```
R version 3.5.2 (2018-12-20) -- "Eggshell Igloo"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> rm(list=ls())
> setwd("C:/Users/Samsung/Desktop/project")
> #Load Libraries
> x = c("ggplot2", "corrgram", "DMwR", "caret", "randomForest", "unbalanced",
"C50", "dummies", "e1071", "Information",
+       "MASS", "rpart", "gbm", "ROSE", 'sampling', 'DataCombine', 'inTrees',
'ggplot2','readxl')
> #install.packages(x)
> lapply(x, require, character.only = TRUE)

[[1]]
[1] TRUE

[[2]]
[1] TRUE

[[3]]
[1] TRUE

[[4]]
[1] TRUE

[[5]]
[1] TRUE

[[6]]
[1] TRUE

[[7]]
[1] TRUE

[[8]]
[1] TRUE

[[9]]
[1] TRUE

[[10]]
[1] TRUE

[[11]]
[1] TRUE

[[12]]
[1] TRUE
```

```
[[13]]
[1] TRUE

[[14]]
[1] TRUE

[[15]]
[1] TRUE

[[16]]
[1] TRUE

[[17]]
[1] TRUE

[[18]]
[1] TRUE

[[19]]
[1] TRUE
```

```
> library(readxl)
> data <- read_excel("Absenteeism.xls")
> View(data)
> # See thew dimensions
> dim(data)
[1] 740  21
> #Strcucture of the data
> str(data)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':     740 obs. of  21 variables:
 $ ID                              : num  11 36 3 7 11 3 10 20 14 1 ...
 $ Reason for absence              : num  26 0 23 7 23 23 22 23 19 22 ...
 $ Month of absence                : num  7 7 7 7 7 7 7 7 7 7 ...
 $ Day of the week                 : num  3 3 4 5 5 6 6 6 2 2 ...
 $ Seasons                         : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Transportation expense          : num  289 118 179 279 289 179 NA 260 155 2
35 ...
 $ Distance from Residence to Work: num  36 13 51 5 36 51 52 50 12 11 ...
 $ Service time                    : num  13 18 18 14 13 18 3 11 14 14 ...
 $ Age                             : num  33 50 38 39 33 38 28 36 34 37 ...
 $ Work load Average/day           : num  239554 239554 239554 239554 239554 .
..
 $ Hit target                      : num  97 97 97 97 97 97 97 97 97 97 ...
 $ Disciplinary failure            : num  0 1 0 0 0 0 0 0 0 0 ...
 $ Education                       : num  1 1 1 1 1 1 1 1 1 3 ...
 $ Son                             : num  2 1 0 2 2 0 1 4 2 1 ...
 $ Social drinker                  : num  1 1 1 1 1 1 1 1 1 0 ...
 $ Social smoker                   : num  0 0 0 1 0 0 0 0 0 0 ...
 $ Pet                             : num  1 0 0 0 1 0 4 0 0 1 ...
 $ Weight                          : num  90 98 89 68 90 89 80 65 95 88 ...
 $ Height                          : num  172 178 170 168 172 170 172 168 196
172 ...
 $ Body mass index                 : num  30 31 31 24 30 31 27 23 25 29 ...
 $ Absenteeism time in hours       : num  4 0 2 4 2 NA 8 4 40 8 ...
