Test 3. DSCI 35600. Spring 2021.

Problem 1. (10 pts) Write **True** or **False** next to each of the following statements.

a) When comparing potential cuts, a decision tree algorithm selects the cut with the highest Gini score. False

b) When using bagging or pasting to create an ensemble model, individual models are trained on a randomly selected subset of the training set. True

c) When using bagging, sampling is performed without replacement. False

d) PCA is a technique for engineering new relevant features to be added to the original data set. False

e) Each component of a PCA decomposition explains a larger proportion of the variance in the original dataset than any later components in the decomposition. True

f) When using the RandomForestClassifier class to create an ensemble of decision trees, each tree in the ensemble is trained on a different subset of the training set. True

g) Assume we are using the RandomForestClassifier class to create an ensemble of decision trees. When constructing a particular tree model in the ensemble, each splitting step is performed on the feature that will generate the best results at that step. False

h) Increasing the value of the hyper-parameter K in a KNN model will typically make it less likely that the model will overfit. True

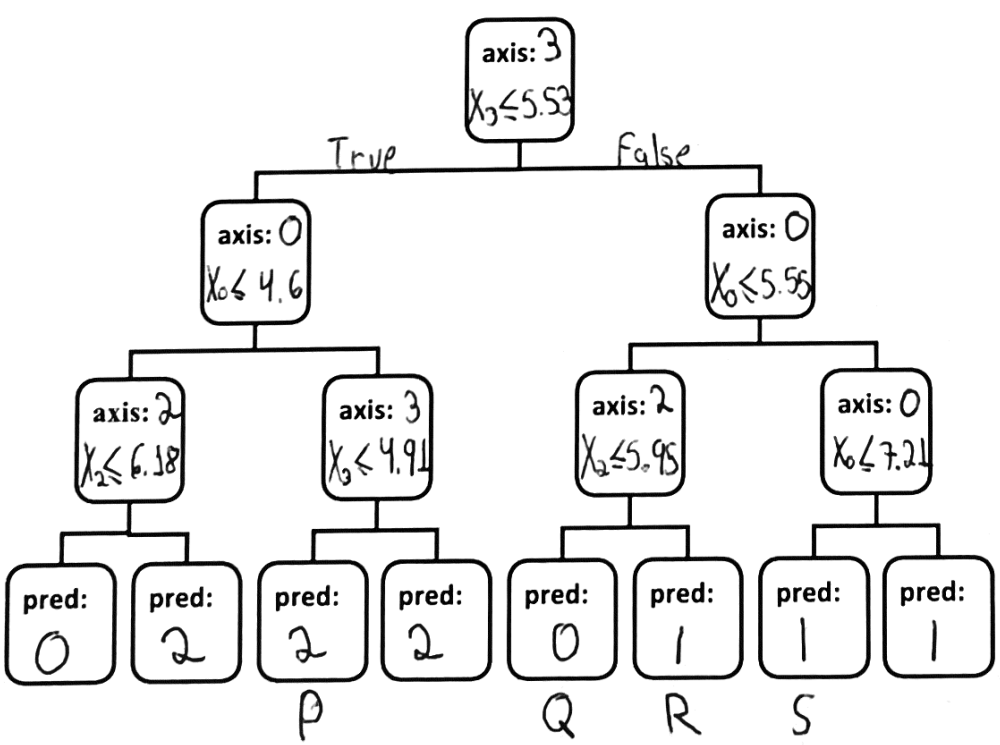
i) Soft-voting can only be used in an ensemble model in which each of the individual models is capable of producing probability estimates. True

j) When selecting a value for the parameter *C* in a logistic regression model, we typically choose the value that results in the greatest accuracy for the training set. True

