CSC 35600 – Machine Learning – HW 07 (30 pts) **Name: KEY**

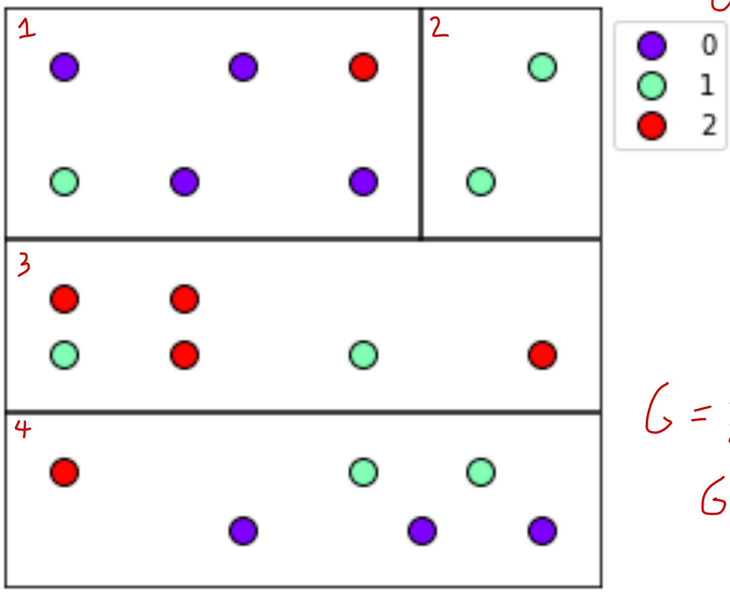
Problem 1. (10 pts) The two plots below show a dataset used to train a decision tree classification model. Each image shows the regions determined by two different tree models. Calculate the Gini index for each model as follows:

- 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.

$1 - B - G - R$

Tree Model 1:



$$G_1 = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{1}{6}\right)^2 - \left(\frac{1}{6}\right)^2 = \frac{1}{2}$$

$$G_2 = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 - \left(\frac{0}{2}\right)^2 = 0$$

$$G_3 = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2 = \frac{4}{9}$$

$$G_4 = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{2}{6}\right)^2 - \left(\frac{1}{6}\right)^2 = \frac{11}{18}$$

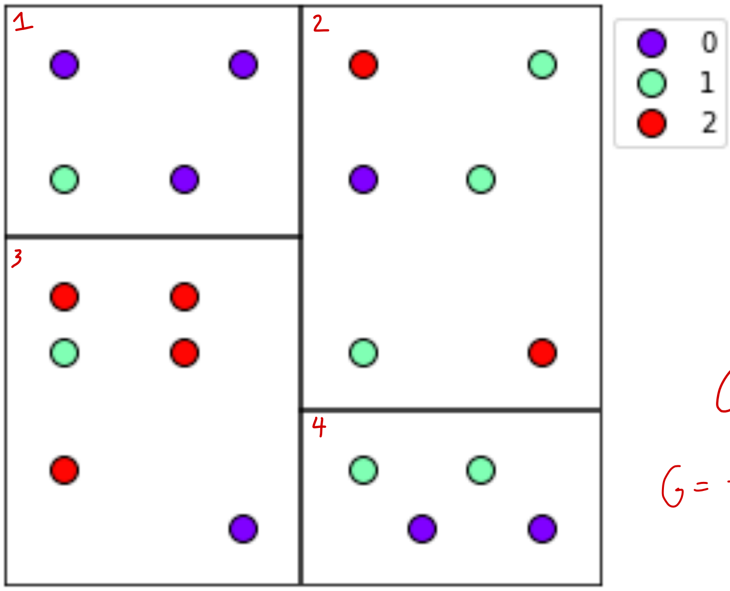
$$G = \frac{1}{20}\left[6\left(\frac{1}{2}\right) + 2(0) + 6\left(\frac{4}{9}\right) + 6\left(\frac{11}{18}\right)\right]$$

$$G = 1 + 0 + \frac{8}{3} + \frac{11}{3}$$

Gini Index for Model 1: 0.4667

Tree Model 2:

$1 - B - G - R$



$$G_1 = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 - (0)^2 = \frac{3}{8}$$

$$G_2 = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{3}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = \frac{11}{18}$$

$$G_3 = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{1}{6}\right)^2 - \left(\frac{4}{6}\right)^2 = \frac{1}{2}$$

$$G_4 = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 - 0^2 = \frac{1}{2}$$

$$G = \frac{1}{20}\left(4\left(\frac{3}{8}\right) + 6\left(\frac{11}{18}\right) + 6\left(\frac{1}{2}\right) + 4\left(\frac{1}{2}\right)\right)$$

