CENG414-Introduction to Data Mining

Programming HW1

Task 2-Decision Tree

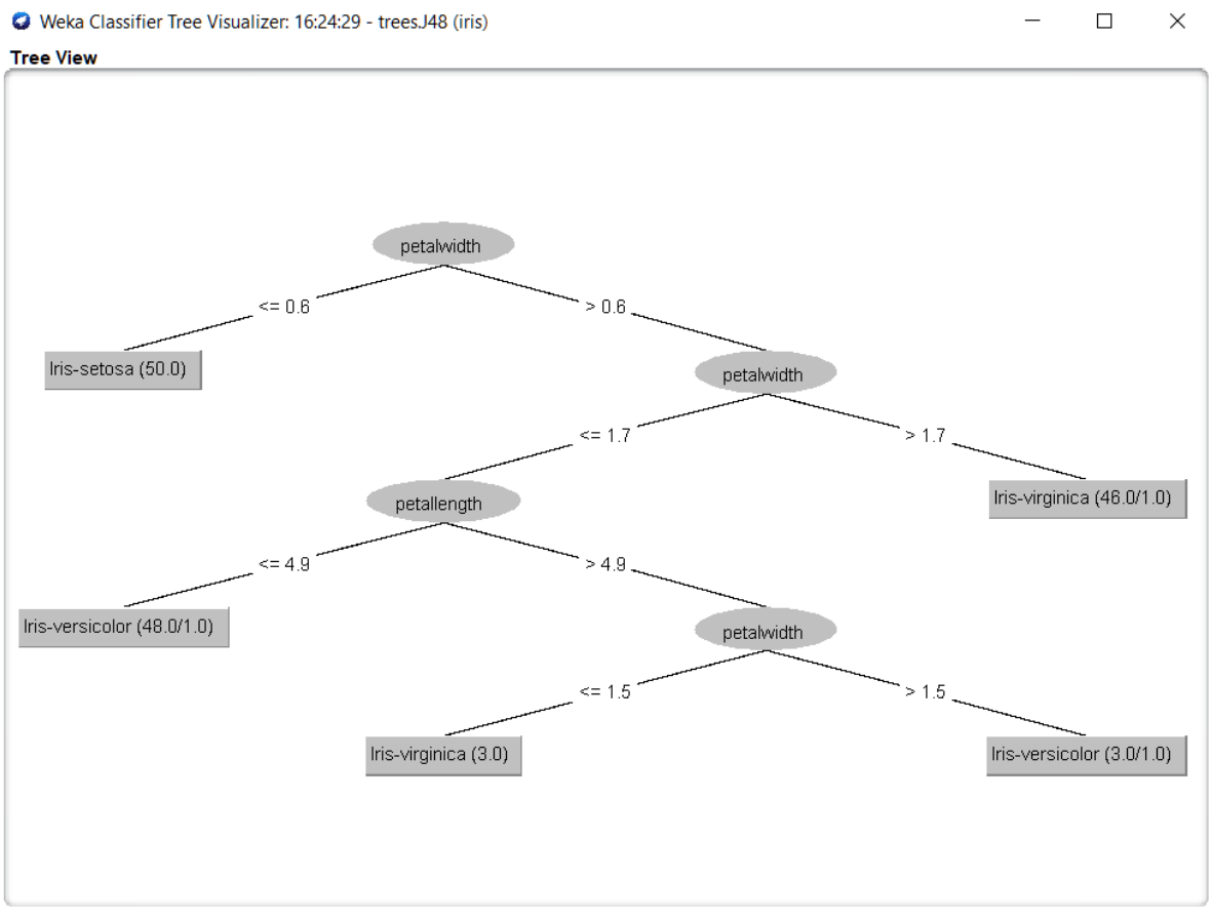
Zeynepsu Kesim

2292340

Report J48 pruned tree, Summary and Detailed Accuracy By Class.