Problem 2. (16 pts) A decision tree model has been trained on a data set 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:3, Cut: 5.53
  * Size: 217 [79, 14, 124], Gini: 0.54, Axis:0, Cut: 4.6
    * Size: 135 [77, 8, 50], Gini: 0.53, Axis:2, Cut: 6.18
      * Size: 112 [77, 2, 33], Gini: 0.44, Predicted Class: 0
      * Size: 23 [0, 6, 17], Gini: 0.39, Predicted Class: 2
    * Size: 82 [2, 6, 74], Gini: 0.18, Axis:3, Cut: 4.91
      * Size: 67 [1, 2, 64], Gini: 0.09, Predicted Class: 2
      * Size: 15 [1, 4, 10], Gini: 0.48 Predicted Class: 2
  * Size: 183 [55, 119, 9], Gini: 0.48, Axis:0, Cut: 5.55
    * Size: 111 [53, 50, 8], Gini: 0.56, Axis:2, Cut: 5.95
      * Size: 54 [44, 10, 0], Gini: 0.3, Predicted Class: 0
      * Size: 57 [9, 40, 8], Gini: 0.46, Predicted Class: 1
    * Size: 72 [2, 69, 1], Gini: 0.08, Axis:0, Cut: 7.21
      * Size: 19 [2, 16, 1], Gini: 0.28, Predicted Class: 1
      * Size: 53 [0, 53, 0], Gini: 0.0, Predicted Class: 1
```

Use the information in the above printout to fill out the tree below and 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= [5, 9, 2, 3], Q = [1, 0, 4, 7], R = [3, 1, 6, 9], S = [7, 4, 0, 8].**
Hint: **in each node (except the leaf nodes) write the cut, e.g., X3<=5.53**



Problem 3. (12 pts) A dataset X_train, y_train contains 200 samples split over 3 classes. A random forest classifier is trained on this data using the following code:

