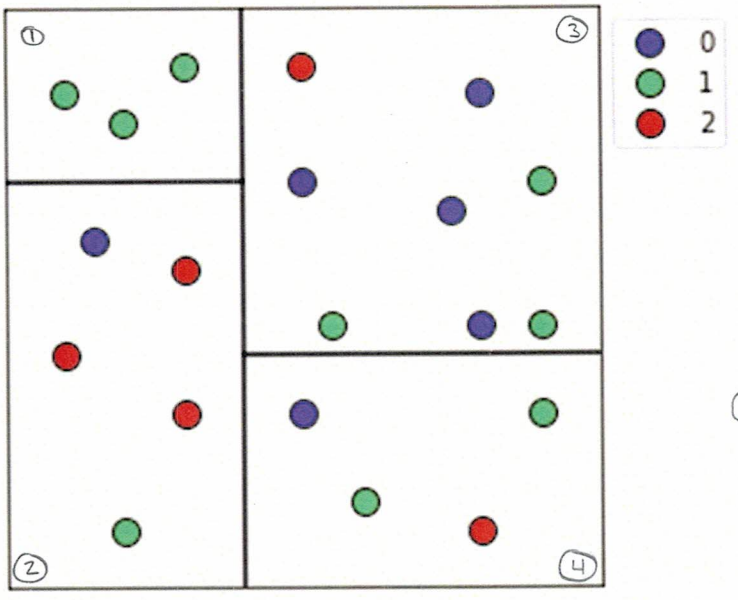
DSCI 35600 – HW 07 (36 pts) Due: Monday, 3/04 **Name** Key

Problem 1. (12 pts) The two plots below show a dataset used to train a decision tree classification model. Each image shows the regions determined by two different tree models. Calculate the Gini index for each model as follows:

- Calculate the Gini index for each leaf node in the model.
- Find the Gini index for the model as a whole by taking a weighted average of the Gini index of the leaf nodes, using the number of observations in each node as the weights.

Round all answers on this problem to 4 decimal places.

Tree Model 1:



$G_1 = 1 - 0^2 - \left(\frac{3}{3}\right)^2 - 0^2 = 0$

$G_2 = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{1}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = \frac{14}{25}$

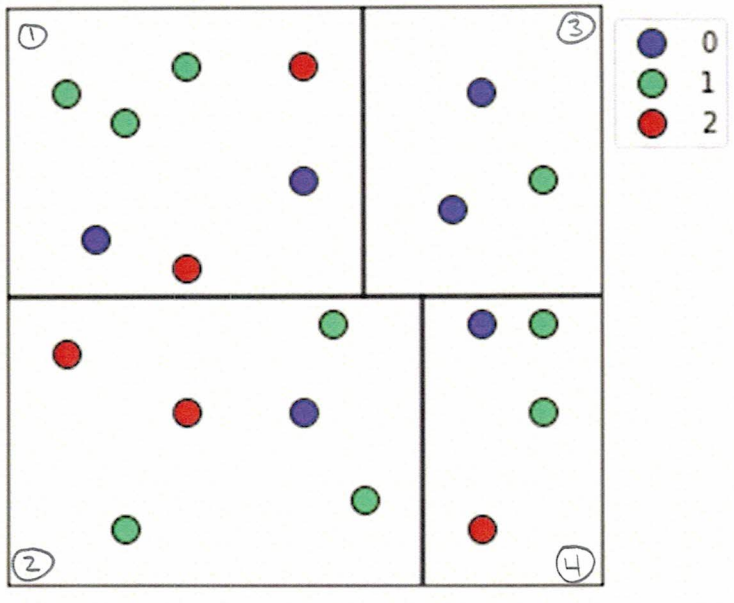
$G_3 = 1 - \left(\frac{1}{8}\right)^2 - \left(\frac{3}{8}\right)^2 - \left(\frac{4}{8}\right)^2 = \frac{38}{64}$

$G_4 = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{2}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = \frac{10}{16}$

$G = \frac{1}{20}\left[3(0) + 5\left(\frac{14}{25}\right) + 8\left(\frac{38}{64}\right) + 4\left(\frac{10}{16}\right)\right]$

Gini Index for Model 1: 0.5025

Tree Model 2:



$G_1 = 1 - \left(\frac{2}{7}\right)^2 - \left(\frac{3}{7}\right)^2 - \left(\frac{2}{7}\right)^2 = \frac{32}{49}$

$G_2 = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{2}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = \frac{22}{36}$

$G_3 = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 - 0^2 = \frac{4}{9}$

$G_4 = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{2}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = \frac{10}{16}$

$G = \frac{1}{20}\left[7\left(\frac{32}{49}\right) + 6\left(\frac{22}{36}\right) + 3\left(\frac{4}{9}\right) + 4\left(\frac{10}{16}\right)\right]$

