spark-task-6-decision-tree.R

SURABHI. S

2021-02-11

```
#data importing
data("iris")
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1
1 1 1 1 ...
```

```
#this will give idea about the data
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

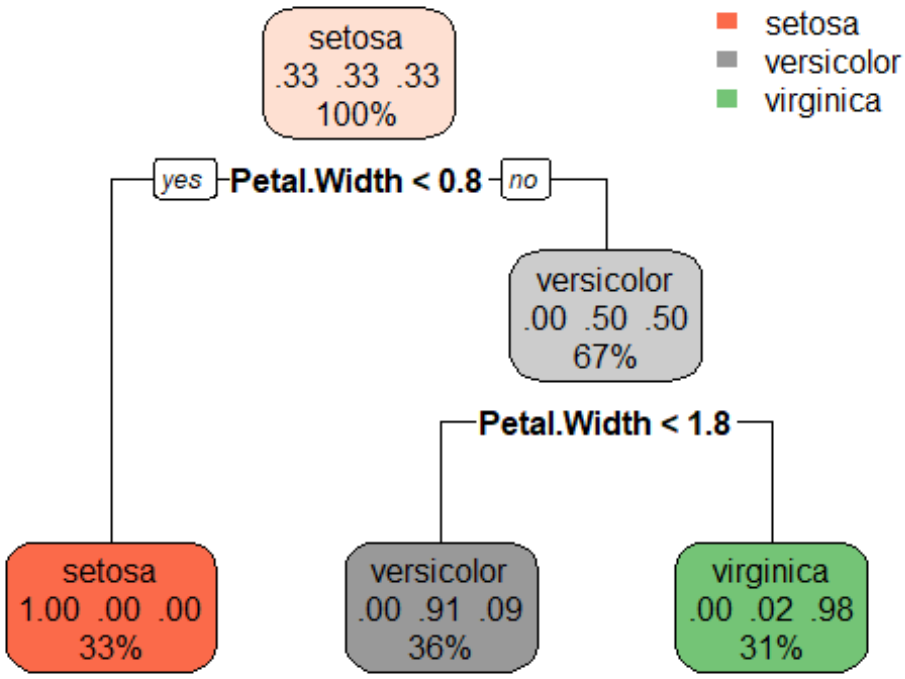
```
#decision tree by
library(rpart)
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 4.0.3
```

```
#decision tree
tree=rpart(Species~Sepal.Length+Sepal.Width+Petal.Width+Petal.Length,
           data=iris,method="class")
tree
```

```
## n= 150
##
```

```
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 150 100 setosa (0.33333333 0.33333333 0.33333333)
##   2) Petal.Width< 0.8 50   0 setosa (1.00000000 0.00000000 0.00000000) *
##   3) Petal.Width>=0.8 100  50 versicolor (0.00000000 0.50000000
0.50000000)
##     6) Petal.Width< 1.75 54   5 versicolor (0.00000000 0.90740741
0.09259259) *
##     7) Petal.Width>=1.75 46   1 virginica (0.00000000 0.02173913
0.97826087) *

#new data
unseen=data.frame(Sepal.Length=c(5.3,7.2),
                  Sepal.Width=c(2.9,3.9),
                  Petal.Length=c(1.7,5.4),
                  Petal.Width=c(0.8,2.3))
unseen

##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1          5.3         2.9          1.7         0.8
## 2          7.2         3.9          5.4         2.3

#classification of new data
predict(tree,unseen,type="class")

##          1          2
## versicolor  virginica
## Levels: setosa versicolor virginica

#Graphical representation
rpart.plot(tree)
```


setosa
versicolor
virginica
setosa
.33 .33 .33
100%
yes
Petal.Width < 0.8
no
versicolor
.00 .50 .50
67%
Petal.Width < 1.8
setosa
1.00 .00 .00
33%
versicolor
.00 .91 .09
36%
virginica
.00 .02 .98
31%