> #head(data)
> ###################################Missing Values Analysis##################
##############################
> missing_val = data.frame(apply(data,2,function(x){sum(is.na(x))}))
> missing_val$Columns = row.names(missing_val)
> names(missing_val)[1] =  "Missing_percentage"
> missing_val$Missing_percentage = (missing_val$Missing_percentage/nrow(data)
) * 100
```

```
> missing_val = missing_val[order(-missing_val$Missing_percentage),]
> row.names(missing_val) = NULL
> missing_val = missing_val[,c(2,1)]
> missing_val
                           Columns Missing_percentage
1                    Body mass index          4.1891892
2          Absenteeism time in hours          2.9729730
3                             Height          1.8918919
4              Work load Average/day          1.3513514
5                          Education          1.3513514
6             Transportation expense          0.9459459
7                         Hit target          0.8108108
8               Disciplinary failure          0.8108108
9                                Son          0.8108108
10                     Social smoker          0.5405405
11                Reason for absence          0.4054054
12 Distance from Residence to Work          0.4054054
13                       Service time          0.4054054
14                                Age          0.4054054
15                     Social drinker          0.4054054
16                                Pet          0.2702703
17                  Month of absence          0.1351351
18                             Weight          0.1351351
19                                 ID          0.0000000
20                   Day of the week          0.0000000
21                           Seasons          0.0000000
> write.csv(missing_val, "Miisingfile.csv", row.names = F)
> # Rename  column names
> names(data)[1]<-"ID"
> names(data)[2]<-"Reasonforabsence"
> names(data)[3]<-"Monthofabsence"
> names(data)[4]<-"Dayofweek"
> names(data)[5]<-"Seasons"
> names(data)[6]<-"Transportationexpense"
> names(data)[7]<-"Distancefromresidence"
> names(data)[8]<-"Servicetime"
> names(data)[9]<-"Age"
> names(data)[10]<-"WorkloadAverage"
> names(data)[11]<-"Hittarget"
> names(data)[12]<-"Disciplinaryfailure"
> names(data)[13]<-"Education"
> names(data)[14]<-"Son"
> names(data)[15]<-"Socialdrinker"
> names(data)[16]<-"Socialsmoker"
> names(data)[17]<-"Pet"
> names(data)[18]<-"Weight"
> names(data)[19]<-"Height"
> names(data)[20]<-"Bodymassindex"
> names(data)[21]<-"Absenteesmtimeinhours"
> colnames(data)
 [1] "ID"                    "Reasonforabsence"      "Monthofabsence"
"Dayofweek"             "Seasons"
 [6] "Transportationexpense" "Distancefromresidence" "Servicetime"
"Age"                   "WorkloadAverage"
[11] "Hittarget"                 "Disciplinaryfailure"   "Education"
"Son"                   "Socialdrinker"
[16] "Socialsmoker"              "Pet"                   "Weight"
"Height"                "Bodymassindex"
[21] "Absenteesmtimeinhours"
> #KNN Imputation
> library("DMwR")
> data <- as.data.frame(data)
> data = knnImputation(data, k = 3)
```

```
> sum(is.na(data))
[1] 0
> colnames(data)
 [1] "ID"                    "Reasonforabsence"      "Monthofabsence"
"Dayofweek"             "Seasons"
 [6] "Transportationexpense" "Distancefromresidence" "Servicetime"
"Age"                   "WorkloadAverage"
[11] "Hittarget"             "Disciplinaryfailure"   "Education"
"Son"                   "Socialdrinker"
[16] "Socialsmoker"          "Pet"                   "Weight"
"Height"                "Bodymassindex"
[21] "Absenteesmtimeinhours"
> View(data)
> ##############################################Outlier Analysis###############
##############################
> # ## BoxPlots - Distribution and Outlier Check
> continuous_vars=c("Transportationexpense",
+                   "Distancefromresidence","Servicetime","Age",
+                   "WorkloadAverage","Hittarget","Weight",
+                   "Height","Bodymassindex","Absenteesmtimeinhours")
> continuous_vars
 [1] "Transportationexpense" "Distancefromresidence" "Servicetime"
"Age"                   "WorkloadAverage"
 [6] "Hittarget"             "Weight"                "Height"
"Bodymassindex"         "Absenteesmtimeinhours"
> Categorical_vars=c('ID','Reasonforabsence','Monthofabsence','Dayofweek','Se
asons',
+                    'Disciplinaryfailure','Socialsmoker','Socialdrinker','So
n','Pet','Education')
> Categorical_vars
