zl9901 homework5 CSCI620

1 a) The Gini index for the overall collection of training examples is

$$1-(\frac{1}{2})^2-(\frac{1}{2})^2=0.5$$

b) The Gini index for the Gender attribute is

Male:

C0	6
C1	4
Gini=0.48	

$$1-(\frac{6}{6+4})^2-(\frac{4}{6+4})^2=0.48$$

Female:

C0	4
C1	6
Gini=0.48	

$$1-(\frac{4}{4+6})^2-(\frac{6}{4+6})^2=0.48$$

$$\text{GenderGini}=\frac{1}{2}*0.48+\frac{1}{2}*0.48=0.48$$

c) The Gini index for the Car Type attribute is

Family:

C0	1
C1	3
Gini=0.375	

$$1-(\frac{1}{4})^2-(\frac{3}{4})^2=0.375$$

Sports:

C0	8
C1	0
Gini=0.0	

$$1-(\frac{8}{8})^2-(\frac{0}{8})^2=0.0$$

Luxury:

C0	1
C1	7
Gini=0.21875	

$$1-(\frac{1}{8})^2-(\frac{7}{8})^2=0.21875$$

$$\text{CarTypeGini} = \frac{4}{20} * 0.375 + \frac{8}{20} * 0.21875 = 0.1625$$

d) The Gini index for the Shirt Size attribute is

Small:

C0	3
C1	2
Gini=0.48	

$$1-(\frac{3}{5})^2-(\frac{2}{5})^2=0.48$$

Medium:

C0	3
C1	4
Gini=0.4898	

$$1-(\frac{3}{7})^2-(\frac{4}{7})^2=0.4898$$

Large:

C0	2
C1	2
Gini=0.5	

$$1-(\frac{2}{4})^2-(\frac{2}{4})^2=0.5$$

Extra Large:

C0	2
C1	2
Gini=0.5	

$$1-(\frac{2}{4})^2-(\frac{2}{4})^2=0.5$$

$$\text{ShirtSizeGini} = \frac{5}{20} * 0.48 + \frac{7}{20} * 0.4898$$

$$+ \frac{4}{20} * 0.5 + \frac{4}{20} * 0.5 = 0.4914$$

e)

Since Gini index for Car Type is the lowest among 4 attributes, Car Type attribute is better.

CarTypeGini < GenderGini < ShirtSizeGini

2

x	0.5	0.3	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	-	-	+	+	+	-	-	+	-	-
distance	20.25	22.09	0.25	0.16	0.01	0.04	0.09	0.25	4.0	20.25
1/distance	0.049	0.045	4	6.25	100	25	11.11	4	0.25	0.049

a) Using majority vote

1- nearest neighbors: 4.9 should be only one nearest neighbor, the prediction should be '+'

3- nearest neighbors: According to majority vote, the final prediction should be '-'

4.9	5.2	5.3
+	-	-

5- nearest neighbors: According to majority vote, the final prediction should be '+'

4.6	4.9	5.2	5.3	5.5
+	+	-	-	+

Or

4.5	4.6	4.9	5.2	5.3
+	+	+	-	-

9- nearest neighbors: According to majority vote, the final prediction should be '-'

0.5	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
-	+	+	+	-	-	+	-	-

b) Using distance-weighted voting approach

1- nearest neighbors: 4.9 should be only one nearest neighbor, the prediction should be '+'

Positive Weights=100

3- nearest neighbors: According to distance-weighted, the final prediction should be '+'

4.9	5.2	5.3
+	-	-

Positive Weights=100

Negative Weights=36.11

5- nearest neighbors: According to majority vote, the final prediction should be '+'

