STAT 5650 Statistical Learning and Data Mining 1 Spring 2020

Name: Michael Huber

Submission Date: 03/13/2020

Homework 3 (03/13/2020)

100 Points — Due Friday 03/13/2020 (via Canvas by 11:59pm)

(i) **Question 1:** This is a continuation of the analyses on the data for three bird species—(Northern) Flicker, (Mountain) Chickadee, and (Red-naped) Sapsucker—plus a bunch of sites at which none of these species of birds are nesting. In the previous homework you analyzed these data using logistic regression and LDA/QDA; in this question I would like you to apply k-Nearest Neighbor classification to the data and to compare your results with the results for LDA, QDA, and logistic regression.

(a) Apply k-NN classification to the combined dataset for all 3 species using 'Nest' as the response variable. What value of k did you select. Compare the results of this classification with the results you obtained for LDA, QDA, and logistic regression in the previous homework.

```
      predicted
actual   1   2
     0 100   6
     1  10  97
```

```
 [,1]                              [,2]
 "Percent Correctly Classified = " "92.49"
 "Specificity = "                  "94.34"
 "Sensitivity = "                  "90.65"
 "Kappa ="                         "0.8498"
 "AUC= "                           "0.9748"
```

Comparison of Methods:
For this KNN on the nest data setting of the value of k to 2, 4 and 5 all returned similar results, but 5 had the highest accuracy at 92.49%.

Below I have included the scores I got from the three different methods we are comparing to KNN. Looking at the scores below and the output of the KNN model it is clear to see that KNN outperforms the other three models. KNN has an accuracy over 92%, while the other three are all around 80%. The Sensitivity and Specificity of the KNN also outperform the other three models. So in this instance, it appears KNN would be the best model to choose for classification of this model.

- LDA: Accuracy = 78.4% Sensitivity = 74.53% SPecificity = 82.24%
- QDA: Accuracy = 81.22% Sensitivity = 80.19% SPecificity = 82.24%

- Logistic Regression: Accuracy = 79.34% Sensitivity = 76.42% Specificity = 82.24%

(b) Now apply k-NN classification to the three datasets for the individual. What values of k did you select. Compare the results of this classification with the results you obtained for LDA, QDA, and logistic regression in the previous homework.

Chickadee:

```
      predicted
actual   1   2
     0 100   6
     1  13  29
```

```
 [,1]                               [,2]
 "Percent Correctly Classified = "  "87.16"
 "Specificity = "                   "94.34"
 "Sensitivity = "                   "69.05"
 "Kappa ="                          "0.6675"
 "AUC= "                            "0.8784"
```

Comparison of Methods:

For k I chose the value 3 because it had the highest accuracy rate at 87.16% given the seed value of 12345

Looking at the results of KNN vs the other three methods it outperforms the other three by around 7% in its accuracy score. However KNN does under perform in the area of Sensitivity. It comes in about 8% below LDA and QDA and about 15% under Logistic Regression. With Specificity however, KNN does much better than the other three. At its high it is around 94% while the other three are in the 60% range.

- KNN, K=3: Accuracy = 87.16% Sensitivity = 69.05% Specificity = 94.34%
- LDA: Accuracy = 80.41% Sensitivity = 85.85% Specificity = 66.67%
- QDA: Accuracy = 79.73% Sensitivity = 86.79% Specificity = 61.90%
- Logistic Regression: Accuracy = 79.73% Sensitivity = 87.74% Specificity = 59.52%

Flicker:

```
      predicted
actual   1   2
     0 105   1
     1  12  11
```

```
 [,1]                              [,2]
 "Percent Correctly Classified = " "89.92"
 "Specificity = "                  "98.11"
 "Sensitivity = "                  "52.17"
 "Kappa ="                         "0.5938"
 "AUC= "                           "0.8997"
```

Comparison of Methods:

In deciding what value of k to use for this data set I tested numbers from 2 to 10 to see how the accuracy performed. All of the accuracy rates were very similar between all of the groups. But each time I ran the test the accuracy rate would change. But some values of k which were 4, 8, 9, and 10 got up over 89% The rest staid between 84% and 88%. I chose 4 as the number for k since it had the greatest accuracy at 89.92%.

Comparing KNN with the other three methods, in this instance shows KNN having the greatest accuracy of the three. but QDA is also up in the 89% range so QDA may be a suitable choice to classify this data, but you would want to look and see if you wanted to be more focused on the Sensitivity or Specificity for the model. If you were more concerned about the Sensitivity than QDA would be the model to choose since it comes in at 98.11% while KNN comes in at 56.52%. However, if you were want to focus more on Specificity then KNN would be the model to choose since it has a value of 97.17%, while QDA has a value of 47.83%.