 [1] "ID"                   "Reasonforabsence"    "Monthofabsence"      "Dayof
week"           "Seasons"
 [6] "Disciplinaryfailure" "Socialsmoker"        "Socialdrinker"       "Son"
"Pet"
[11] "Education"
> colnames(data)
 [1] "ID"                    "Reasonforabsence"      "Monthofabsence"
"Dayofweek"             "Seasons"
 [6] "Transportationexpense" "Distancefromresidence" "Servicetime"
"Age"                   "WorkloadAverage"
[11] "Hittarget"             "Disciplinaryfailure"   "Education"
"Son"                   "Socialdrinker"
[16] "Socialsmoker"          "Pet"                   "Weight"
"Height"                "Bodymassindex"
[21] "Absenteesmtimeinhours"
> library(ggplot2)
> for (i in 1:length(continuous_vars))
+ {
+   assign(paste0("gn",i), ggplot(aes_string(y = (continuous_vars[i]), x = "A
bsenteesmtimeinhours"), data = subset(data))+
+            stat_boxplot(geom = "errorbar", width = 0.5) +
+            geom_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=
18,
+                         outlier.size=1, notch=FALSE) +
+            theme(legend.position="bottom")+
+            labs(y=continuous_vars[i],x="Absenteesmtimeinhours")+
+            ggtitle(paste("Box plot of Absenteeism for",continuous_vars[i]))
)
+ }
> # ## Plotting plots together
> gridExtra::grid.arrange(gn1,gn5,gn2,ncol=3)
> gridExtra::grid.arrange(gn6,gn7,gn4,gn3,ncol=4)
> gridExtra::grid.arrange(gn8,gn9,gn10,ncol=3)
```

```
> # # #loop to remove from all variables
> for(i in continuous_vars){
+   print(i)
+   val = data[,i][data[,i] %in% boxplot.stats(data[,i])$out]
+   print(length(val))
+   data = data[which(!data[,i] %in% val),]
+   data[,i][data[,i] %in% val] = NA
+ }
[1] "Transportationexpense"
[1] 3
[1] "Distancefromresidence"
[1] 0
[1] "Servicetime"
[1] 5
[1] "Age"
[1] 8
[1] "WorkloadAverage"
[1] 29
[1] "Hittarget"
[1] 19
[1] "Weight"
[1] 0
[1] "Height"
[1] 106
[1] "Bodymassindex"
[1] 0
[1] "Absenteesmtimeinhours"
[1] 26
> #for(i in continuous_vars){
>  # val = data[,i][data[,i] %in% boxplot.stats(data[,i])$out]
>   #print(length(val))
>   #data[,i][data[,i] %in% val] = NA
> #}
> table(is.na(data))

FALSE
11424
> #Imputing missing values
> data=knnImputation(data,k=3)
> ###################################Feature Selection#########################
#######################
> ## Correlation Plot
> library(corrgram)
> corrgram(data[,continuous_vars], order = F,
+          upper.panel=panel.pie, text.panel=panel.txt, main = "Correlation P
lot")
> ## ANOVA test for Categprical variable
> summary(aov(formula = Absenteesmtimeinhours~ID,data = data))
             Df Sum Sq Mean Sq F value  Pr(>F)
ID            1     91   90.61   8.397 0.00391 **
Residuals   542   5849   10.79
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov(formula = Absenteesmtimeinhours~Reasonforabsence,data = data))
                  Df Sum Sq Mean Sq F value   Pr(>F)
Reasonforabsence   1    132  132.39   12.36 0.000476 ***
Residuals        542   5807   10.71
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov(formula = Absenteesmtimeinhours~Monthofabsence,data = data))
               Df Sum Sq Mean Sq F value Pr(>F)
Monthofabsence   1      3   3.268   0.298  0.585
Residuals      542   5936  10.952
> summary(aov(formula = Absenteesmtimeinhours~Dayofweek,data = data))
             Df Sum Sq Mean Sq F value Pr(>F)
Dayofweek     1     40   40.32   3.705 0.0548 .
Residuals   542   5899   10.88
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov(formula = Absenteesmtimeinhours~Seasons,data = data))
             Df Sum Sq Mean Sq F value Pr(>F)
Seasons       1      4   3.801   0.347  0.556
Residuals   542   5936  10.951
> summary(aov(formula = Absenteesmtimeinhours~Disciplinaryfailure,data = data
))
                     Df Sum Sq Mean Sq F value   Pr(>F)
Disciplinaryfailure   1    344   344.4   33.36 1.29e-08 ***
Residuals           542   5595    10.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov(formula = Absenteesmtimeinhours~Education,data = data))
             Df Sum Sq Mean Sq F value Pr(>F)
Education     1      1   1.007   0.092  0.762
Residuals   542   5938  10.956
> summary(aov(formula = Absenteesmtimeinhours~Socialdrinker,data = data))
              Df Sum Sq Mean Sq F value  Pr(>F)
Socialdrinker   1    105  104.90   9.745 0.00189 **
Residuals     542   5834   10.76
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov(formula = Absenteesmtimeinhours~Socialsmoker,data = data))
              Df Sum Sq Mean Sq F value Pr(>F)
Socialsmoker   1     34   34.04   3.124 0.0777 .