4.6	4.9	5.2	5.3	5.5
+	+	-	-	+

Or

4.5	4.6	4.9	5.2	5.3
+	+	+	-	-

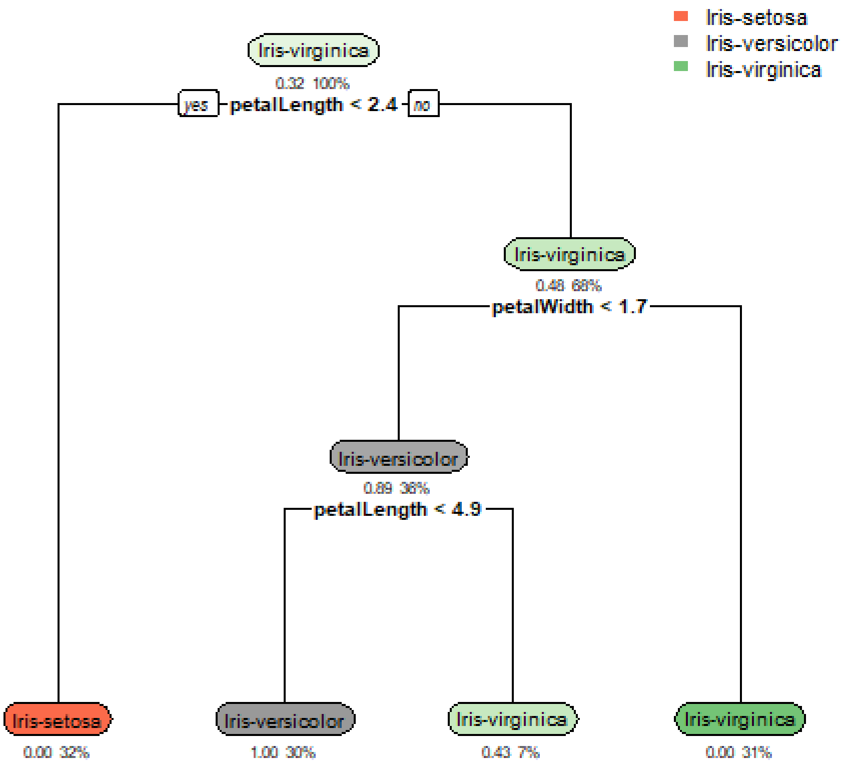
For first table:
Positive Weights=110.25
Negative Weights=36.11
For second table:
Positive Weights=110.25
Negative Weights=36.11

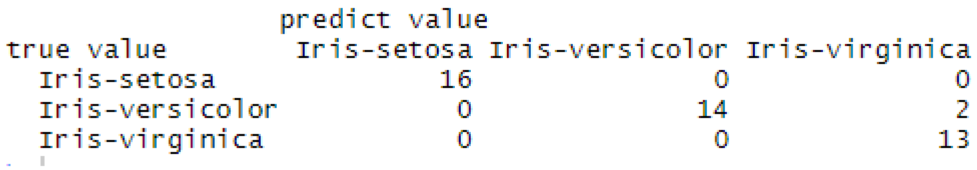
9- nearest neighbors: According to majority vote, the final prediction should be ‘+’

0.5	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
-	+	+	+	-	-	+	-	-

Positive Weights=4+6.25+100+4=114.25
Negative Weights=0.049+25+11.11+0.25+0.049=36.458

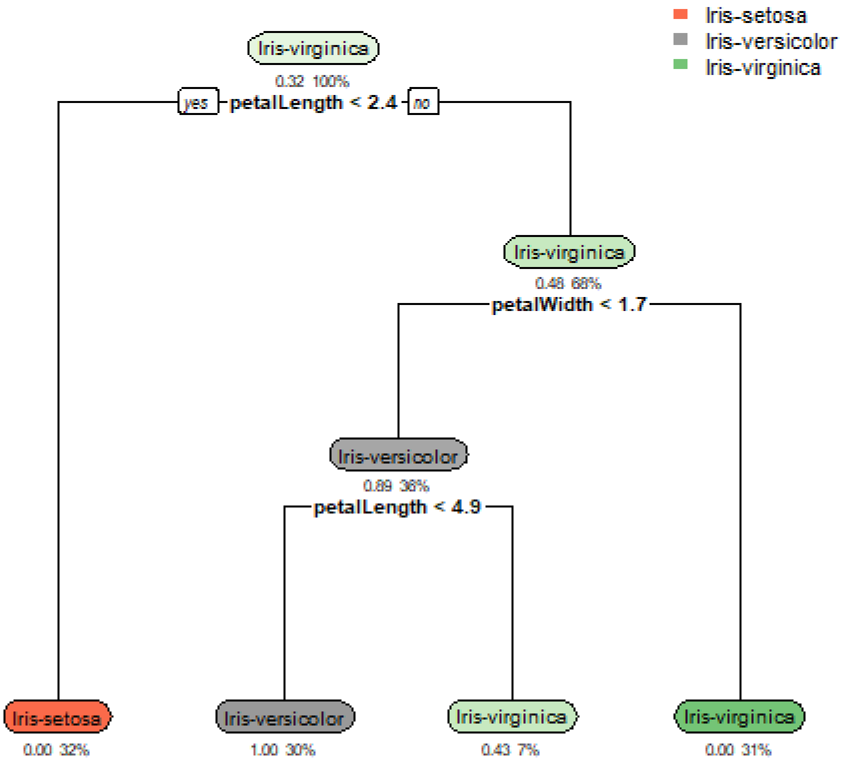
before pruning





```
                 predict value
true value        Iris-setosa Iris-versicolor Iris-virginica
  Iris-setosa              16               0              0
  Iris-versicolor           0              14              2
  Iris-virginica            0               0             13
```

after pruning



Iris-setosa
Iris-versicolor
Iris-virginica
Iris-virginica
0.32 100%
yes
petalLength < 2.4
no
Iris-virginica
0.48 68%
petalWidth < 1.7
Iris-versicolor
0.89 36%
petalLength < 4.9
Iris-setosa
0.00 32%
Iris-versicolor
1.00 30%
Iris-virginica
0.43 7%
Iris-virginica
0.00 31%