3c - Aniket Maheshwari

UB Person Number : 50360266

Setting up our environment and importing important libraries:

```
### Clear the environment
rm(list = ls())


### First we will set the directory of the R script
setwd("C:/Users/anike/Desktop/Sem 1/EAS 506 Statistical Data
Mining/Homework/Homework 3")


## Loading all the libraries
library(ISLR)
library(corrplot)

## corrplot 0.90 loaded

library(MASS)
library(klaR)
library(leaps)
library(lattice)
library(ggplot2)
library(corrplot)
library(car)

## Loading required package: carData

library(caret)
library(class)
library(MVN)
```

Importing dataset:

```
load("Diabetes.RData")
dim(Diabetes)

## [1] 145   6

str(Diabetes)

## 'data.frame':    145 obs. of  6 variables:
##  $ relwt  : num  0.81 0.95 0.94 1.04 1 0.76 0.91 1.1 0.99 0.78 ...
##  $ glufast: int  80 97 105 90 90 86 100 85 97 97 ...
##  $ glutest: int  356 289 319 356 323 381 350 301 379 296 ...
##  $ instest: int  124 117 143 199 240 157 221 186 142 131 ...
##  $ sspg   : int  55 76 105 108 143 165 119 105 98 94 ...
##  $ group  : Factor w/ 3 levels "Normal","Chemical_Diabetic",..: 1 1 1 1 1
1 1 1 1 1 ...

summary(Diabetes)

##      relwt          glufast       glutest          instest     
##  Min.   :0.7100   Min.   : 70   Min.   : 269.0   Min.   : 10.0  
##  1st Qu.:0.8800   1st Qu.: 90   1st Qu.: 352.0   1st Qu.:118.0  
##  Median :0.9800   Median : 97   Median : 413.0   Median :156.0  
##  Mean   :0.9773   Mean   :122   Mean   : 543.6   Mean   :186.1  
##  3rd Qu.:1.0800   3rd Qu.:112   3rd Qu.: 558.0   3rd Qu.:221.0  
##  Max.   :1.2000   Max.   :353   Max.   :1568.0   Max.   :748.0  
##       sspg                      group   
##  Min.   : 29.0   Normal           :76  
##  1st Qu.:100.0   Chemical_Diabetic:36  
##  Median :159.0   Overt_Diabetic   :33  
##  Mean   :184.2                         
##  3rd Qu.:257.0                         
##  Max.   :480.0
```

So , the diabetes dataset has 6 features. 5 of them are numeric and our target varaible is a categorical variable with 3 categories : ‘Normal’, ‘Chemical Diabetes’ and ‘Overt Diabetic’.

Let’s find out if there is any missing value in our dataset:

```
NAmat1 = matrix(as.numeric(is.na(Diabetes)) , ncol = 6)
nonNAdx1 = which(rowSums(NAmat1) == 0)
length(nonNAdx1)

## [1] 145

dim(Diabetes)

## [1] 145   6

## so no rows have empty or null values.
```

So the dataset has no null value.

Now, the value of the different feature are in different size of scale. For example, relwt is between 0.71 - 1.2 range whereas glutest is in between 269-1568 range. So i need to normalize this dataset so that all the features are in one scale before working on the dataset.

Normalize:

```
normalize <- function(x) {
  (x -min(x)) / (max(x) - min(x))

}
```

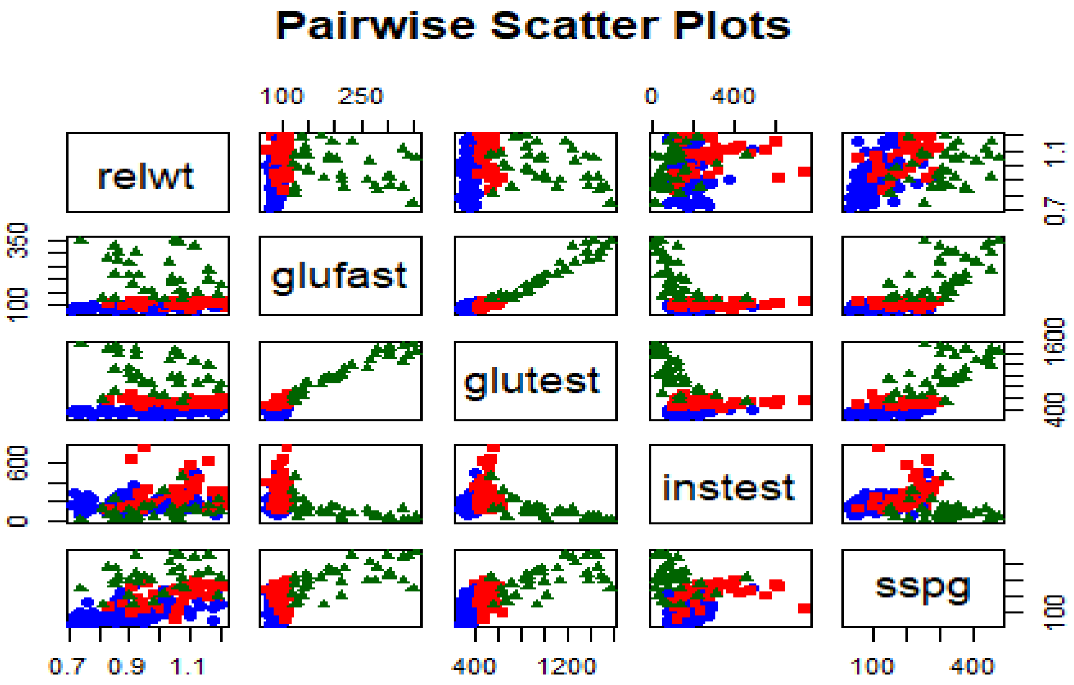
```
Diabetes_Norm <- as.data.frame(lapply(Diabetes[1:5], normalize))
full_dataset_Diabetes <- cbind(Diabetes_Norm , Diabetes)
full_dataset_Diabetes <- subset(full_dataset_Diabetes , select = -c(6:10))
head(full_dataset_Diabetes , 2)

##       relwt    glufast    glutest   instest       sspg  group
## 1 0.2040816 0.03533569 0.06697460 0.1544715 0.05764967 Normal
## 2 0.4897959 0.09540636 0.01539646 0.1449864 0.10421286 Normal
```