Residuals    542   5905   10.90
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov(formula = Absenteesmtimeinhours~Son,data = data))
             Df Sum Sq Mean Sq F value   Pr(>F)
Son           1    240  240.41   22.86 2.24e-06 ***
Residuals   542   5699   10.51
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov(formula = Absenteesmtimeinhours~Pet,data = data))
             Df Sum Sq Mean Sq F value Pr(>F)
Pet           1      5   5.226   0.477   0.49
Residuals   542   5934  10.949
> ## Dimension Reduction
> data = subset(data, select = -c(Weight))
> dim(data)
[1] 544  20
> #*******************************Feature Scalling*****************************
*
> #Update coninuou and categorical variables
> continuous_vars=c("Transportationexpense",
+                   "Distancefromresidence","Servicetime","Age",
+                   "WorkloadAverage","Hittarget",
+                   "Height","Bodymassindex")
> continuous_vars
[1] "Transportationexpense" "Distancefromresidence" "Servicetime"           "
Age"                    "WorkloadAverage"
```

```
[6] "Hittarget"             "Height"                "Bodymassindex"
> colnames(continuous_vars)
NULL
> Categorical_vars=c('ID','Reasonforabsence','Monthofabsence','Dayofweek','Se
asons',
+                    'Disciplinaryfailure','Socialsmoker','Socialdrinker','So
n','Pet','Education')
> Categorical_vars
 [1] "ID"                 "Reasonforabsence"    "Monthofabsence"      "Dayof
week"           "Seasons"
 [6] "Disciplinaryfailure" "Socialsmoker"        "Socialdrinker"       "Son"
"Pet"
[11] "Education"
> # #Standardisation
> for(i in continuous_vars){
+   print(i)
+   data[,i] = (data[,i] - mean(data[,i]))/sd(data[,i])
+ }
[1] "Transportationexpense"
[1] "Distancefromresidence"
[1] "Servicetime"
[1] "Age"
[1] "WorkloadAverage"
[1] "Hittarget"
[1] "Height"
[1] "Bodymassindex"
> View(data)
> #Divide data into train and test using stratified sampling method
> set.seed(123)
> train.index = sample(1:nrow(data), 0.8 * nrow(data))
> train = data[ train.index,]
> test  = data[-train.index,]
> #**********************Model Development Phase*****************************
****
> dim(train)
[1] 435  20
> dim(test)
[1] 109  20
> ##Decision tree for classification
> #Develop Model on training data
> fit_DT = rpart(Absenteesmtimeinhours ~., data = train, method = "anova")
> #Summary of DT model
> summary(fit_DT)
Call:
rpart(formula = Absenteesmtimeinhours ~ ., data = train, method = "anova")
  n= 435

           CP nsplit rel error    xerror       xstd
1  0.15284136      0 1.0000000 1.0030408 0.09035805
2  0.03175268      2 0.6943173 0.6994610 0.07524030
3  0.02802241      3 0.6625646 0.6983665 0.07934922
4  0.02090210      4 0.6345422 0.6933803 0.07807169
5  0.01875426      5 0.6136401 0.6825453 0.07876560
6  0.01575145      7 0.5761316 0.6874324 0.08127554
7  0.01293613      8 0.5603801 0.6983549 0.08354162
8  0.01265739      9 0.5474440 0.7129042 0.08388699
9  0.01028298     11 0.5221292 0.7071314 0.08409765
10 0.01000000     12 0.5118462 0.7097408 0.08266846

Variable importance
     Reasonforabsence   Disciplinaryfailure Transportationexpense
ID                Height
```

```
                  29                      15                      10
7                    5
        Bodymassindex                     Son         Monthofabsence
Pet            Servicetime
                   5                       4                       4
4                    4
                  Age Distancefromresidence                 Seasons
Hittarget           Socialsmoker
                   3                       3                       3
2                    2
            Education
                   1

Node number 1: 435 observations,    complexity param=0.1528414
  mean=4.309009, MSE=11.63178
  left son=2 (259 obs) right son=3 (176 obs)
  Primary splits:
      Reasonforabsence      < 22.5         to the right, improve=0.15244340,
