CSE3506 Essentials of Data Analytics

Name : **Sparsh Raj**
Reg. No. : **19BPS1028**

Lab Exercise 10: Random Forest

Objective: To perform Random Forest Classifier on any dataset.

Question:

```
rm(list=ls())
library('stats19')
## Warning: package 'stats19' was built under R version 4.1.3
## Data provided under OGL v3.0. Cite the source and link to:
## www.nationalarchives.gov.uk/doc/open-government-licence/version/3/
library('dplyr')
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
library('randomForest')
## Warning: package 'randomForest' was built under R version 4.1.3
## randomForest 4.7-1
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
## The following object is masked from 'package:dplyr':
##
##     combine
```

Read the dataset.We have taken the iris dataset.

```
mydata=iris
head(mydata)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
summary(mydata)
```

```
##    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
##
##
##
```

Split into Training and Testing data

```
index=sample(2,nrow(mydata), replace=TRUE,prob=c(0.7,0.3))
training=mydata[index==1,]
testing=mydata[index==2,]
```

Random Forest Implementation

```
RFM <- randomForest(Species ~ .,data=training, importance=T, proximity=T)
Species_Pred=predict(RFM,testing)
testing$Species_Pred=Species_Pred
```

```
head(testing)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species Species_Pre
d
## 4           4.6         3.1          1.5         0.2  setosa       setos
a
## 8           5.0         3.4          1.5         0.2  setosa       setos
a
## 12          4.8         3.4          1.6         0.2  setosa       setos
a
## 13          4.8         3.0          1.4         0.1  setosa       setos
a
## 14          4.3         3.0          1.1         0.1  setosa       setos
a
## 21          5.4         3.4          1.7         0.2  setosa       setos
a
```

```
CFM=table(testing$Species,testing$Species_Pred)
CFM
```

```
##
##              setosa versicolor virginica
##   setosa         17          0         0
##   versicolor      0         15         1
##   virginica       0          2        15
```