```
random_forest = RandomForestClassifier(n_estimators=3, max_depth=2, bootstrap='True')
random_forest.fit(X_train,  y_train)
```

The three trees created by the classifier are described below. Consider a new sample given by [2, 1,- 2]. Find the probability of this sample being in each of the three classes, as predicted by the random forest classifier. That is, find the output of the following code:

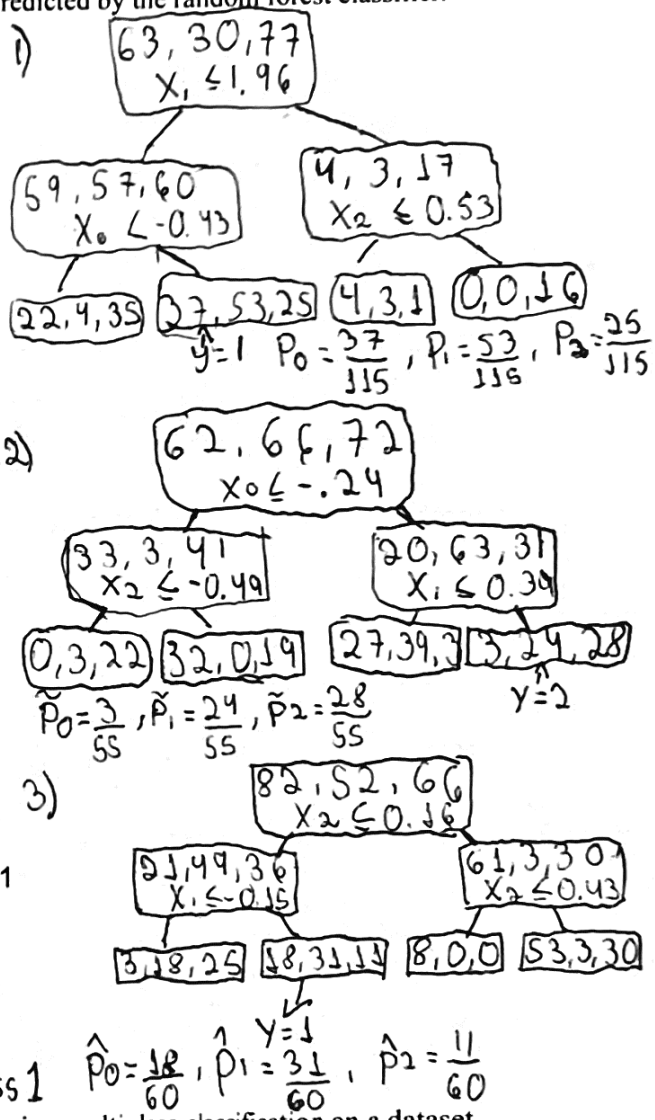
```
Xnew = [[ 2, 1, -2 ]]
random_forest.predict_proba(Xnew )
```

Provide your answer as an array with shape (3,).

1)
- Size: 200 [63, 30, 77], Axis:1 Cut: 1.96
  - Size: 176 [59, 57, 60], Axis: 0 Cut: -0.43
    - Size: 61 [22, 4, 35], Predicted Class: 2
    - Size: 115 [37, 53, 25], Predicted Class: 1
  - Size: 24 [4, 3, 17], Axis: 2 Cut: 0.53
    - Size: 8 [4, 3, 1], Predicted Class: 0
    - Size: 16 [0, 0, 16], Predicted Class: 2

2)
- Size: 200 [62, 66, 72], Axis: 0 Cut: -.24
  - Size: 76 [32, 3, 41], Axis: 2 Cut: -0.49
    - Size: 25 [0, 3, 22], Predicted Class: 2
    - Size: 51 [32, 0, 19], Predicted Class: 0
  - Size: 124 [20, 63, 31], Axis: 1 Cut: 0.39
    - Size: 69 [27, 39, 3], Predicted Class: 1
    - Size: 55 [3, 24, 28], Predicted Class: 2

3)
- Size: 200 [82, 52, 66], Axis: 2 Cut: 0.16
  - Size: 106 [21, 49, 36], Axis: 1 Cut: -0.15
    - Size: 46 [3, 18, 25], Predicted Class: 2
    - Size: 60 [18, 31, 11], Predicted Class: 1
  - Size: 94 [61, 3, 30], Axis: 2 Cut: 0.43
    - Size: 8 [8, 0, 0], Predicted Class: 0
    - Size: 86 [53, 3, 30], Predicted Class: 0



$\left[\frac{1}{3}(P_0+\tilde{P}_0+\hat{P}_0), \frac{1}{3}(P_1,\tilde{P}_1,\hat{P}_1), \frac{1}{3}(P_2,\tilde{P}_2,\hat{P}_2)\right]$