- KNN, K=4: Accuracy = 89.92% Sensitivity = 56.52% Specificity = 97.17%
- LDA: Accuracy = 85.27% Sensitivity = 93.40% SPecificity = 47.83%
- QDA: Accuracy = 89.15% Sensitivity = 98.11% SPecificity = 47.83%
- Logistic Regression: Accuracy = 82.95% Sensitivity = 93.40% Specificity = 34.78%

Sapsucker:

```
      predicted
```

```
actual  1  2
     0 95 11
     1 11 31
```

```
 [,1]                               [,2]
 "Percent Correctly Classified = " "86.49"
 "Specificity = "                  "85.85"
 "Sensitivity = "                  "88.1"
 "Kappa ="                         "0.6899"
 "AUC= "                           "0.8911"
```

Comparison of Methods:

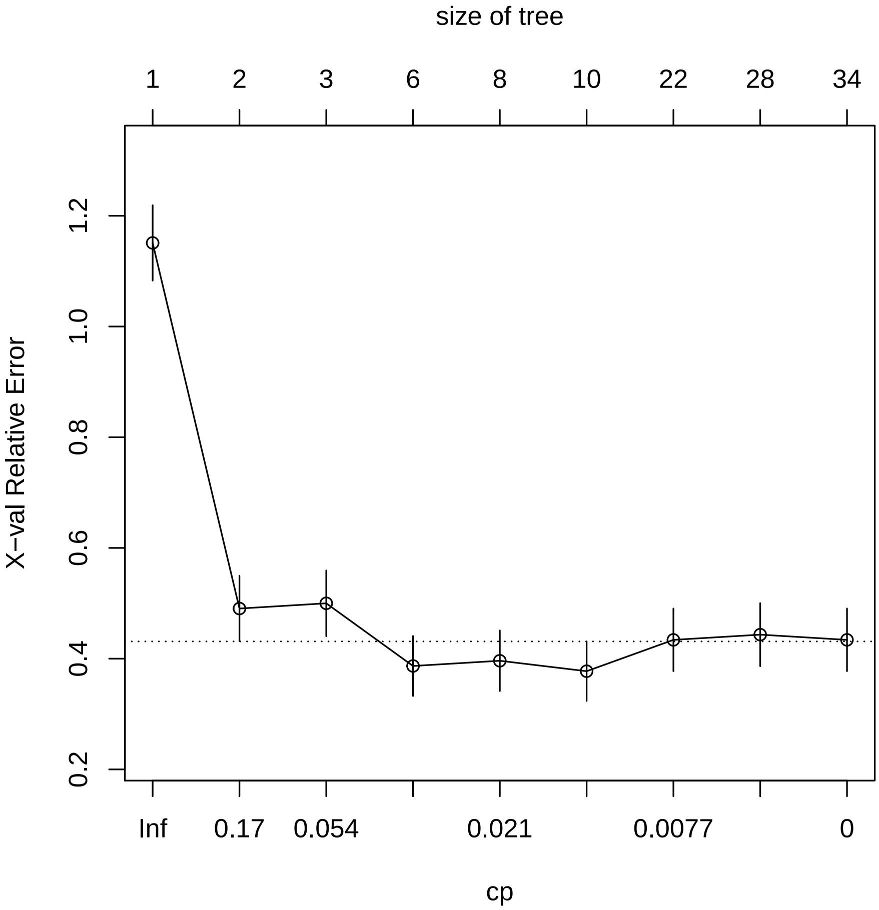
In deciding what value of k to use for this data set I tested numbers from 2 to 10 to see how the accuracy performed. All of the accuracy rates were very similar between all of the groups. Averaging between 82% and 86%. I chose 2 as the value for k since it returned the high of 86.49%. Below I have listed the fro results listed below I also included the values for LDA, QDA and Logistic Regression I got in the previous homework.

Comparing the 4 models, KNN performs the best at 86.49%, It also has the high in Specificity at 91.51%. But it does have the low value of the four models for Sensitivity at 71.43%. While the other three models are up in the 80's for Sensitivity.