Gini Index for Model 2: .5083

Which model has the better Gini index? Model 1

Problem 2. (10 pts) A decision tree model has been trained on a dataset with 400 observations and 4 features. The **print_tree()** method of the tree is called, generating the following output.

```
* Size: 400 [134, 133, 133], Gini: 0.67, Axis:3, Cut: 4.3
  * Size: 127 [90, 36, 1], Gini: 0.42, Axis:0, Cut: 6.64
    * Size: 94 [89, 4, 1], Gini: 0.1, Axis:1, Cut: 5.03
      * Size: 91 [89, 1, 1], Gini: 0.04, Predicted Class: 0
      * Size: 3 [0, 3, 0], Gini: 0.0, Predicted Class: 1
    * Size: 33 [1, 32, 0], Gini: 0.06, Axis:2, Cut: 3.8
      * Size: 3 [1, 2, 0], Gini: 0.44, Predicted Class: 1
      * Size: 30 [0, 30, 0], Gini: 0.0, Predicted Class: 1
  * Size: 273 [44, 97, 132], Gini: 0.61, Axis:2, Cut: 2.68
    * Size: 61 [0, 2, 59], Gini: 0.06, Axis:1, Cut: 1.74
      * Size: 3 [0, 2, 1], Gini: 0.44, Predicted Class: 1
      * Size: 58 [0, 0, 58], Gini: 0.0, Predicted Class: 2
    * Size: 212 [44, 95, 73], Gini: 0.64, Axis:2, Cut: 6.39
      * Size: 180 [44, 94, 42], Gini: 0.61, Predicted Class: 1
      * Size: 32 [0, 1, 31], Gini: 0.06, Predicted Class: 2
```

Use the information in this print-out to classify the four observations provided below.

- [4, 3, 2, 7]

  2

- [8, 4, 7, 3]

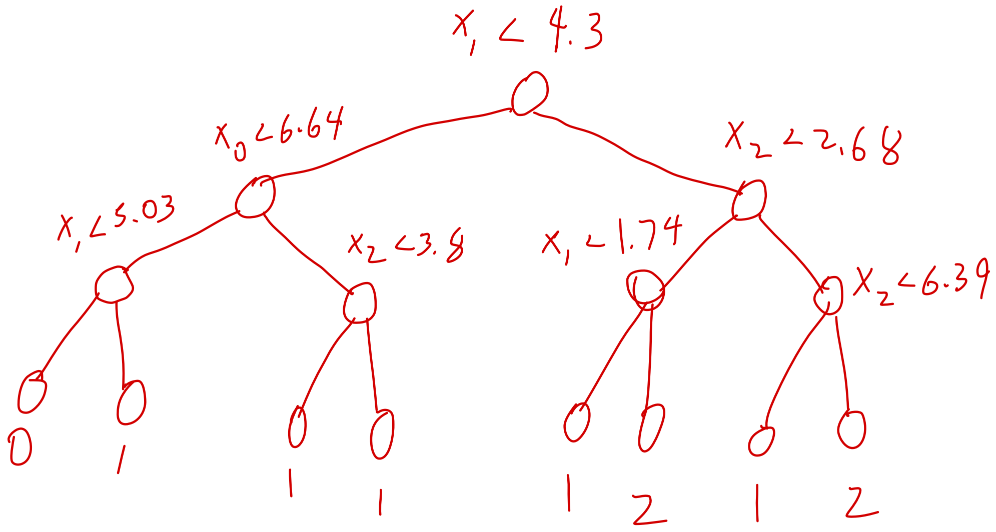
  1

- [3, 2, 7, 2]

  0

- [5, 8, 4, 7]