(0 missing)
      Disciplinaryfailure   < 0.8301272    to the right, improve=0.06549184,
(0 missing)
      Transportationexpense < 0.09707757   to the left,  improve=0.05104341,
(0 missing)
      Son                   < 1.5          to the left,  improve=0.04309360,
(0 missing)
      ID                    < 27.5         to the right, improve=0.03737734,
(0 missing)
  Surrogate splits:
      Transportationexpense < 1.127482     to the left,  agree=0.646, adj=0.1
25, (0 split)
      Disciplinaryfailure   < 0.3301272    to the left,  agree=0.639, adj=0.1
08, (0 split)
      Height                < 0.7740529    to the left,  agree=0.639, adj=0.1
08, (0 split)
      Pet                   < 3            to the left,  agree=0.637, adj=0.1
02, (0 split)
      Socialsmoker          < 0.5          to the left,  agree=0.632, adj=0.0
91, (0 split)

Node number 2: 259 observations,    complexity param=0.0209021
  mean=3.211308, MSE=5.790235
  left son=4 (249 obs) right son=5 (10 obs)
  Primary splits:
      Bodymassindex         < 1.926834     to the left,  improve=0.07052281,
(0 missing)
      Reasonforabsence      < 26.5         to the right, improve=0.05709426,
(0 missing)
      Son                   < 1.5          to the left,  improve=0.04964950,
(0 missing)
      Distancefromresidence < -0.5787794   to the right, improve=0.04763964,
(0 missing)
      ID                    < 27.5         to the right, improve=0.04202330,
(0 missing)

Node number 3: 176 observations,    complexity param=0.1528414
  mean=5.924375, MSE=15.84556
  left son=6 (21 obs) right son=7 (155 obs)
  Primary splits:
      Reasonforabsence    < 0.5          to the left,  improve=0.27802610, (0
missing)
      Disciplinaryfailure < 0.3301272    to the right, improve=0.24631140, (0
missing)
```

```
      Monthofabsence      < 7.5          to the right, improve=0.02955676, (0
missing)
      Seasons             < 3.5          to the right, improve=0.02909522, (0
missing)
      WorkloadAverage     < 1.001764     to the left,  improve=0.02835051, (0
missing)
  Surrogate splits:
      Disciplinaryfailure < 0.3301272    to the right, agree=0.989, adj=0.905
, (0 split)
      Monthofabsence      < 0.5          to the left,  agree=0.892, adj=0.095
, (0 split)
      Bodymassindex       < 1.466141     to the right, agree=0.886, adj=0.048
, (0 split)

Node number 4: 249 observations,    complexity param=0.01875426
  mean=3.083248, MSE=5.347429
  left son=8 (198 obs) right son=9 (51 obs)
  Primary splits:
      Son                   < 1.5          to the left,  improve=0.07062439,
(0 missing)
      Reasonforabsence      < 26.5         to the right, improve=0.04115649,
(0 missing)
      Transportationexpense < 0.4746304    to the left,  improve=0.03907225,
(0 missing)
      ID                    < 27.5         to the right, improve=0.03280397,
(0 missing)
      Hittarget             < -1.09702     to the right, improve=0.03192362,
(0 missing)
  Surrogate splits:
      Height                < -0.7964976   to the right, agree=0.908, adj=0.5
49, (0 split)
      Transportationexpense < 0.3645108    to the left,  agree=0.896, adj=0.4
90, (0 split)
      Age                   < 1.226003     to the left,  agree=0.859, adj=0.3
14, (0 split)
      Bodymassindex         < 1.12062      to the left,  agree=0.851, adj=0.2
75, (0 split)
      Socialsmoker          < 0.5          to the left,  agree=0.819, adj=0.1
18, (0 split)

Node number 5: 10 observations
  mean=6.4, MSE=6.24

Node number 6: 21 observations
  mean=0.2220402, MSE=0.4825977

Node number 7: 155 observations,    complexity param=0.03175268
  mean=6.69695, MSE=12.92464
  left son=14 (52 obs) right son=15 (103 obs)
  Primary splits:
      Transportationexpense < -0.6265653   to the left,  improve=0.08019842,
(0 missing)
      ID                    < 25           to the right, improve=0.05357337,
(0 missing)