- KNN, K=2: Accuracy = 86.49% Sensitivity = 71.43% Specificity = 91.51%
- LDA: Accuracy = 81.76% Sensitivity = 85.85% SPecificity = 71.43%
- QDA: Accuracy = 80.41% Sensitivity = 86.79% SPecificity = 64.29%
- Logistic Regression: Accuracy = 80.41% Sensitivity = 84.91% Specificity = 69.05%

(ii) **Question 2:** This question is also a continuation of the analyses on the data for three bird species— (Northern) Flicker, (Mountain) Chickadee, and (Red-naped) Sapsucker—plus a bunch of sites at which none of these species of birds are nesting. In the previous homework you analyzed these data using logistic regression and LDA/QDA; in this homework I would like you to apply classification trees to the data. The first priority is to come up with accurate classifications of the nest sites, the second priority is to determine important variables to the birds in selecting nest sites, and the third priority is to determine whether the three species can be treated as one species (with regard to selection of bird nest sites) or need to be treated separately.

(a)) First, fit a classification tree to all the data treating the three birds as a single species. Compute the accuracy of your classification using 10-fold cross-validation, and compare it with cross–validated accuracy rates for LDA, QDA, and logistic regression that you computed before.



```
      predicted
actual  1  2
     0 89 17
     1 25 82
```

```
 [,1]                              [,2]
 "Percent Correctly Classified = " "80.28"
 "Specificity = "                  "83.96"
 "Sensitivity = "                  "76.64"
 "Kappa ="                         "0.6058"
 "AUC= "                           "0.8115"
```

Comparison of Methods:

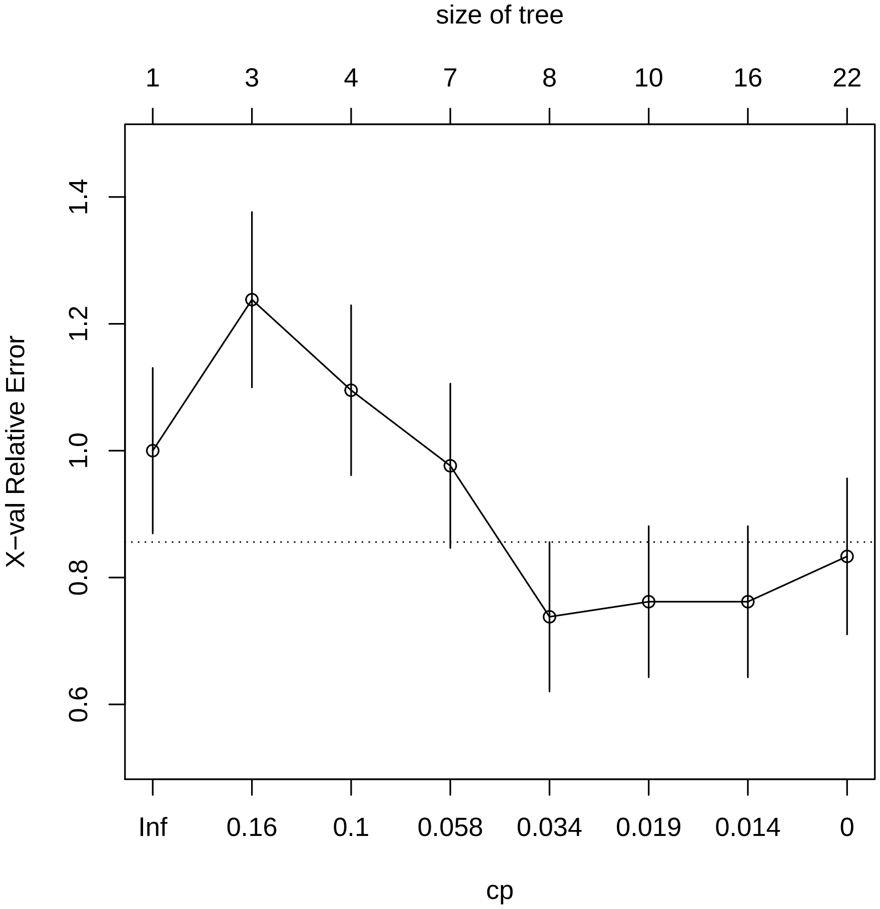
After plotting the cp values I decided to set the cp value to 0.035 as shown on the graph above. When I entered this value it gave me the cross-validated accuracy for the classification tree as 80.28%, shown below.

Comparing this with the other three methods of classifying the data it seems to perform just as well as the other three. They all are around an 80% accuracy score. QDA is the highest at 81.22% and LDA is the lowest at 78.4% but they are all very close to one another.

