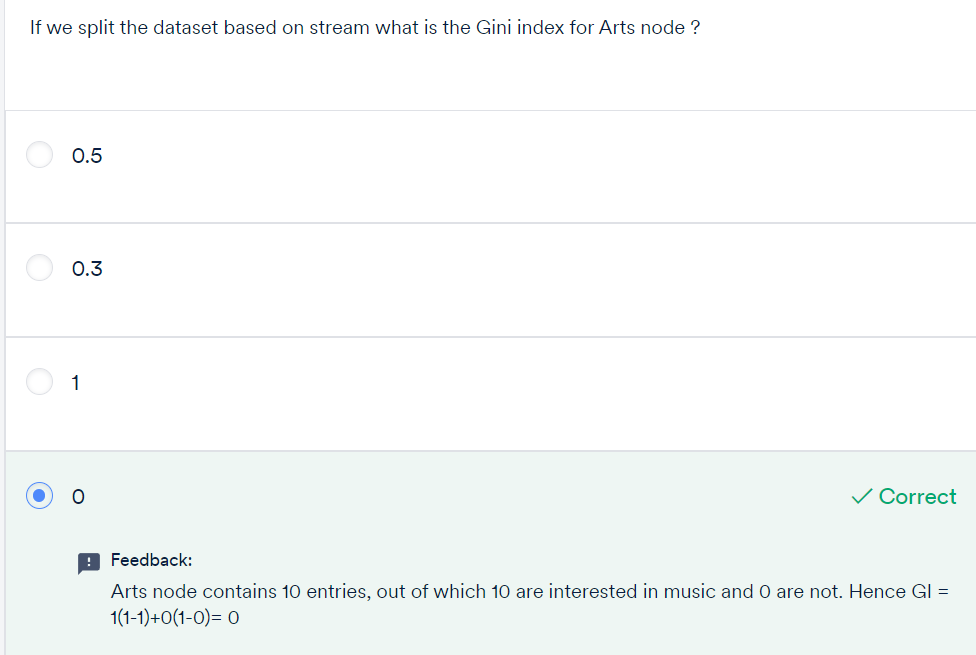


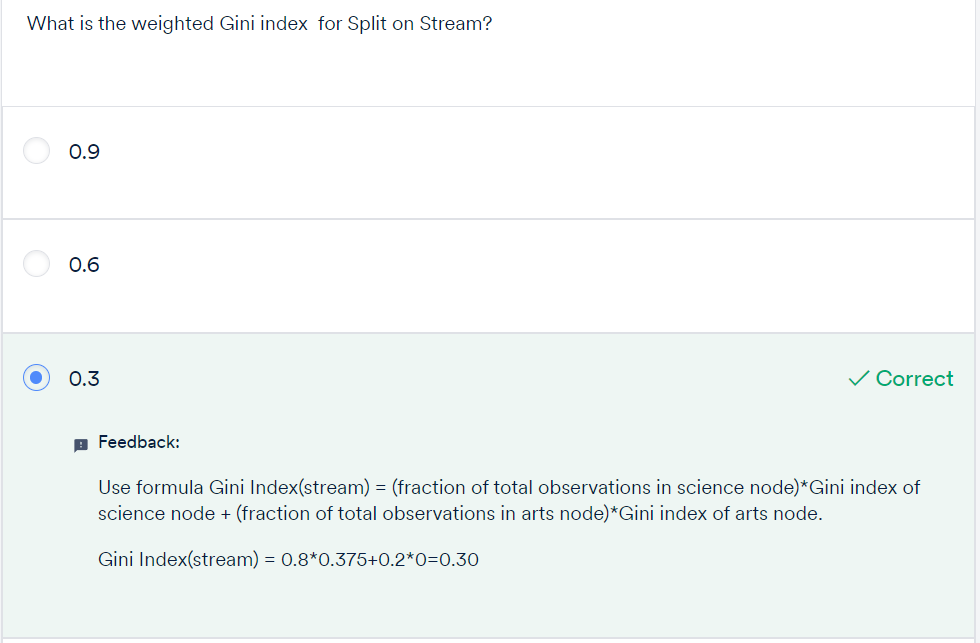


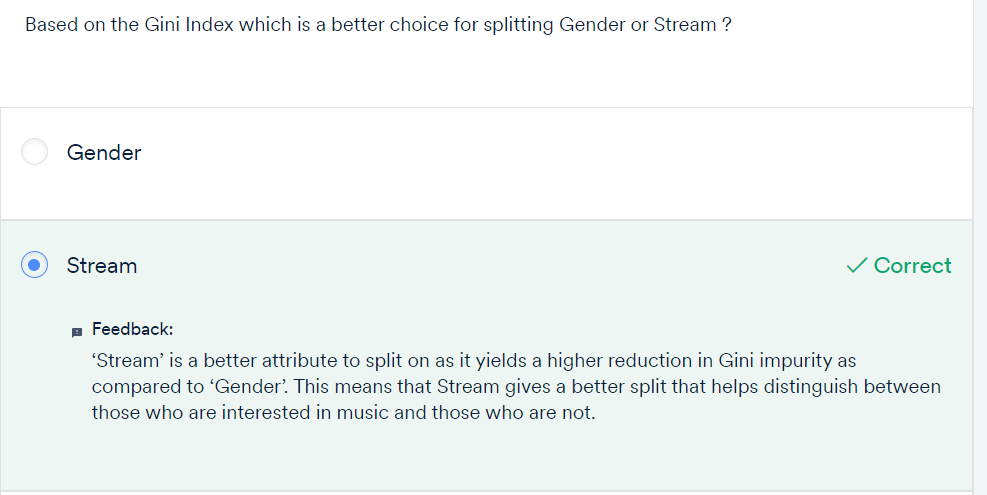










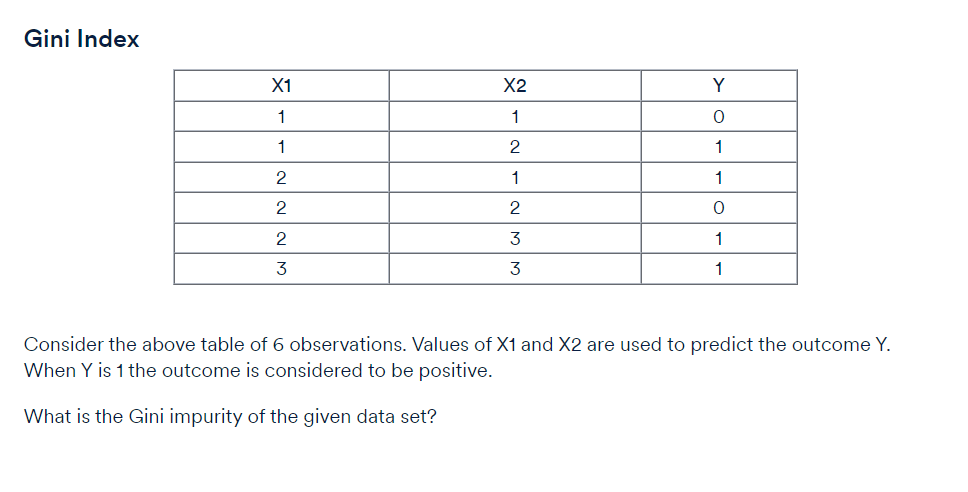


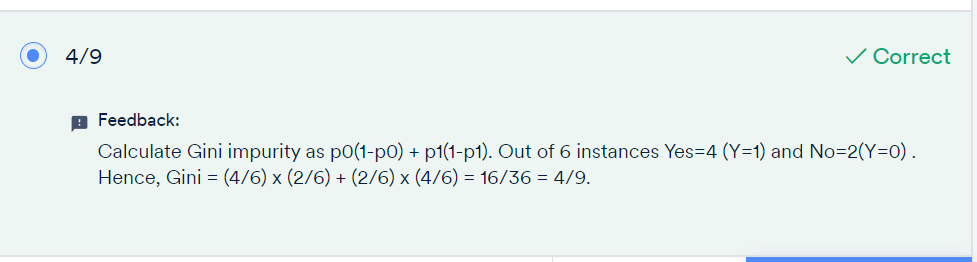
Decision trees help in quantifying the importance of each feature by calculating the reduction in the impurity for each feature at a node. The feature that results in a **significant reduction in the impurity is the important variable**, and the one that results in **less impurity reduction is the less important variable**.

