Exercise 1.1 – (a) How many levels are there in the decision tree?

Answer: Using the rpart package, we trained the model using the whole iris dataset. Once, we plot the decision tree by utilizing the rpart.plot package, we notice that there are 3 levels in the decision tree.

- ➔ Level 1: Splits the total data based on Petal.length

  We get a pure node for Iris Setosa on the left side of the tree. On the right side of the tree, we get an impure node of equal distribution of Iris Versicolor and Iris Virginica.

- ➔ Level 2: Splits the right-side data of Level 1 based on Petal.width

  At level 2, we get two impure nodes. The left node with 91% and 9% split range of Iris Versicolor and Iris Virginica respectively. Similarly on the right node with 2% and 98% split range of Iris Versicolor and Iris Virginica respectively.

- ➔ Level 3: At level 3, we have two terminal impure nodes as described in level 2.

Exercise 1.1 – (b) What is the default class label associated with each vertex?

Answer: For each vertex, we have a default class label associated with it. Those are as follows:

1) Level 1, Vertex 1: Default class label is Setosa

   Node number 1: 150 observations, complexity param=0.5

   predicted class=setosa, expected loss=0.6666667, P(node) =1

2) Level 2, Vertex 1: Default class label is Setosa

   Node number 2: 50 observations

   predicted class=setosa, expected loss=0, P(node) =0.3333333

3) Level 2, Vertex 2: Default class label is Versicolor

   Node number 3: 100 observations, complexity param=0.44

   predicted class=versicolor, expected loss=0.5, P(node) =0.6666667

4) Level 3, Vertex 1: Default class label is Versicolor

   Node number 6: 54 observations

   predicted class=versicolor, expected loss=0.09259259, P(node) =0.36

5) Level 3, Vertex 2: Default class label is Virginica

   Node number 7: 46 observations

predicted class=virginica, expected loss=0.02173913, P(node) =0.3066667

Exercise 1.1 – (c) Starting from the root note, what is the name of the first attribute used for a decision, and what are the split points?

Answer: From the root node, we use attributes and split points for the classification process. Those are as follows:

- ➔ Level 1, split on attribute: Petal.length

  Split points:

  Petal.length < 2.5 left subtree: pure node of setosa [class label = setosa]

  Petal.length >= 2.5 right subtree: impure node of versicolor and virginica [class label = versicolor]

- ➔ Level 2, split on attribute: Petal.width

  Split points:

  Petal.width < 1.8 left subtree: impure node of versicolor and virginica [class label = versicolor]

  Petal.width >= 1.8 right subtree: impure node of versicolor and virginica [class label = virginica]

Exercise 1.1 – (d) Each vertex has three lines.

Exercise 1.1 – (d) – (i) At each vertex, what do the three numbers in the middle line signify?

Answer: When we use type=4 in rpart.plot, we get each vertex with 3 lines. The three numbers in the middle line depict the predicted probability of each class.

Exercise 1.1 – (d) – (ii) At each vertex, what does the last line signify?
Answer: When we use type=4 in rpart.plot, we get each vertex with 3 lines. The last line depicts the percentage of observations in the node with respect to total number of observations.