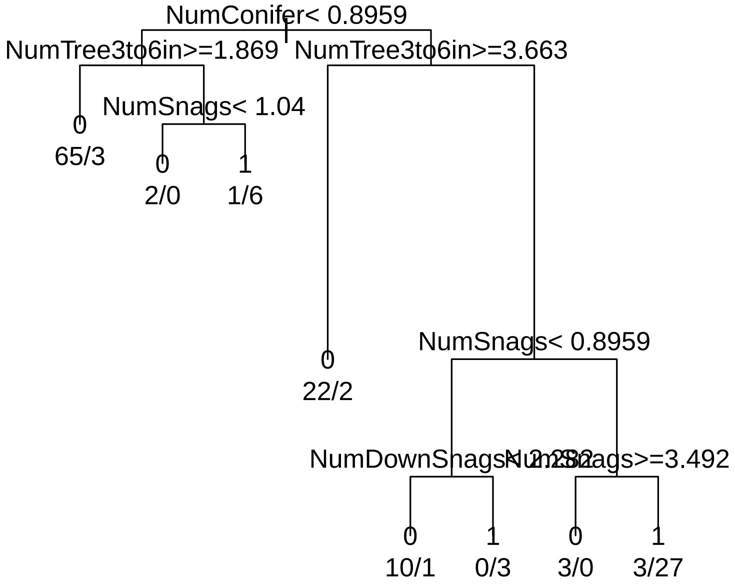
Moving on and looking at their sensitivity scores QDA still performs the best at 80.19%, and LDA has the lowest value at 74.53%. Then with Specificity the Classification tree performs the best at 83.96%, while the other three methods all have the same value of 82.24%

- Classification Tree: Accuracy = 80.28% Sensitivity = 76.64% Specificity = 83.96%
- LDA: Accuracy = 78.4% Sensitivity = 74.53% SPecificity = 82.24%
- QDA: Accuracy = 81.22% Sensitivity = 80.19% SPecificity = 82.24%
- Logistic Regression: Accuracy = 79.34% Sensitivity = 76.42% Specificity = 82.24%

(b) Fit classification trees for each bird species separately and, again, compute estimates of the accuracies by 10-fold cross validation. Qualitatively compare the classification trees for the three species.

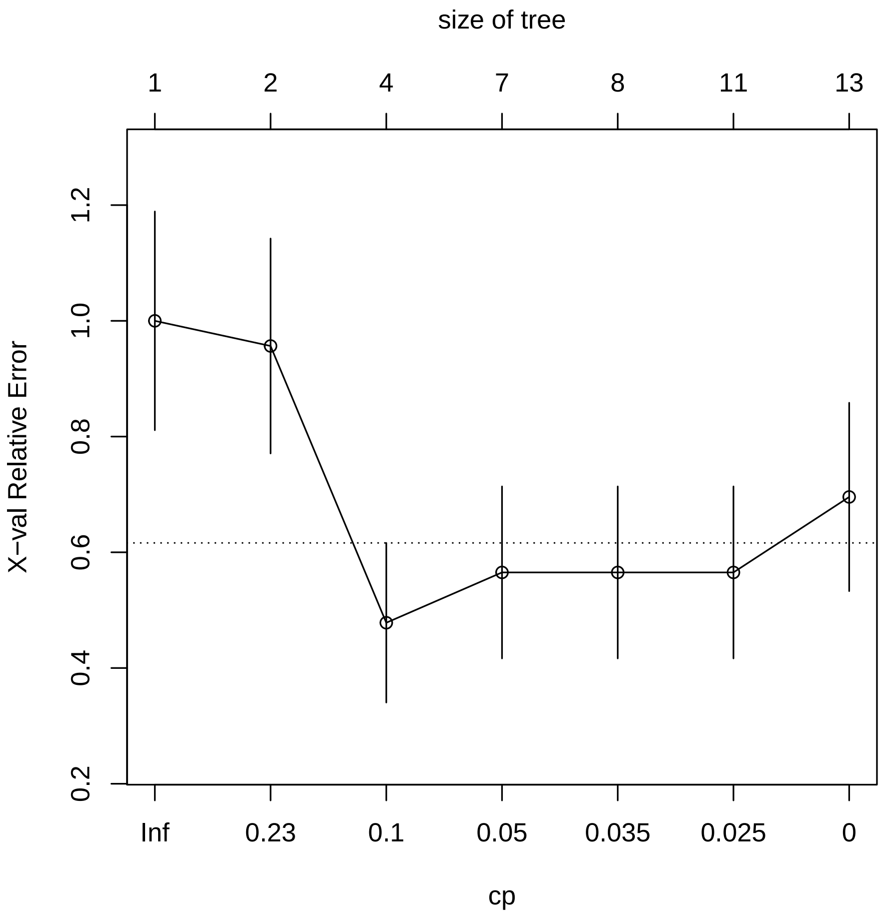

size of tree
1
3
4
7
8
10
16
22
X−val Relative Error
1.4
1.2
1.0
0.8
0.6
Inf
0.16
0.1
0.058
0.034
0.019
0.014
0
cp