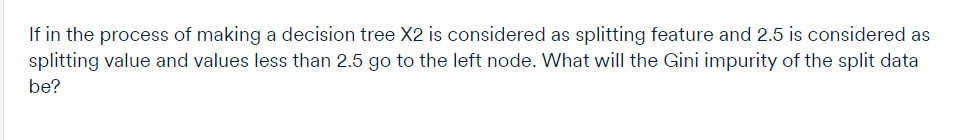
Now, let’s summarise your learnings so far:

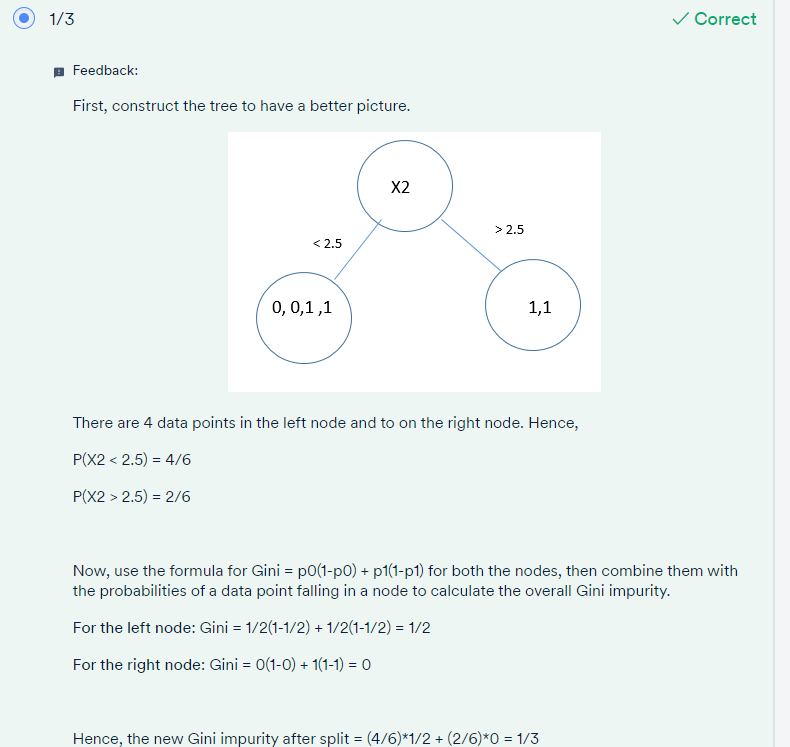
* A decision tree first decides on an attribute to split on.
* To select this attribute, it measures the homogeneity of the nodes before and after the split.
* You can measure homogeneity in various ways with metrics like Gini index and entropy.
* The attribute that results in the increase of homogeneity the most is then selected for splitting.
* Then, this whole cycle is repeated until you obtain a sufficiently homogeneous data set.