      Socialdrinker         < 0.5          to the left,  improve=0.05241698,
(0 missing)
      Reasonforabsence      < 18.5         to the left,  improve=0.05167860,
(0 missing)
      Height                < -0.7964976   to the right, improve=0.04730483,
(0 missing)
  Surrogate splits:
      Son           < 0.5          to the left,  agree=0.884, adj=0.654, (0 s
plit)
```

```
      Pet           < 0.5          to the left,  agree=0.845, adj=0.538, (0 s
plit)
      Servicetime   < 1.04741      to the right, agree=0.826, adj=0.481, (0 s
plit)
      Bodymassindex < -0.9525024   to the left,  agree=0.794, adj=0.385, (0 s
plit)
      ID            < 33.5         to the right, agree=0.755, adj=0.269, (0 s
plit)

Node number 8: 198 observations
  mean=2.771357, MSE=4.408508

Node number 9: 51 observations,    complexity param=0.01875426
  mean=4.294118, MSE=7.148789
  left son=18 (31 obs) right son=19 (20 obs)
  Primary splits:
      Height               < -0.7964976   to the left,  improve=0.2626237, (
0 missing)
      ID                   < 16.5         to the right, improve=0.2519027, (
0 missing)
      Age                  < -0.2015348   to the right, improve=0.2446273, (
0 missing)
      Transportationexpense < 0.9465714    to the left,  improve=0.2446273, (
0 missing)
      Bodymassindex        < 1.005447     to the right, improve=0.2289551, (
0 missing)
  Surrogate splits:
      ID                   < 18.5         to the right, agree=0.941, adj=0.8
5, (0 split)
      Transportationexpense < 0.9465714    to the left,  agree=0.922, adj=0.8
0, (0 split)
      Age                  < -0.2015348   to the right, agree=0.882, adj=0.7
0, (0 split)
      Distancefromresidence < -0.5787794   to the right, agree=0.725, adj=0.3
0, (0 split)
      Servicetime          < -0.2475697   to the left,  agree=0.686, adj=0.2
0, (0 split)

Node number 14: 52 observations
  mean=5.264072, MSE=12.97011

Node number 15: 103 observations,    complexity param=0.02802241
  mean=7.420344, MSE=11.34185
  left son=30 (24 obs) right son=31 (79 obs)
  Primary splits:
      ID               < 25           to the right, improve=0.12137240, (0 mi
ssing)
      Pet              < 0.5          to the right, improve=0.06457637, (0 mi
ssing)
      Reasonforabsence < 18.5         to the left,  improve=0.06247727, (0 mi
ssing)
      Socialdrinker    < 0.5          to the left,  improve=0.05971176, (0 mi
ssing)
      Bodymassindex    < -0.7221554   to the right, improve=0.05144540, (0 mi
ssing)
  Surrogate splits:
      Transportationexpense < 0.09707757   to the left,  agree=0.854, adj=0.3
75, (0 split)
      Age                  < 1.226003     to the right, agree=0.825, adj=0.2
50, (0 split)
      Height               < -2.10529     to the left,  agree=0.825, adj=0.2
50, (0 split)
```

```
      Bodymassindex         < 1.005447     to the right, agree=0.796, adj=0.1
25, (0 split)

Node number 18: 31 observations
  mean=3.193548, MSE=3.575442

Node number 19: 20 observations,    complexity param=0.01293613
  mean=6, MSE=7.9
  left son=38 (11 obs) right son=39 (9 obs)
  Primary splits:
      Seasons          < 2.5          to the left,  improve=0.41426930, (0 mi
ssing)
      Reasonforabsence < 25.5         to the left,  improve=0.18987340, (0 mi
ssing)
      Hittarget        < -0.457189    to the right, improve=0.10680380, (0 mi
ssing)
      WorkloadAverage  < 0.0002658436 to the right, improve=0.03196522, (0 mi
ssing)
      ID               < 12           to the left,  improve=0.02109705, (0 mi
ssing)
  Surrogate splits:
      Monthofabsence        < 8.5          to the left,  agree=0.85, adj=0.66
7, (0 split)
      ID                    < 12           to the left,  agree=0.75, adj=0.44
4, (0 split)
      Transportationexpense < 1.111751     to the left,  agree=0.75, adj=0.44
4, (0 split)
      Distancefromresidence < -0.8552114   to the right, agree=0.75, adj=0.44
4, (0 split)