Chickadee:

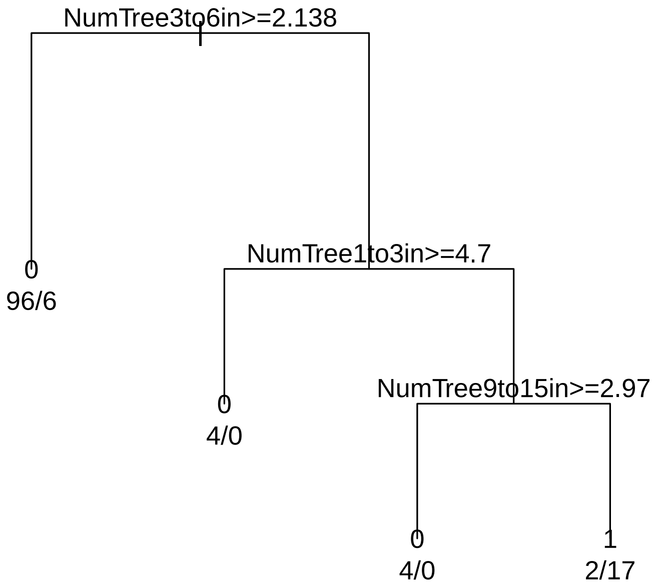


```
      predicted
actual  1  2
     0 96 10
     1 20 22
```

```
 [,1]                               [,2]
 "Percent Correctly Classified = " "79.73"
 "Specificity = "                  "90.57"
 "Sensitivity = "                  "52.38"
 "Kappa ="                         "0.4627"
 "AUC= "                           "0.7345"
```

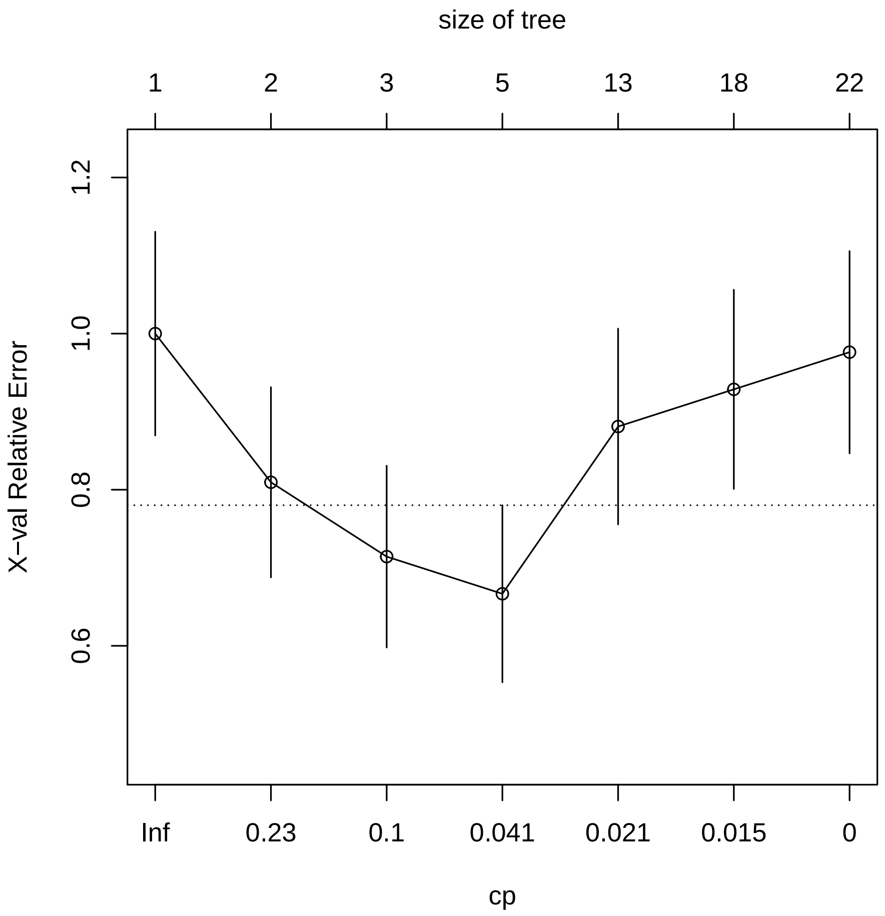

size of tree
1
2
4
7
8
11
13
X−val Relative Error
0.2
0.4
0.6
0.8
1.0
1.2
Inf
0.23
0.1
0.05
0.035
0.025
0
cp

