Show your working in all calculations.

1. You find yourself stranded on a desert island, and you need to start gathering food to survive. You notice that there are two types of berries growing on this island, but you don't know if the fruit is edible. After walking for some time you come across an abandoned wooden hut, where find some measuring tools and notes from an old explorer - including a diagram indicating how to differentiate between poisonous and edible berries, based on fruit size and bush height. The diagram is shown below:

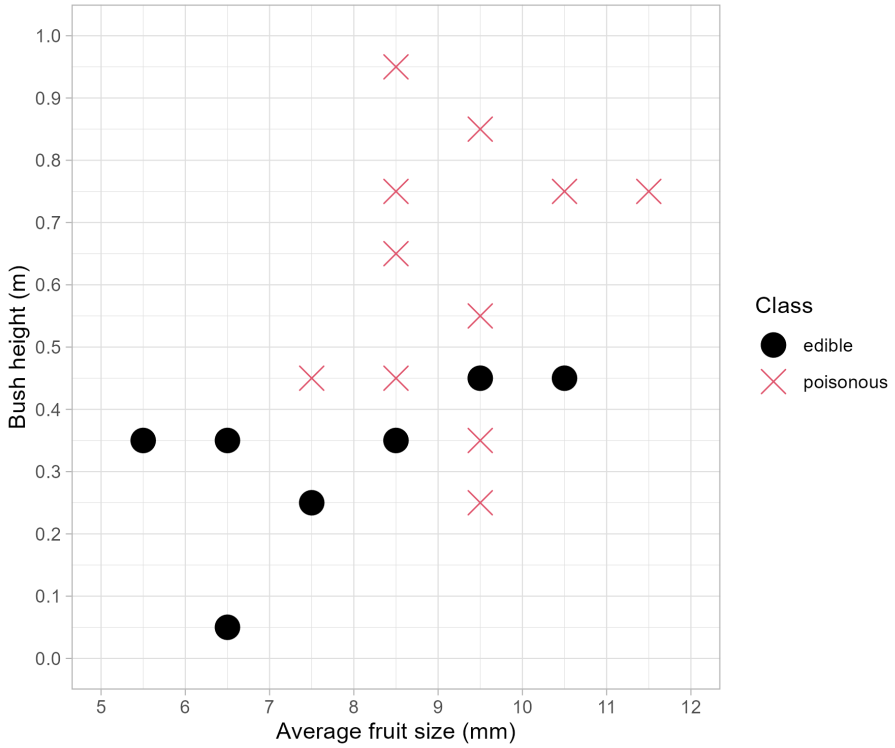

Figure 1: Diagram found in the abandoned hut.

Your Data Mining knowledge immediately comes back to mind, and you decide to build a simple decision tree model to classify fruit as edible or not before you start foraging for food - after all your life depends on it! Fortunately, the old explorer 's notes also contained some formulas that might be useful (available at the end of this question).

Your tasks are:

a) Based on the diagram in Figure 1, calculate the total **Entropy** of the this initial data. [4%]

b) Suppose you decide that your root node will have the expression $(Bush\ height > 0.4m)$. Calculate the **Information gain** of this split. [4%]

c) For the subset of data that have $(Bush\ height > 0.4m) : True$, add a *leaf node*. Use the class of the majority of points that reached this node to determine the node's predicted class. [2%]

d) For the subset of data that have $(Bush\ height > 0.4m) : False$, add an internal node with expression $(Average\ fruit\ size < 9mm)$. Calculate the **Information gain** of this split [3%] and add two leaf nodes after this internal node. Use the class of the majority of points that reached each leaf node to determine the node's predicted class. [4%]

e) Draw your decision tree diagram, including (1) the split criteria in the root and internal nodes, and (2) the number of points from each class and the predicted classes at the leaf nodes. [4%]

f) Calculate the numbers of **True Positives**, **True Negatives**, **False Positives** and **False Negatives** from your decision tree, based on the counts of points at the leaf nodes from the previous item. Assume that "edible" is the positive class. [4%]

g) Draw the **Confusion Matrix** for your model, and estimate the model's **Accuracy**. [4%]

h) Based on the confusion matrix, calculate the **Precision** of your model - i.e., the probability that a fruit is actually edible, if the model predicts it as being edible. [4%]

i) Imagine that you find a plant with $Bush\ height = 0.6m$ and $Average\ fruit\ size = 10mm$. Use your decision tree to decide whether you should eat the berries from this bush or not. [2%]

j) In this exercise we used the same data for training and evaluating our model. Discuss why this may be a problem, and suggest a simple strategy to mitigate it. [5%]

(Total: 40% of the exam marks)

Old Explorer's additional notes:

- *Entropy of a set of class labels:*

$$Entropy(t) = -\left[\ P(yes)\log_2\left(P(yes)\right) + P(no)\log_2\left(P(no)\right)\ \right]$$

- *Conversion from natural logarithm to base-2 logarithm:*

$$\log_2(x) = 0.693\ln(x)$$

- Definition of $0 \times \log_2(0)$:

$$0 \times \log_2(0) = 0$$

- *Impurity after split (n_t is the number of data points in child node t, n_p is the total number of data points in the parent node):*

$$I(children) = \sum_{t=1}^{k} \frac{n_t}{n_p} Entropy(t)$$

- *Information Gain:*

$$Gain = Entropy(parent) - I(children)$$

- *Bayes' formula:*

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$