**IDS 572 Data Mining**

**Assignment 2**

**Submitted by:**

Bhalerao, Amit (abhale2)

Joshi, Madhura (mjosh4)

Singh, Vivek (vsingh34**)**

**Problem 2:**

Considering the dataset for classifying whether a type of food is appealing or not.

*(a) What is the information gain associated with choosing Temperature as root? What about Taste and Size? Briefly write the steps of your calculations.*

Using the formula for Entropy, for Temperature,

Ehot = -(2/5)log2(2/5) – (3/5)log2(3/5) = 0.97

Ecold = -(3/5)log2(3/5) – (2/5)log2(2/5) = 0.97

Weighted Entropyafter(Temperature) = (5/10)\*0.97 + (5/10)\*0.97 = 0.97

Entropybefore(Temperature) = -(5/10)log2(5/10) – (5/10)log2(5/10) = 1

Therefore, Gain = Entropybefore(Temperature) - Entropyafter(Temperature) = 1 - 0.97 = **0.03**

Similarly,

Gain(Size) = 1 – 0.72 = **0.28**

Gain(Taste) = 1 – 0.4 = **0.6**

*(b) Draw a full decision tree for the dataset by choosing appropriate features for splitting. Justify each split.*

1. Using the values of the Information Gain calculated in step (a), ‘Taste’ is selected as the root node since it has the highest gain value.

Based on the values of the attribute Taste 🡪 Salty, Sweet, Sour, the first level of the Decision tree is constructed.

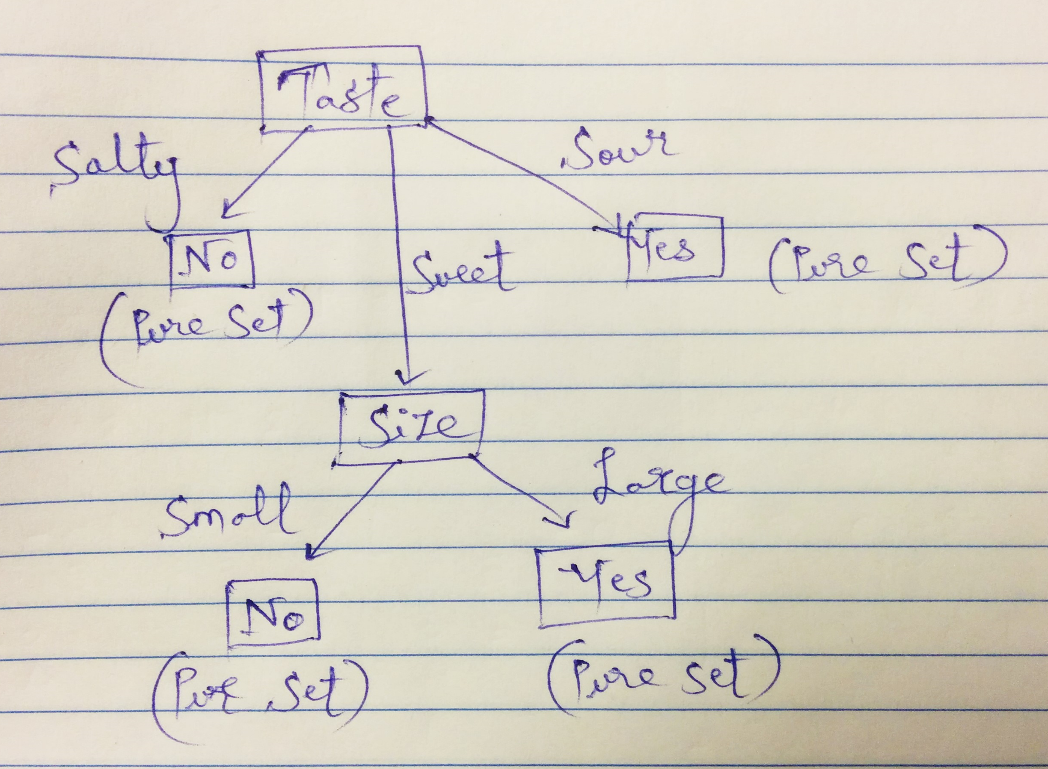
1. For ‘Salty’ and ‘Sour’ tastes, the leaf nodes are a pure set and the tree needs no more expansion.
2. For ‘Sweet’, the decision cannot be made without further splitting. Hence, the Gain values for Temperature and Size are calculated:

Gain(Temperature) = 0

Gain(Size) = 0.5

So, ‘Size’ is selected as the next attribute for splitting.

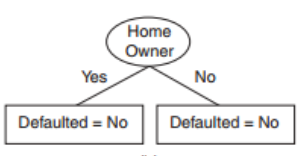
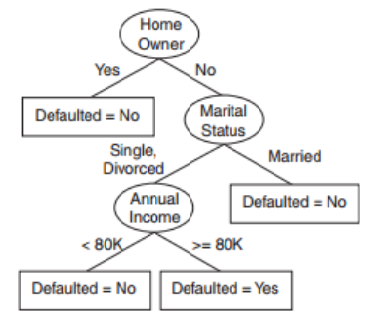
The decision tree can be drawn as below:



**Problem 3:**

Please briefly justify your answer to the following questions in one or two sentences.

*(a) Consider two decision trees that always yield the same class label, given the same test sample. Do both the trees need to have the same number of nodes? You can justify your answer with an example.*

Considering the example above, if two decision trees yield the same class label for the same test sample, the trees do not need to have the same number of nodes.

The number of attributes used in the tree to arrive at a particular decision could be different but the decision would be the same for a given data set.

*(b) Describe the purpose of separating the data into training and testing data.*

It is crucial that data is partitioned into ‘Training’ and ‘Testing’ sets in order to evaluate various Data Mining models. Most of the data (usually 65%) is used for the Training set based on which a model is built. The remaining (35%) data is used for Testing or measuring the performance on unseen data.

After a model has been processed using the Training set, it is tested by making predictions against the Test set. Because the data in the Testing set already contains known values for the attributes to be predicted, it is easy to determine whether the model is accurate.

*(c) Which problem do we try to address when using pruning? How do we do that? Please explain.*

The problem that pruning tries to address is ‘Overfitting’ which is a difficulty for decision tree models. Overfitting happens when the learning algorithm continues to develop hypotheses that reduce training set error at the cost of an increased test set error.

The decision trees become complex, sometimes having branches that fit data too specifically. This results in an unstable model.

Pruning is an approach that helps avoid overfitting while building decision trees.

* Pre-pruning stops growing the tree earlier, before it perfectly classifies the training set.
* Post-pruning allows the tree to perfectly classify the training set and then prunes the tree.
  + Subtree replacement selects a subtree and replaces it with a single leaf
  + Subtree raising selects a subtree and replaces it with the child one i.e. a ‘sub-subtree’ replaces its parent

Post pruning is a better approach since it is not easy to precisely estimate when to stop growing the tree.