Flicker:

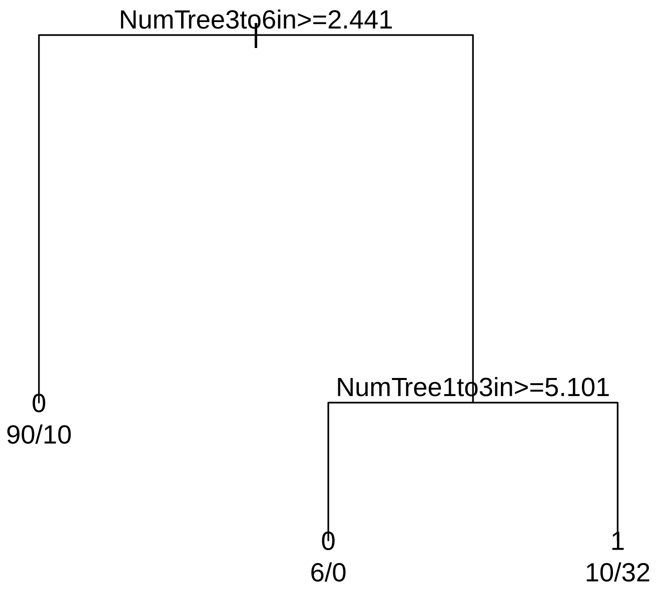


```
      predicted
actual   1   2
     0 100   6
     1   7  16
```

```
 [,1]                               [,2]
 "Percent Correctly Classified = " "89.92"
 "Specificity = "                  "94.34"
 "Sensitivity = "                  "69.57"
 "Kappa ="                         "0.6501"
 "AUC= "                           "0.7555"
```


size of tree
1
2
3
5
13
18
22
X−val Relative Error
1.2
1.0
0.8
0.6
Inf
0.23
0.1
0.041
0.021
0.015
0
cp

Sapsucker:



```
      predicted
actual  1  2
     0 96 10
     1 11 31
```

```
 [,1]                                 [,2]
 "Percent Correctly Classified = " "85.81"
 "Specificity = "                  "90.57"
 "Sensitivity = "                  "73.81"
 "Kappa ="                         "0.6484"
 "AUC= "                           "0.7613"
```

Comparison of the Three Species

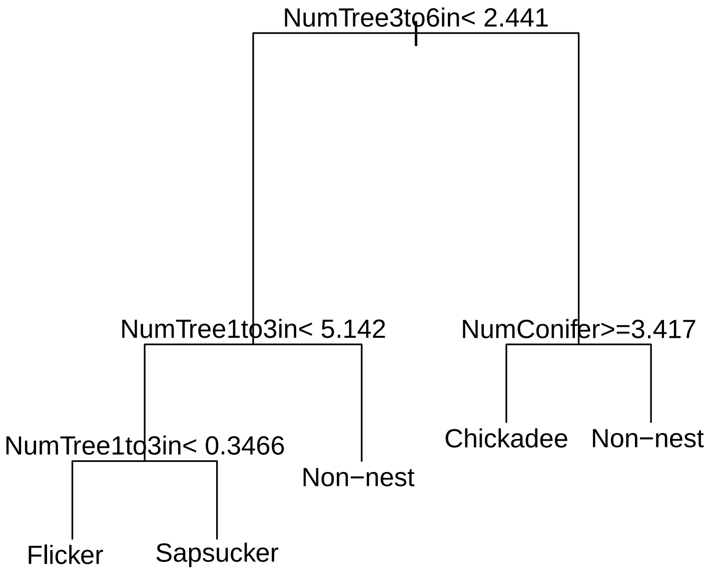
Looking at the output of the three cross validated classification trees Flicker has the greatest accuracy at 89.92%. Looking at the the Tree for flicker it only uses 4 nodes, with Breaks along the variable NumTree3to6in as the first break, NumTree1to3in as the second break and NumTree9to15in as the final break.3 of the nodes are 0 and the final node classifies 1.

Sapsucker has the second greatest accuracy rate at 85.81%, and it ends in three nodes and splits on two variables. NumTree3to6in, and NumTree1to3in. The first two nodes classify Nest values that have a 0 and one node that classifies 1 as a Nest variable.