$[0.225, 0.471, 0.303] \Rightarrow$ class 1

Problem 4. (8 pts) Assume we are creating a model for performing multi-class classification on a dataset for which there are 10 categorical labels.

a) If a one-versus-one (OVO) classification scheme is used, how many binary classification models must be created?

$\frac{n(n-1)}{2} = \frac{10(9)}{2} = 45$ models

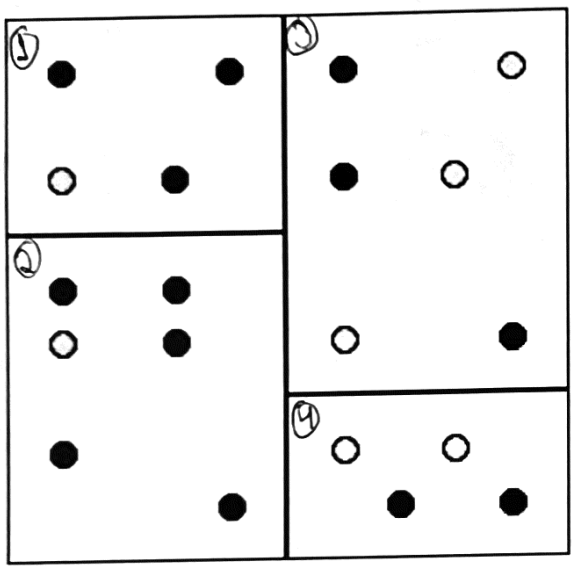
b) If a one-versus-rest (OVR) classification scheme is used, how many binary classification models must be created?

n = 10 models

Problem 5. (12 pts) The two plots below show a dataset used to train a decision tree classification model. The image shows the regions determined by the tree model. Calculate the Gini index for this 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 your answer on this problem to 4 decimal places.



● 0
○ 1
● 2

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

$G_2 = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{1}{6}\right)^2 - \left(\frac{4}{6}\right)^2 = \frac{1}{2}$

$G_3 = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{3}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = \frac{11}{18}$

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

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

$G = 0.508$

Gini index: 0.508

Problem 6. (12 pts) Five probabilistic models are used to create a single ensemble model for a classification task with 4 categorical labels. A single sample is fed into the ensemble, generating the following probability distributions:

	Class 0	Class 1	Class 2	Class 3
Model 1	0.30	0.20	(0.35)	0.15
Model 2	0.15	0.30	(0.35)	0.20
Model 3	(0.40)	0.05	0.25	0.20
Model 4	(0.30)	0.25	0.20	0.25
Model 5	(0.35)	0.25	0.30	0.10

