# assignment_10.1_MunjewarSheetal

Sheetal M

2023-02-18

## Install and Load required packages :

```
# Package names
# packages <- c("ggplot2","dplyr","tidyr","magrittr","tidyverse","purrr")
# Package Rweka - to call read.arff()

packages <- c("broom","dplyr","RWeka")

# Install packages not yet installed
installed_packages <- packages %in% rownames(installed.packages())
if (any(installed_packages == FALSE)) {
  install.packages(packages[!installed_packages])
}

# Packages loading
invisible(lapply(packages, library, character.only = TRUE))
```

```
##
## 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
```

## Problem statement : Predict one year life expectancy of lung cancer patients post surgery.

## Set the working directory to the root of your DSC 520 directory

setwd("E:\Data_Science_DSC510\DSC520-Statistics\dsc520")

```
## Set the working directory to the root of your DSC 520 directory
setwd("E:\\Data_Science_DSC510\\DSC520-Statistics\\dsc520\\data")
```

```
## Load data from "ThoraricSurgery.arff"
pat_data <- read.arff("ThoraricSurgery.arff")
str(pat_data)
```

```
## 'data.frame':    470 obs. of  17 variables:
##  $ DGN    : Factor w/ 7 levels "DGN3","DGN2",..: 2 1 1 1 1 1 1 2 1 1 ...
##  $ PRE4   : num  2.88 3.4 2.76 3.68 2.44 2.48 4.36 3.19 3.16 2.32 ...
##  $ PRE5   : num  2.16 1.88 2.08 3.04 0.96 1.88 3.28 2.5 2.64 2.16 ...
##  $ PRE6   : Factor w/ 3 levels "PRZ2","PRZ1",..: 2 3 2 3 1 2 2 2 1 2 ...
##  $ PRE7   : Factor w/ 2 levels "T","F": 2 2 2 2 2 2 2 2 2 2 ...
##  $ PRE8   : Factor w/ 2 levels "T","F": 2 2 2 2 1 2 2 2 2 2 ...
##  $ PRE9   : Factor w/ 2 levels "T","F": 2 2 2 2 2 2 2 2 2 2 ...
##  $ PRE10  : Factor w/ 2 levels "T","F": 1 2 1 2 1 1 1 1 1 1 ...
##  $ PRE11  : Factor w/ 2 levels "T","F": 1 2 2 2 1 2 2 2 1 2 ...
##  $ PRE14  : Factor w/ 4 levels "OC11","OC14",..: 2 3 1 1 1 1 3 1 1 1 ...
##  $ PRE17  : Factor w/ 2 levels "T","F": 2 2 2 2 2 2 1 2 2 2 ...
##  $ PRE19  : Factor w/ 2 levels "T","F": 2 2 2 2 2 2 2 2 2 2 ...
##  $ PRE25  : Factor w/ 2 levels "T","F": 2 2 2 2 2 2 2 1 2 2 ...
##  $ PRE30  : Factor w/ 2 levels "T","F": 1 1 1 2 1 2 1 1 1 1 ...
##  $ PRE32  : Factor w/ 2 levels "T","F": 2 2 2 2 2 2 2 2 2 2 ...
##  $ AGE    : num  60 51 59 54 73 51 59 66 68 54 ...
##  $ Risk1Yr: Factor w/ 2 levels "T","F": 2 2 2 2 1 2 1 1 2 2 ...
```

```
# nrow(pat_data)
# Alternate option for reference -
# install.packages("foreign")
# thoracic.df <- foreign::read.arff("data/ThoraricSurgery.arff")
```

### Generalized Linear Model

```
pat_mod01 <- glm(Risk1Yr ~ ., data = pat_data, family = "binomial")
```

### Model Summary

```
summary(pat_mod01)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ ., family = "binomial", data = pat_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4929   0.2762   0.4199   0.5439   1.6084
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.604e+01  2.333e+03   0.011 0.991093
## DGNDGN2     -5.557e-01  4.128e-01  -1.346 0.178199
## DGNDGN4     -4.278e-01  4.733e-01  -0.904 0.366122
## DGNDGN6      1.377e+01  1.178e+03   0.012 0.990671
## DGNDGN5     -2.201e+00  6.113e-01  -3.600 0.000318 ***
## DGNDGN8     -3.852e+00  1.550e+00  -2.485 0.012959 *
## DGNDGN1      1.418e+01  2.400e+03   0.006 0.995285
## PRE4         2.272e-01  1.849e-01   1.229 0.219094
## PRE5         3.030e-02  1.786e-02   1.697 0.089715 .
## PRE6PRZ1     1.490e-01  5.783e-01   0.258 0.796647
## PRE6PRZ0    -2.937e-01  7.907e-01  -0.371 0.710303
## PRE7F        7.153e-01  5.556e-01   1.288 0.197884
## PRE8F        1.743e-01  3.892e-01   0.448 0.654188
## PRE9F        1.368e+00  4.868e-01   2.811 0.004942 **
## PRE10F       5.770e-01  4.826e-01   1.196 0.231855
## PRE11F       5.162e-01  3.965e-01   1.302 0.192948
## PRE14OC14   -1.653e+00  6.094e-01  -2.713 0.006675 **
## PRE14OC12   -4.394e-01  3.301e-01  -1.331 0.183177
## PRE14OC13   -1.179e+00  6.165e-01  -1.913 0.055799 .
## PRE17F       9.266e-01  4.445e-01   2.085 0.037092 *
## PRE19F      -1.466e+01  1.654e+03  -0.009 0.992928
## PRE25F      -9.789e-02  1.003e+00  -0.098 0.922273
## PRE30F       1.084e+00  4.990e-01   2.172 0.029840 *
## PRE32F      -1.398e+01  1.645e+03  -0.008 0.993219
## AGE          9.506e-03  1.810e-02   0.525 0.599442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 395.61  on 469  degrees of freedom
## Residual deviance: 341.19  on 445  degrees of freedom
## AIC: 391.19
##
## Number of Fisher Scoring iterations: 15
```

### Variables with significance

1. DGNDGN5 Most Significant
2. PRE9F Significant
3. PRE14OC14 Significant

### Dataframe with new predicted column predict_Risk

```
#mod_plus <- augment(pat_mod01, type type.predict="response")
#class(mod_plus)
pat_mod01_predict <- augment(pat_mod01, type.predict="response") %>% mutate(predict_Risk = round(.fitte
# Name additional columns and check class.
# class(mod_plus)
# names(mod_plus)
```

### Confusion matrix to calculate accurracy

```
pat_mod01_predict %>% select(Risk1Yr, predict_Risk) %>% table()
```

```
##        predict_Risk
## Risk1Yr   0   1
##       T   3  67
##       F  10 390
```

### c. Accuracy of the Model

accuracy = correctly predicted / total Predicted * 100

```
accuracy <- (3 + 390) / (3 + 10 + 67 + 390)
accuracy <- accuracy * 100
print(paste(round(accuracy), "%"))
```

```
## [1] "84 %"
```