Finally Chickadee has the lowest accuracy score at 79.73%, but out of the three species it has the most nodes at 8. The tree splits along the variables NumConifer, NumTree3to6in 2x, NumSnags 2x, NumDownSnags, and then one more. The end nodes are 0, 0, 1, 0, 0, 1, 0, 1.

(c) Another way to get at the issue of the similarity of the bird species might be to do the following:

- Fit a classification tree to the combined data using Species as the response variable.



- Look at the cross-validated confusion matrix for the classification tree to see where the misclassifications are occurring.

```
Confusion Matrix and Statistics

           actual
predicted   Chickadee Flicker Non-nest Sapsucker
  Chickadee         0       3        2        10
  Flicker           3       5        1         3
  Non-nest         20       6       94        12
  Sapsucker        19       9        9        17

Overall Statistics

               Accuracy : 0.5446
                 95% CI : (0.4752, 0.6128)
```

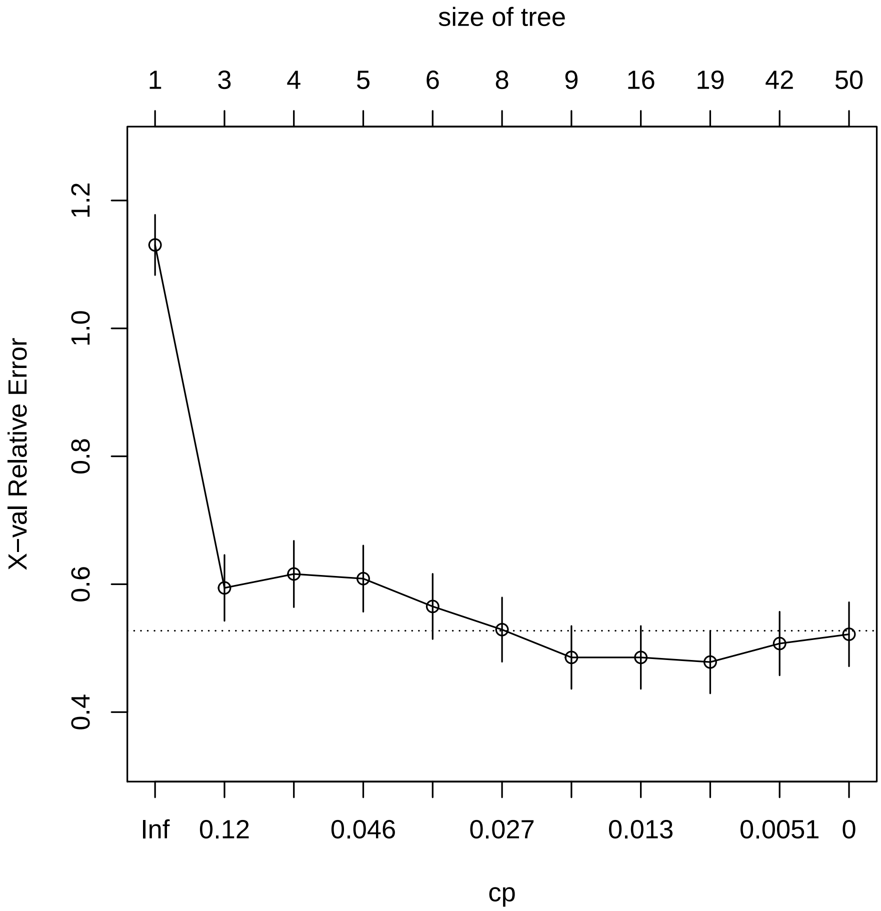
```
    No Information Rate : 0.4977
    P-Value [Acc > NIR] : 0.0964325

                  Kappa : 0.2674

 Mcnemar's Test P-Value : 0.0004188

Statistics by Class:

                     Class: Chickadee Class: Flicker Class: Non-nest
Sensitivity                   0.00000        0.21739          0.8868
Specificity                   0.91228        0.96316          0.6449
Pos Pred Value                0.00000        0.41667          0.7121
Neg Pred Value                0.78788        0.91045          0.8519
Prevalence                    0.19718        0.10798          0.4977
Detection Rate                0.00000        0.02347          0.4413
Detection Prevalence          0.07042        0.05634          0.6197
Balanced Accuracy             0.45614        0.59027          0.7658
                     Class: Sapsucker
Sensitivity                   0.40476
Specificity                   0.78363
Pos Pred Value                0.31481
Neg Pred Value                0.84277
Prevalence                    0.19718
Detection Rate                0.07981
Detection Prevalence          0.25352
Balanced Accuracy             0.59419
```