Car Dataset

```
data1 <- read.csv("cars.csv", header = TRUE)
head(data1)
```

```
##   car_ID                  CarName doornumber wheelbase carlength carwidt
h
## 1      1       alfa-romero giulia        two      88.6     168.8      64.
1
## 2      2      alfa-romero stelvio        two      88.6     168.8      64.
1
## 3      3 alfa-romero Quadrifoglio        two      94.5     171.2      65.
5
## 4      4              audi 100 ls       four      99.8     176.6      66.
2
## 5      5               audi 100ls       four      99.4     176.6      66.
4
## 6      6                 audi fox        two      99.8     177.3      66.
3
```

```
##   carheight curbweight cylindernumber enginesize stroke horsepower citym
pg
## 1      48.8       2548           four        130   2.68        111
21
## 2      48.8       2548           four        130   2.68        111
21
## 3      52.4       2823            six        152   3.47        154
19
## 4      54.3       2337           four        109   3.40        102
24
## 5      54.3       2824           five        136   3.40        115
18
## 6      53.1       2507           five        136   3.40        110
19
##   highwaympg price fueltype
## 1         27 13495      gas
## 2         27 16500      gas
## 3         26 16500      gas
## 4         30 13950      gas
## 5         22 17450      gas
## 6         25 15250      gas
```

```
str(data1)
```

```
## 'data.frame':    205 obs. of  16 variables:
##  $ car_ID        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ CarName       : chr  "alfa-romero giulia" "alfa-romero stelvio" "alfa
-romero Quadrifoglio" "audi 100 ls" ...
##  $ doornumber    : chr  "two" "two" "two" "four" ...
##  $ wheelbase     : num  88.6 88.6 94.5 99.8 99.4 ...
##  $ carlength     : num  169 169 171 177 177 ...
##  $ carwidth      : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.
9 ...
##  $ carheight     : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52
...
##  $ curbweight    : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 305
3 ...
##  $ cylindernumber: chr  "four" "four" "six" "four" ...
##  $ enginesize    : int  130 130 152 109 136 136 136 136 131 131 ...
##  $ stroke        : num  2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 3.4 ...
##  $ horsepower    : int  111 111 154 102 115 110 110 110 140 160 ...
##  $ citympg       : int  21 21 19 24 18 19 19 19 17 16 ...
```

```
##  $ highwaympg    : int  27 27 26 30 22 25 25 25 20 22 ...
##  $ price         : num  13495 16500 16500 13950 17450 ...
##  $ fueltype      : chr  "gas" "gas" "gas" "gas" ...
```

```
summary(data1)
```

```
##      car_ID      CarName           doornumber          wheelbase
##  Min.   :  1   Length:205         Length:205         Min.   : 86.60
##  1st Qu.: 52   Class :character   Class :character   1st Qu.: 94.50
##  Median :103   Mode  :character   Mode  :character   Median : 97.00
##  Mean   :103                                         Mean   : 98.76
##  3rd Qu.:154                                         3rd Qu.:102.40
##  Max.   :205                                         Max.   :120.90
##    carlength        carwidth       carheight       curbweight
##  Min.   :141.1   Min.   :60.30   Min.   :47.80   Min.   :1488
##  1st Qu.:166.3   1st Qu.:64.10   1st Qu.:52.00   1st Qu.:2145
##  Median :173.2   Median :65.50   Median :54.10   Median :2414
##  Mean   :174.0   Mean   :65.91   Mean   :53.72   Mean   :2556
##  3rd Qu.:183.1   3rd Qu.:66.90   3rd Qu.:55.50   3rd Qu.:2935
##  Max.   :208.1   Max.   :72.30   Max.   :59.80   Max.   :4066
##  cylindernumber       enginesize        stroke        horsepower
##  Length:205         Min.   : 61.0   Min.   :2.070   Min.   : 48.0
##  Class :character   1st Qu.: 97.0   1st Qu.:3.110   1st Qu.: 70.0
##  Mode  :character   Median :120.0   Median :3.290   Median : 95.0
##                     Mean   :126.9   Mean   :3.255   Mean   :104.1
##                     3rd Qu.:141.0   3rd Qu.:3.410   3rd Qu.:116.0
##                     Max.   :326.0   Max.   :4.170   Max.   :288.0
##     citympg        highwaympg        price          fueltype
##  Min.   :13.00   Min.   :16.00   Min.   : 5118   Length:205
##  1st Qu.:19.00   1st Qu.:25.00   1st Qu.: 7788   Class :character
##  Median :24.00   Median :30.00   Median :10295   Mode  :character
##  Mean   :25.22   Mean   :30.75   Mean   :13277
##  3rd Qu.:30.00   3rd Qu.:34.00   3rd Qu.:16503
##  Max.   :49.00   Max.   :54.00   Max.   :45400
```

```
data1$fueltype<-as.factor(data1$fueltype)  # convert numeric to factor
str(data1)
```

```
## 'data.frame':    205 obs. of  16 variables:
```

```
##  $ car_ID        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ CarName       : chr  "alfa-romero giulia" "alfa-romero stelvio" "alfa
-romero Quadrifoglio" "audi 100 ls" ...
##  $ doornumber    : chr  "two" "two" "two" "four" ...
##  $ wheelbase     : num  88.6 88.6 94.5 99.8 99.4 ...
##  $ carlength     : num  169 169 171 177 177 ...
##  $ carwidth      : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.
9 ...
##  $ carheight     : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52
...
##  $ curbweight    : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 305
3 ...
##  $ cylindernumber: chr  "four" "four" "six" "four" ...
##  $ enginesize    : int  130 130 152 109 136 136 136 136 131 131 ...
##  $ stroke        : num  2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 3.4 ...
##  $ horsepower    : int  111 111 154 102 115 110 110 110 140 160 ...
##  $ citympg       : int  21 21 19 24 18 19 19 19 17 16 ...
##  $ highwaympg    : int  27 27 26 30 22 25 25 25 20 22 ...
##  $ price         : num  13495 16500 16500 13950 17450 ...
##  $ fueltype      : Factor w/ 2 levels "diesel","gas": 2 2 2 2 2 2 2 2 2
2 ...
```

Split into Training and Testing data

```
index_1=sample(2,nrow(data1), replace=TRUE,prob=c(0.7,0.3))
train=data1[index_1==1,]
test=data1[index_1==2,]
```

Random Forest Implementation

```
RFM1<- randomForest(fueltype~ .,data=train, importance=T, proximity=T)
fueltype_Pred1=predict(RFM1,test)
test$fueltype_Pred1=fueltype_Pred1
head(test)
```

```
##    car_ID               CarName doornumber wheelbase carlength carwidth
## 4       4          audi 100 ls       four      99.8     176.6     66.2
## 7       7           audi 100ls       four     105.8     192.7     71.4
## 18     18               bmw x3       four     110.0     197.0     70.9
## 19     19     chevrolet impala        two      88.4     141.1     60.3
```

```
## 20     20 chevrolet monte carlo        two      94.5     155.9      63.6
## 22     22          dodge rampage       two      93.7     157.3      63.8
##    carheight curbweight cylindernumber enginesize stroke horsepower city
mpg
## 4       54.3       2337           four        109   3.40        102
24
## 7       55.7       2844           five        136   3.40        110
19
## 18      56.3       3505            six        209   3.39        182
15
## 19      53.2       1488          three         61   3.03         48
47
## 20      52.0       1874           four         90   3.11         70
38
## 22      50.8       1876           four         90   3.23         68
37
##    highwaympg price fueltype fueltype_Pred1
## 4          30 13950      gas            gas
## 7          25 17710      gas            gas
## 18         20 36880      gas            gas
## 19         53  5151      gas            gas
## 20         43  6295      gas            gas
## 22         41  5572      gas            gas
```

```
CFM_1=table(test$fueltype,test$fueltype_Pred1)
CFM_1
```

```
##
##          diesel gas
##   diesel      5   2
##   gas         0  58
```