      Hittarget             < -1.09702     to the right, agree=0.75, adj=0.44
4, (0 split)

Node number 30: 24 observations,    complexity param=0.01575145
  mean=5.291667, MSE=12.78993
  left son=60 (13 obs) right son=61 (11 obs)
  Primary splits:
      Monthofabsence   < 4.5          to the right, improve=0.2596430, (0 mis
sing)
      Dayofweek        < 4.5          to the right, improve=0.2053075, (0 mis
sing)
      Reasonforabsence < 11.5         to the right, improve=0.1906475, (0 mis
sing)
      Seasons          < 3.5          to the right, improve=0.1117509, (0 mis
sing)
      WorkloadAverage  < -0.6699834   to the left,  improve=0.0537133, (0 mis
sing)
  Surrogate splits:
      Hittarget             < -0.1372732   to the left,  agree=0.792, adj=0.5
45, (0 split)
      Seasons               < 3.5          to the right, agree=0.750, adj=0.4
55, (0 split)
      ID                    < 30.5         to the left,  agree=0.708, adj=0.3
64, (0 split)
      Distancefromresidence < -0.4405635   to the right, agree=0.708, adj=0.3
64, (0 split)
      Servicetime           < 0.2704223    to the left,  agree=0.708, adj=0.3
64, (0 split)

Node number 31: 79 observations,    complexity param=0.01265739
  mean=8.067031, MSE=9.107132
  left son=62 (11 obs) right son=63 (68 obs)
  Primary splits:
```

```
      Education             < 2           to the right, improve=0.07589201,
(0 missing)
      ID                    < 3           to the left,  improve=0.07589201,
(0 missing)
      Height                < -0.7964976  to the right, improve=0.06880102,
(0 missing)
      Age                   < 0.9404955   to the left,  improve=0.06127262,
(0 missing)
      Distancefromresidence < -0.9243194  to the left,  improve=0.05385641,
(0 missing)
  Surrogate splits:
      ID                    < 3           to the left,  agree=1.000, adj=1.0
00, (0 split)
      Distancefromresidence < -1.235305   to the left,  agree=0.962, adj=0.7
27, (0 split)
      Transportationexpense < 0.2622569   to the left,  agree=0.886, adj=0.1
82, (0 split)
      Servicetime           < 0.2704223   to the right, agree=0.886, adj=0.1
82, (0 split)
      Socialdrinker         < 0.5         to the left,  agree=0.873, adj=0.0
91, (0 split)

Node number 38: 11 observations
  mean=4.363636, MSE=8.413223

Node number 39: 9 observations
  mean=8, MSE=0

Node number 60: 13 observations
  mean=3.615385, MSE=6.390533

Node number 61: 11 observations
  mean=7.272727, MSE=13.10744

Node number 62: 11 observations
  mean=6, MSE=7.818182

Node number 63: 68 observations,    complexity param=0.01265739
  mean=8.401403, MSE=8.512675
  left son=126 (30 obs) right son=127 (38 obs)
  Primary splits:
      Transportationexpense < 1.04096     to the right, improve=0.12695050,
(0 missing)
      Servicetime           < 0.01142629  to the left,  improve=0.09161951,
(0 missing)
      Age                   < -0.5822115  to the left,  improve=0.08992323,
(0 missing)
      Height                < -0.7964976  to the right, improve=0.05237312,
(0 missing)
      WorkloadAverage       < -0.6997059  to the left,  improve=0.05210795,
(0 missing)
  Surrogate splits:
      Servicetime           < 0.01142629  to the left,  agree=0.897, adj=0.7
67, (0 split)
      Age                   < -0.5822115  to the left,  agree=0.868, adj=0.7
00, (0 split)
      Pet                   < 3           to the right, agree=0.853, adj=0.6
67, (0 split)
      Distancefromresidence < 1.32169     to the right, agree=0.794, adj=0.5
33, (0 split)
      Son                   < 1.5         to the left,  agree=0.750, adj=0.4
33, (0 split)
```

```
Node number 126: 30 observations
  mean=7.231415, MSE=2.395142

Node number 127: 38 observations,    complexity param=0.01028298
  mean=9.325078, MSE=11.40844
  left son=254 (24 obs) right son=255 (14 obs)
  Primary splits:
      Seasons         < 2.5          to the left,  improve=0.12001750, (0 mis
sing)