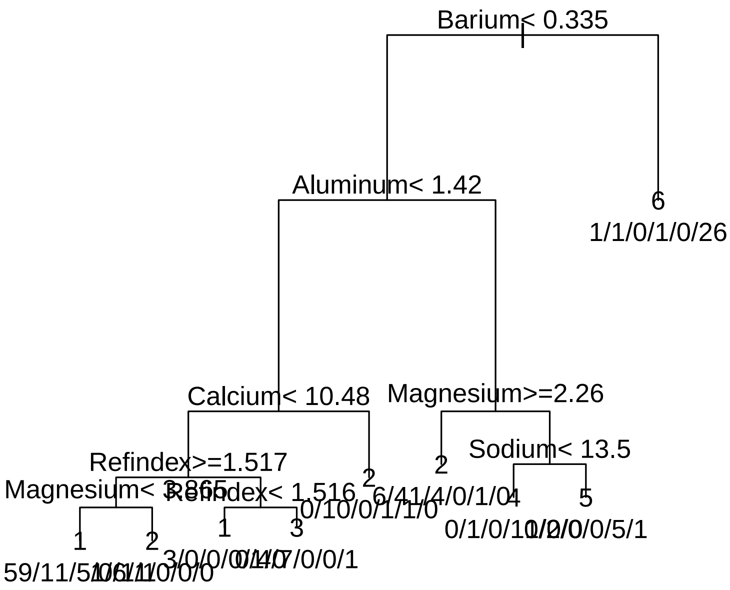
Summary:

Lokking at the balanced accuracy for each of the Species it appears that the only one the model detects with an worth while accuracy is Non-nest and even that only comes in at 76.58%. Overall the model had an accuraccy score of 49.77% which is not very good at all.

(iii) **Question 3:** This problem concerns the forensic glass data set labeled "Glass.csv" in Canvas. There are six different types of glass, coded 1—6, and nine measured variables. The first of the measured variables is the refractive index of the glass, and the remaining eight are weight percentages of eight chemical elements. The purpose of the analysis is to classify the six types of glass using the refractive index and the chemical percentages.

(a) Fit a classification tree to the data using the 1-SE rule or choosing a tree just a little smaller or larger than the one selected by the 1-SE rule. Briefly summarize the tree.





Summary: Based on the above cp plot I chose 0.018 as the value cp to be plugged into the classification tree. Looking at the tree it produces, the first major split is on the variable Barium, when it has a value less than 0.335. If the value is > than 0.335 then it is directed to the glass type 6, otherwise it drops down to the rest of the tree.

The next major split is on Aluminum that is less than 1.42. Going right the tree then branches on Magnesum >=2.26 and then Sodum < 13.5. Traveling back up to where it splits on Aluminum if we follow the left branch the next break is Calcium < 10.48, then going left it breaks on the Refindex >= 1.517 and then it breaks on Magnesum, to the left and the Refindex again to the right. Finally if we were to follow the right branch after the break on Calcium it takes us to the end of the branch that classifies glass in group 2.

In totoal it looks like there are two nodes for glass type 1, three nodes for glass type 2, one node for glass type 3, one node for glass type 4, one node for glass type five and one node for glass type .

(b) Compute the 10-fold cross-validated confusion matrix. If you have trouble doing this, you may have to consider eliminating some types of glass or collapsing categories of glass that may be similar and have small numbers of observations.

```
Confusion Matrix and Statistics

         actual
predicted  1  2  3  4  5  6
        1 51 16  8  0  2  1
        2 16 53  6  3  4  2
        3  2  2  3  0  0  0
        4  0  2  0  8  0  0
        5  0  2  0  1  3  0
        6  1  1  0  1  0 26

Overall Statistics

               Accuracy : 0.6729
                 95% CI : (0.6056, 0.7353)
    No Information Rate : 0.3551
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5434

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6
Sensitivity            0.7286   0.6974  0.17647  0.61538  0.33333   0.8966
Specificity            0.8125   0.7754  0.97970  0.99005  0.98537   0.9838
Pos Pred Value         0.6538   0.6310  0.42857  0.80000  0.50000   0.8966
```

```
Neg Pred Value        0.8603   0.8231  0.93237  0.97549  0.97115   0.9838
Prevalence            0.3271   0.3551  0.07944  0.06075  0.04206   0.1355
Detection Rate        0.2383   0.2477  0.01402  0.03738  0.01402   0.1215
Detection Prevalence  0.3645   0.3925  0.03271  0.04673  0.02804   0.1355
Balanced Accuracy     0.7705   0.7364  0.57808  0.80272  0.65935   0.9402
```