a) Assuming a soft-voting scheme is used, find the probability distribution that this model would return for this sample.

[0.3, 0.21, 0.29, 0.18]

b) Which class would the ensemble predict for this sample if soft-voting is used?

Class 0

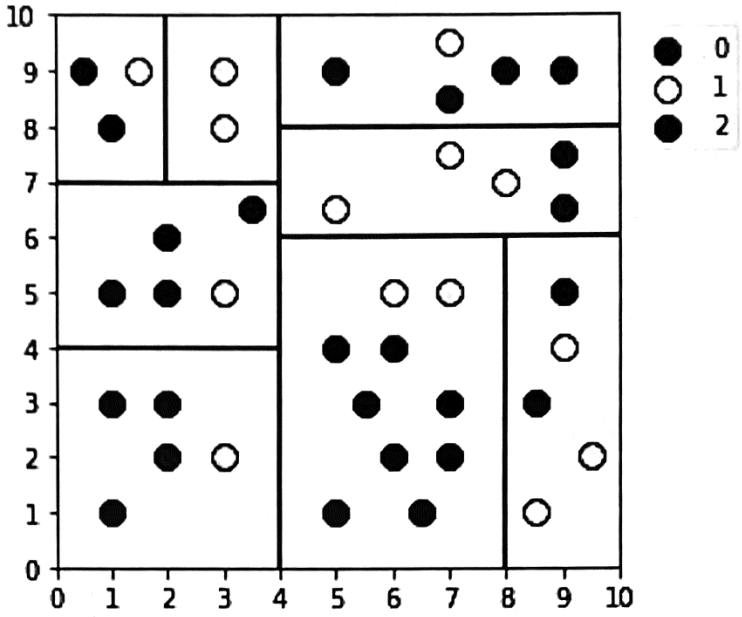
c) Which class would the ensemble predict for this sample if hard-voting is used?

Class 0

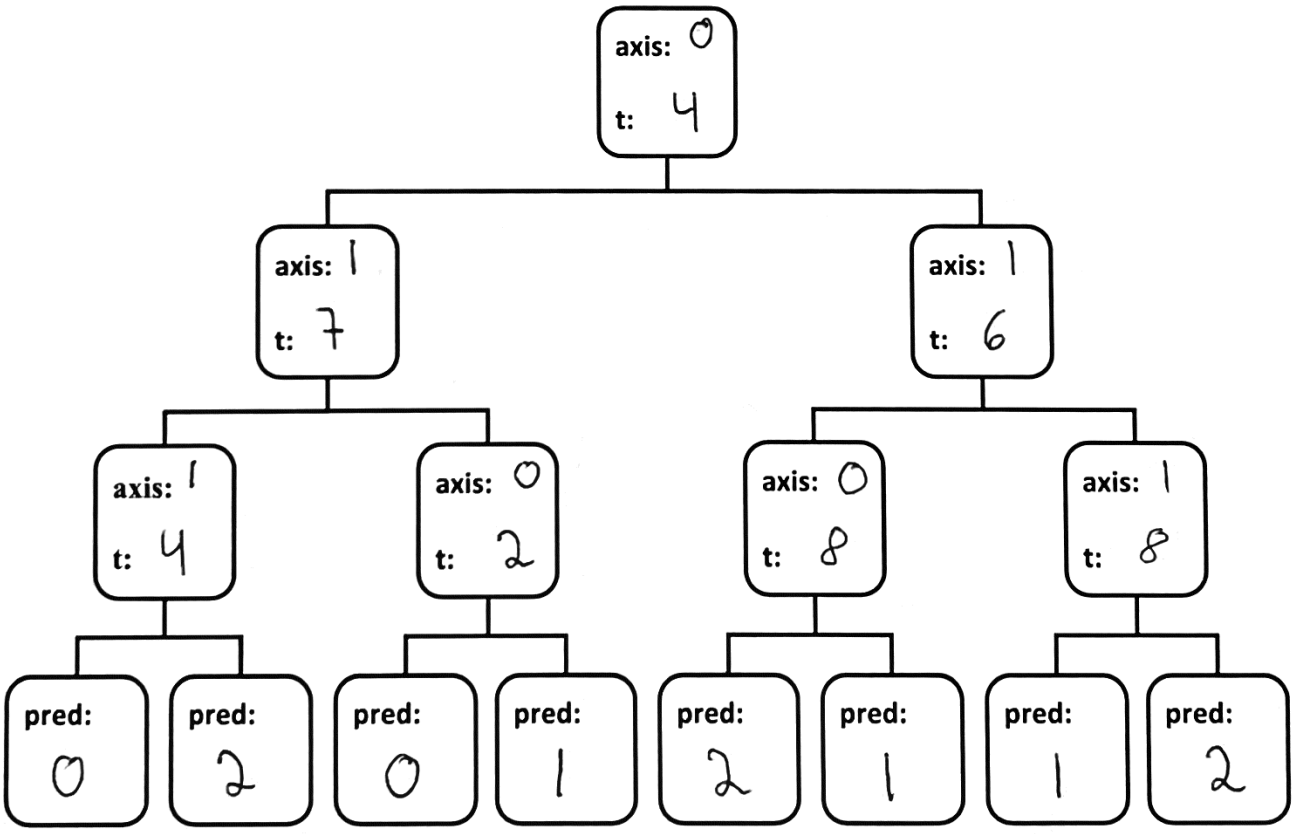
Problem 7. (18 pts) The plot on the right shows a dataset used to train a decision tree classification model. The horizontal and vertical lines represent show where the algorithm decided to split the dataset at each node.

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 = $\frac{24}{40}$



Problem 8. (12 pts) Assume that PCA was performed on a dataset containing 3 features: x1, x2, and x3 . The resulting three principal components are

- pc1 = [-0.920, -0.343, 0.163]
- pc2 = [0.168, 0.080, 0.945]
- pc3 = [0.323, -0.783, -0.138]

The mean of each of the three original features is given by the array [3. 97, 8. 49, 7. 97]

An observation is transformed to new coordinates using the PCA decomposition. The transformed coordinates are given by the array [1. 46, 3. 92, -1. 24].

Convert this observation back into its original x1 , x2, and x3 coordinates.

Round your final answers to 2 decimal places. Box your final answer.

delta1 = 1.46 * pc1 = [-1.3432, -0.50078, 0.23798]

delta2 = 3.92 * pc2 = [0.65856, 0.3136, 3.7044]

delta3 = -1.24 * pc3 = [-0.40052, 0.97092, 0.17112]

$X_1 = 3.97 + delta1 = 2.364$

$X_2 = 8.49 + delta2 = 13.16656$

$X_3 = 7.97 + delta3 = 8.71152$

$[X1, X_2, X_3] = \mu + \Delta1 + \Delta2 + \Delta3$

$= [2.88, 9.27, 12.08$