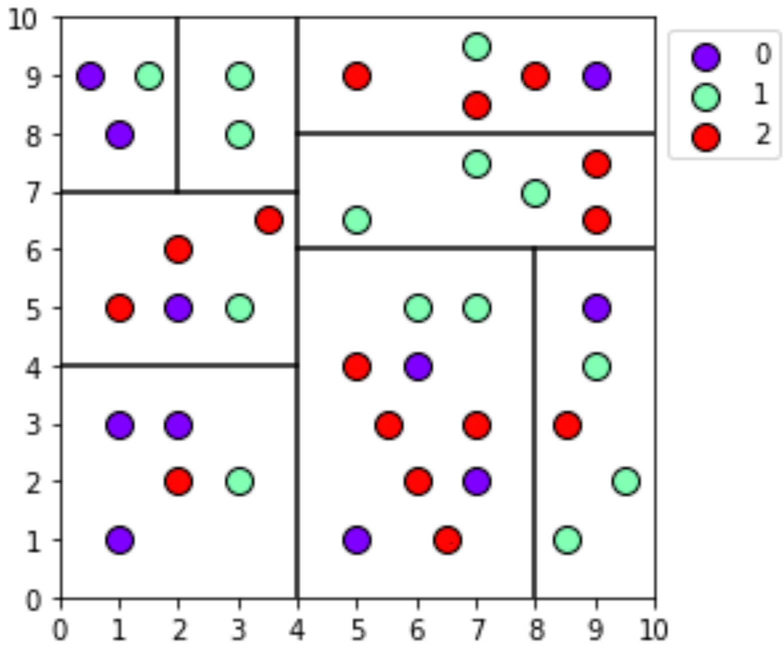
  1



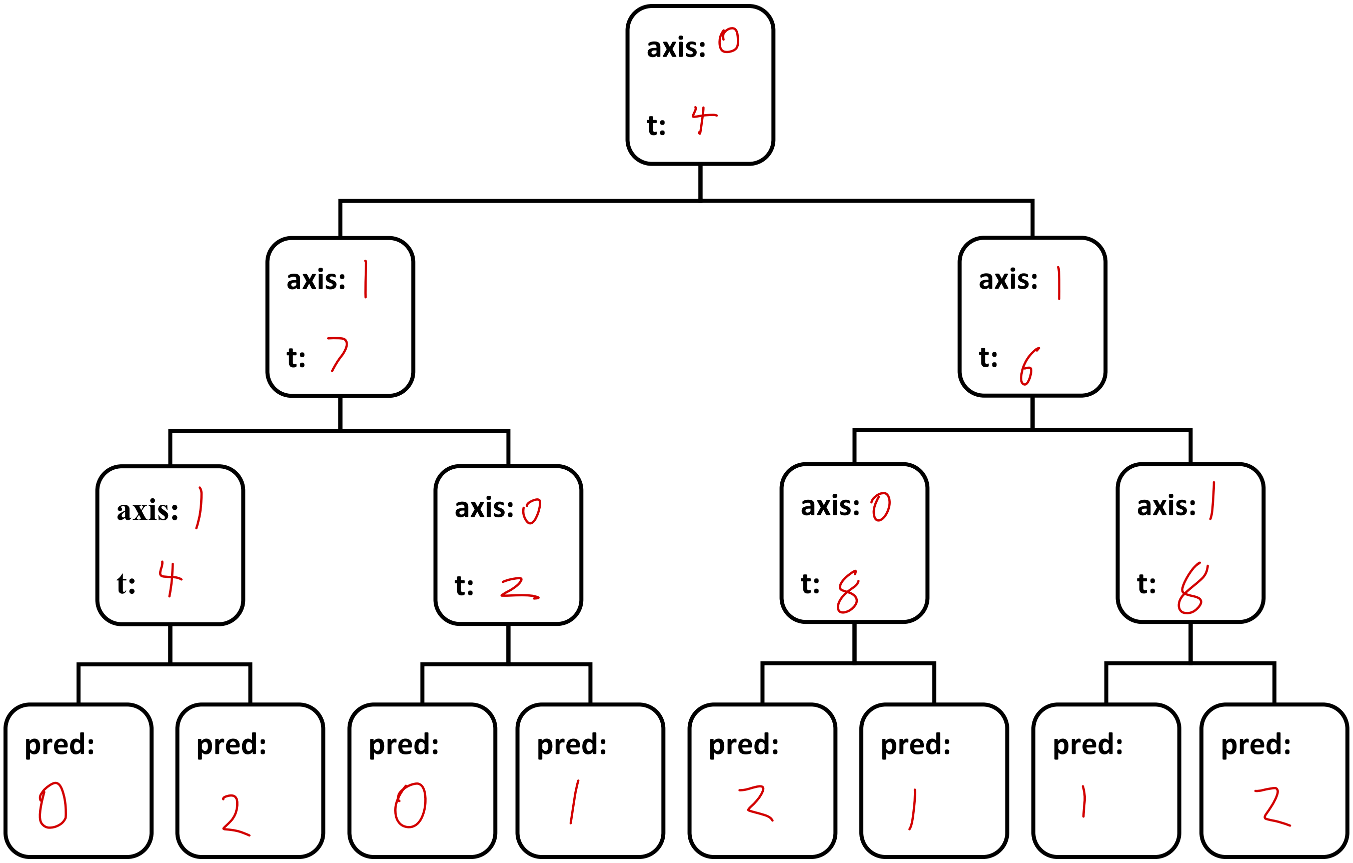
Problem 3. (10 pts) The plot on the right shows a dataset used to train a decision tree classification model. The horizontal and vertical lines represent show where the algorithm decided to split the dataset at each node.

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.

Also, calculate the accuracy of this model, as evaluated on the training set displayed in the image. **Round the accuracy to two decimal places.**



Accuracy = $\frac{24}{40} = 0.6$



axis: 0
t: 4

axis: 1
t: 7

axis: 1
t: 6

axis: 1
t: 4

axis: 0
t: 2

axis: 0
t: 8

axis: 1
t: 8

pred: 0

pred: 2

pred: 0

pred: 1

pred: 2

pred: 1

pred: 1

pred: 2