      Monthofabsence  < 2.5          to the left,  improve=0.08231889, (0 mis
sing)
      WorkloadAverage < -0.6997059   to the left,  improve=0.05308748, (0 mis
sing)
      Age             < 0.8453263    to the left,  improve=0.01852644, (0 mis
sing)
      Son             < 1.5          to the right, improve=0.01852644, (0 mis
sing)
  Surrogate splits:
      Monthofabsence        < 9.5          to the left,  agree=0.763, adj=0.3
57, (0 split)
      Hittarget             < -1.09702     to the right, agree=0.711, adj=0.2
14, (0 split)
      Distancefromresidence < 0.7688264    to the left,  agree=0.658, adj=0.0
71, (0 split)
      Servicetime           < -0.1180717   to the right, agree=0.658, adj=0.0
71, (0 split)
      Son                   < 3            to the left,  agree=0.658, adj=0.0
71, (0 split)

Node number 254: 24 observations
  mean=8.431374, MSE=7.324019

Node number 255: 14 observations
  mean=10.85714, MSE=14.69388

> #write rules into disk
> write(capture.output(summary(fit_DT)), "Rules.txt")
> #Lets predict for training data
> pred_DT_train = predict(fit_DT, train[,names(test) != "Absenteesmtimeinhour
s"])
> #Lets predict for training data
> pred_DT_test = predict(fit_DT,test[,names(test) != "Absenteesmtimeinhours"]
)
> # For training data
> print(postResample(pred = pred_DT_train, obs = train[,20]))
     RMSE  Rsquared       MAE
2.4400173 0.4881538 1.6734371
> # For testing data
> print(postResample(pred = pred_DT_test, obs = test[,20]))
     RMSE  Rsquared       MAE
2.5263757 0.2776217 1.7176231
> #**********************************************Linear Regression*********
*******************************#
> set.seed(123)
> #Develop Model on training data
> fit_LR = lm(Absenteesmtimeinhours ~ ., data = train)
> #Lets predict for training data
> pred_LR_train = predict(fit_LR, train[,names(test) != "Absenteesmtimeinhour
s"])
> #Lets predict for testing data
> pred_LR_test = predict(fit_LR,test[,names(test) != "Absenteesmtimeinhours"]
)
> # For training data
```

```
> print(postResample(pred = pred_LR_train, obs = train[,20]))
     RMSE  Rsquared       MAE
2.8709805 0.2913788 2.1069866
> # For testing data
> print(postResample(pred = pred_LR_test, obs = test[,20]))
     RMSE  Rsquared       MAE
2.6552019 0.1922374 2.0407065
> set.seed(123)
> #Develop Model on training data
> fit_RF = randomForest(Absenteesmtimeinhours~., data = train)
> # For testing data
> print(postResample(pred = pred_LR_test, obs = test[,20]))
     RMSE  Rsquared       MAE
2.6552019 0.1922374 2.0407065
> set.seed(123)
> #Develop Model on training data
> fit_RF = randomForest(Absenteesmtimeinhours~., data = train)
> #Lets predict for training data
> pred_RF_train = predict(fit_RF, train[,names(test) != "Absenteesmtimeinhour
s"])
> #Lets predict for testing data
> pred_RF_test = predict(fit_RF,test[,names(test) != "Absenteesmtimeinhours"]
)
> # For training data
> print(postResample(pred = pred_RF_train, obs = train[,20]))
     RMSE  Rsquared       MAE
1.5070909 0.8409664 1.0210046
> # For testing data
> print(postResample(pred = pred_RF_test, obs = test[,20]))
     RMSE  Rsquared       MAE
2.3013691 0.3534497 1.6552546
> set.seed(123)
> #Develop Model on training data
> fit_XGB = gbm(Absenteesmtimeinhours~., data = train, n.trees = 300, interac
tion.depth = 2)
Distribution not specified, assuming gaussian ...
> #Lets predict for training data
> pred_XGB_train = predict(fit_XGB, train[,names(test) != "Absenteesmtimeinho
urs"], n.trees = 300)
> #Lets predict for testing data
> pred_XGB_test = predict(fit_XGB,test[,names(test) != "Absenteesmtimeinhours
"], n.trees = 300)
> # For training data
> print(postResample(pred = pred_XGB_train, obs = train[,20]))
     RMSE  Rsquared       MAE
2.1142400 0.6261456 1.5013331
> # For testing data
> print(postResample(pred = pred_XGB_test, obs = test[,20]))
     RMSE  Rsquared       MAE
2.2789633 0.3656184 1.6322991
>
```