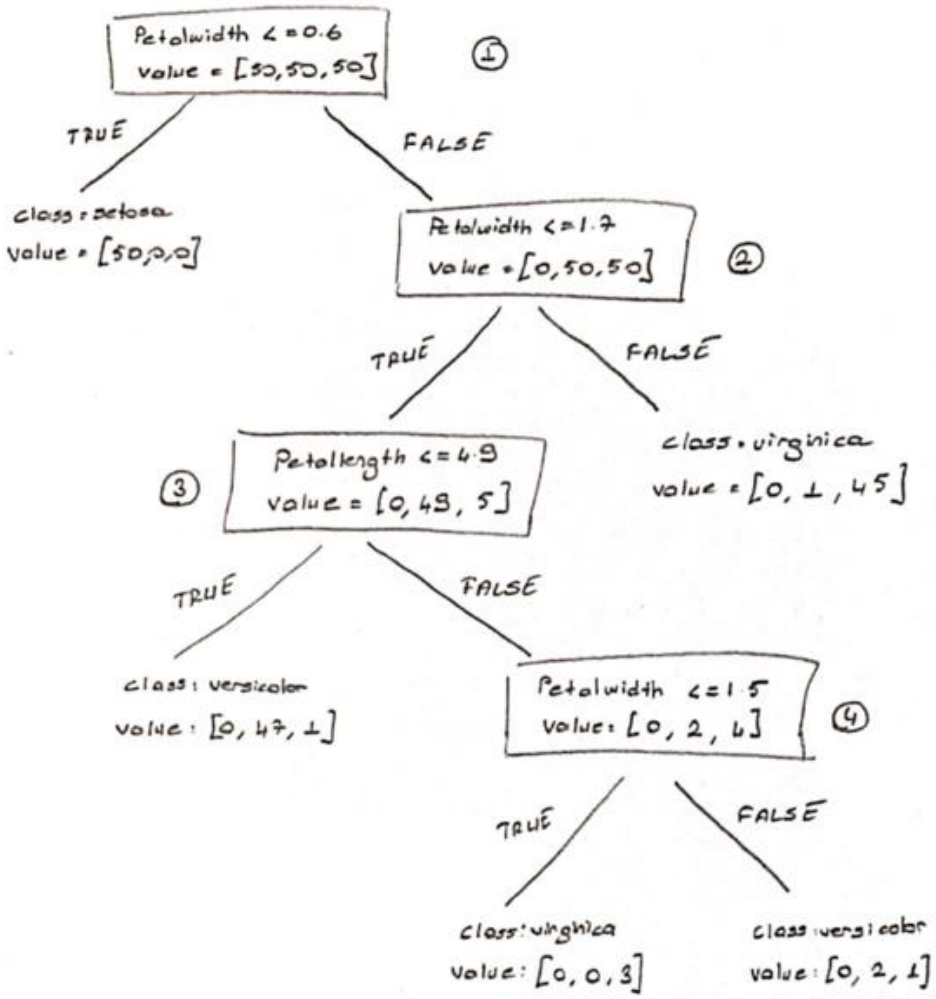
```
J48 pruned tree
------------------

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds


=== Summary ===

Correctly Classified Instances          43               95.5556 %
Kappa statistic                          0.9331
Mean absolute error                      0.0416
Root mean squared error                  0.1682
Relative absolute error                  9.3466 %
Root relative squared error             35.6559 %
Total Number of Instances               45

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     Iris-setosa
                 1,000    0,069    0,889      1,000    0,941      0,910    0,966     0,889     Iris-versicolor
                 0,867    0,000    1,000      0,867    0,929      0,901    0,964     0,931     Iris-virginica
Weighted Avg.    0,956    0,025    0,960      0,956    0,955      0,935    0,976     0,938
```

Visualization of the tree



1. **What is the meaning of information gain?**
   Shows the goodness of split when impurity measure is entropy. It means that how much entropy we have reduced by splitting. Difference between the first entropy before splitting and the final entropies after splitting. The bigger the difference the better information gain, meaning we have a better splitting.

2. **What is the entropy of each non-leaf node in the result tree? Show your calculations step by step, instead of writing final entropy values only.**



* Values are taken as value : [setosa, versicolor, virginica]

$$\text{Entropy(node1)} = E(1) = -\frac{50}{150}\log_2\frac{50}{150} - \frac{50}{150}\log_2\frac{50}{150} - \frac{50}{150}\log_2\frac{50}{150}$$

$$= -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{1}{3}\log_2\frac{1}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) = -1\log_2\frac{1}{3} = \boxed{1.585}$$

$$E(2) = -\frac{50}{100}\log_2\frac{50}{100} - \frac{50}{100}\log_2\frac{50}{100} = -\log\frac{1}{2} = \boxed{1}$$

$$E(3) = -\frac{49}{54}\log_2\frac{49}{54} - \frac{5}{54}\log_2\frac{5}{54} \simeq -0.907 \times -0.14 - 0,092 \times -3.433$$

$$\simeq 0,1269 + 0,3158 \simeq \boxed{0.4427}$$

$$E(4) = -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6} = \frac{1}{3} \times 1.585 + \frac{2}{3} \times 0.585 \simeq 0.526 + 0.39 = \boxed{0.918}$$

3. Trace the tree for predicting the correspondent classes for each given sample below

Sepal Length	Sepal Width	Petal Length	Petal Width	Class
7.2	3.2	5.2	1.5	Iris-Virginica
3.1	2.7	9.0	0.5	Iris-Setosa
4.2	5.8	4.7	0.7	Iris-Versicolor
1.4	3.4	3.2	1.9	Iris-Virginica