# GRADED QUESTIONS:

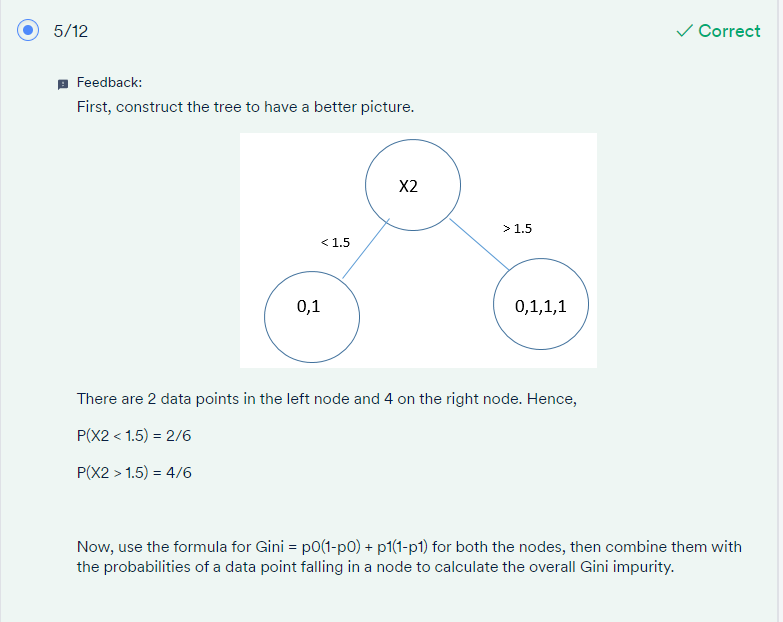


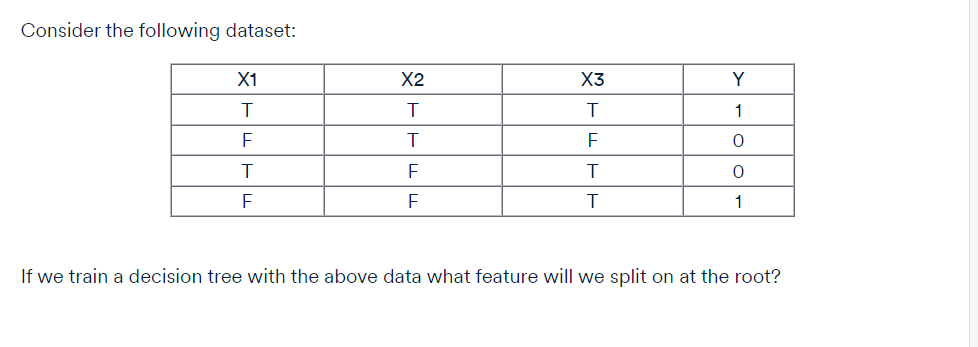


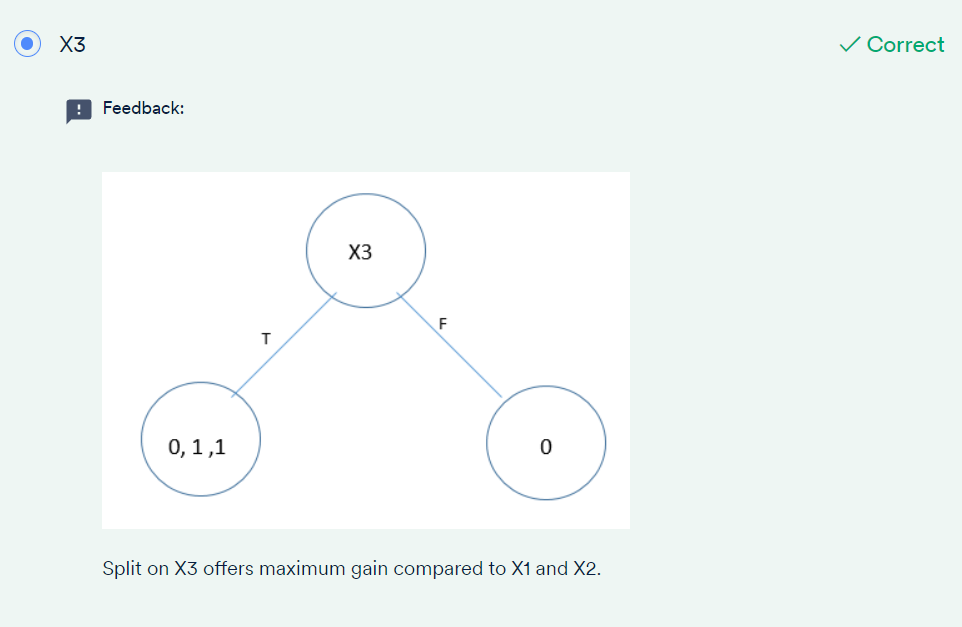




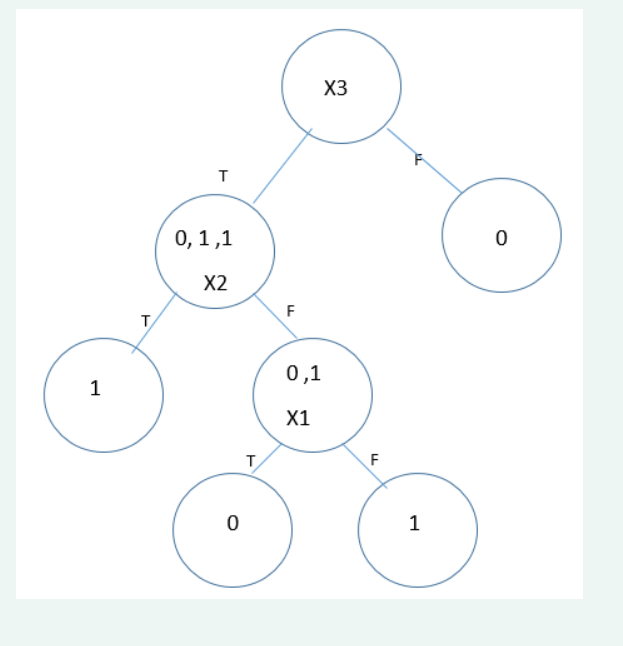
Now, instead of 2.5, if the splitting value of X2 is chosen as 1.5 then what is the Gini impurity of the split data?







Q. If we fully train the decision tree using the root feature (as the answer of the previous question) until each leaf has data points of the same output label (i.e., all leaf nodes will be completely homogenous), what will be the depth of that tree?



After fully training the tree in the previous question what is the training error of the decision tree?