Gini Index for Model 2: 0.6036

Which model has the better Gini index? Model 1

Problem 2. (12 pts) A decision tree model has been trained on a dataset with 400 observations and 4 features. The structure of the tree is provided by the following output:

```
* Size: 400 [134, 133, 133], Gini: 0.67, Axis:1, Cut: 3.43
|-* Size: 188 [10, 68, 110], Gini: 0.52, Axis:2, Cut: 3.75
|  |-* Size: 98 [2, 62, 34], Gini: 0.48, Axis:0, Cut: 5.34
|  |  |-* Size: 33 [0, 2, 31], Gini: 0.11, Predicted Class: 2
|  |  |-* Size: 65 [2, 60, 3], Gini: 0.14, Predicted Class: 1
|  |-* Size: 90 [8, 6, 76], Gini: 0.27, Axis:3, Cut: 1.55
|  |  |-* Size: 4 [0, 4, 0], Gini: 0.0, Predicted Class: 1
|  |  |-* Size: 86 [8, 2, 76], Gini: 0.21, Predicted Class: 2
|-* Size: 212 [124, 65, 23], Gini: 0.55, Axis:1, Cut: 5.16
   |-* Size: 73 [11, 56, 6], Gini: 0.38, Axis:3, Cut: 3.83
   |  |-* Size: 16 [7, 3, 6], Gini: 0.63, Predicted Class: 0
   |  |-* Size: 57 [4, 53, 0], Gini: 0.13, Predicted Class: 1
   |-* Size: 139 [113, 9, 17], Gini: 0.32, Axis:0, Cut: 7.08
      |-* Size: 120 [111, 0, 9], Gini: 0.14, Predicted Class: 0
      |-* Size: 19 [2, 9, 8], Gini: 0.59, Predicted Class: 1
```

Use the information in this print-out to classify the four observations provided below. Also, write the name of each point (P, Q, R, or S) beneath the leaf node to which it belongs in the tree diagram below.

- P = [3, 5, 4, 6]

  1

- Q = [6, 2, 3, 7]

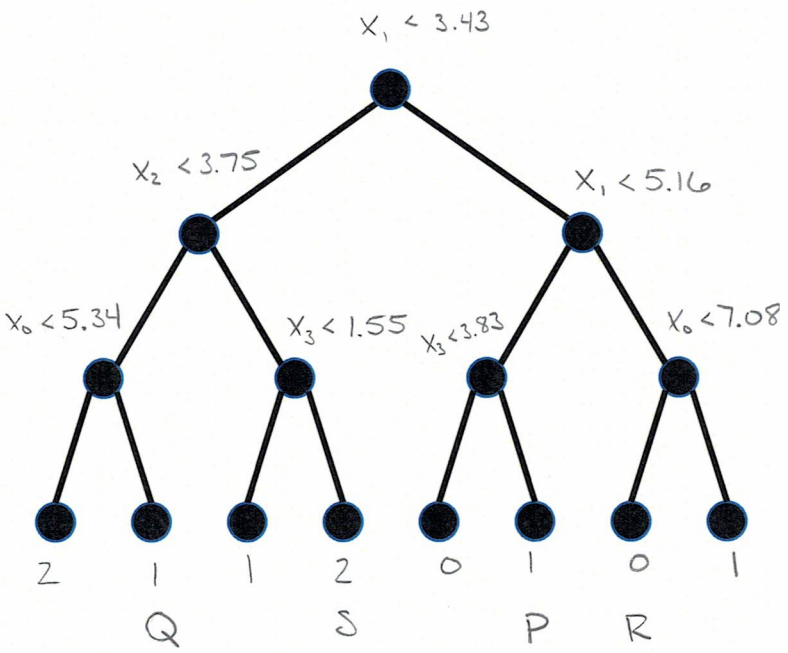
  1

- R = [3, 8, 5, 6]

  0

- S = [7, 1, 4, 3]

  2

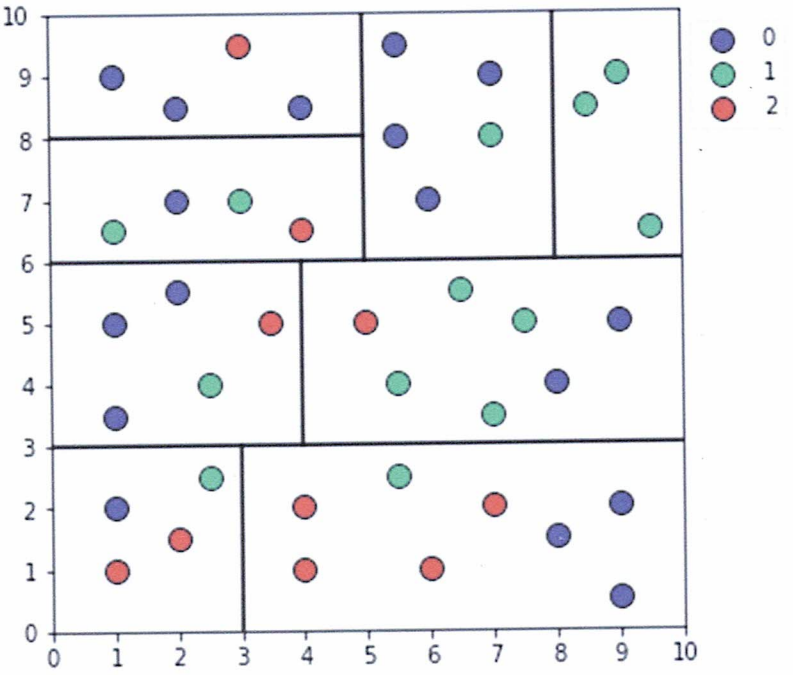


Problem 3. (12 pts) The plot on the right shows a dataset used to train a decision tree classification model. The horizontal and vertical lines represent show where the algorithm decided to split the dataset at each node.

Translate the information contained in this image to the tree structure provided below.

Use 0 to indicate the horizontal axis and 1 to indicate the vertical axis. Use 0, 1, and 2 to denote the classes.

Also, calculate the accuracy of this model, as evaluated on the training set displayed in the image.



Accuracy = $\frac{25}{40} = 0.625$