Now the full dataset is normalized.

Part A)

Plotting:

```
col <- c("blue", "red", "darkgreen")[full_dataset_Diabetes$group]
pch <- c(16,15,17)[full_dataset_Diabetes$group]
plot(Diabetes[,1:5], col=col, pch=pch , main = "Pairwise Scatter Plots")
```



Here, almost most of the variables are correlated and has elliptical shape. So none of the feature has multivariate normal. glufast and glutest has similar spread but it's not multivariate normal. glutest and sspg also has similar spread but it's not multivariate spread either. Just to be confirm, I'll do Multivariate Normality Test that's done by MVN package in R.

Multivariate Test:

```
MVN_test <- mvn(Diabetes[1:5], subset = NULL, mvnTest = "mardia")
MVN_test$multivariateNormality

##              Test        Statistic              p value Result
## 1 Mardia Skewness 402.412745893203 5.38979452469584e-64     NO
## 2 Mardia Kurtosis  11.647662518592                    0     NO
## 3             MVN             <NA>                 <NA>     NO
```

So it is confirmed that the classes is not Multivariate Normal.

Part B)

Splitting into test and train datasets: I'll split my data into train dataset (70% of the data) and test dataset (30% of the data).After splitting the train dataset had 104 row and 6 columns. The test dataset had 41 rows and 6 columns.

```
set.seed(1)
trainIndex <- createDataPartition(full_dataset_Diabetes$group, p = 0.70,list
= FALSE,times = 1)
train_data <- full_dataset_Diabetes[trainIndex,]
test_data <- full_dataset_Diabetes[-trainIndex,]
dim(train_data)

## [1] 104   6

dim(test_data)

## [1] 41  6

dim(full_dataset_Diabetes)

## [1] 145   6
```

LDA: Linear Discriminant analysis is a true decision boundary discovery algorithm. It assumes that the class has common covariance and it's decision boundary is linear separating the class.

```
lda.fit <- lda(group~., data = train_data)
lda.fit

## Call:
## lda(group ~ ., data = train_data)
##
## Prior probabilities of groups:
##            Normal Chemical_Diabetic    Overt_Diabetic
##         0.5192308         0.2500000         0.2307692
##
## Group means:
##                         relwt    glufast    glutest   instest      sspg
## Normal            0.4081633 0.07551368 0.05934765 0.2150959 0.1738934
## Chemical_Diabetic 0.7001570 0.10627888 0.16974596 0.3670002 0.3992836
## Overt_Diabetic    0.5663265 0.49234393 0.55783295 0.1484869 0.6119734
```

```
##
## Coefficients of linear discriminants:
##                 LD1         LD2
## relwt     0.9534108 -1.91989784
## glufast -12.0576474 10.82713615
## glutest  18.7366101 -8.30258477
## instest   0.1990146 -3.74896723
## sspg      1.6168734  0.00506727
##
## Proportion of trace:
##    LD1    LD2
## 0.8736 0.1264
```

So, Prior Probabilities of being in Normal category is 51%, being in Chemical Diabetic category is 25% and being in Overt Diabetic category is 23%.

Fitting the LDA model on test dataset:

```
test_pred <- predict(lda.fit , newdata = test_data)
test_pred_y = test_pred$class
table(test_data$group ,test_pred_y )

##                      test_pred_y
##                       Normal Chemical_Diabetic Overt_Diabetic
##   Normal                  19                 3              0
##   Chemical_Diabetic        0                10              0
##   Overt_Diabetic           0                 0              9

#Error
mean(test_data$group != test_pred_y)

## [1] 0.07317073
```

The Accuracy for LDA model is 73%.

QDA:

```
qda.fit <- qda(group~., data = train_data)
qda.fit

## Call:
## qda(group ~ ., data = train_data)
##
## Prior probabilities of groups:
##            Normal Chemical_Diabetic    Overt_Diabetic
##         0.5192308         0.2500000         0.2307692
##
## Group means:
##                       relwt    glufast    glutest   instest      sspg
## Normal            0.4081633 0.07551368 0.05934765 0.2150959 0.1738934
## Chemical_Diabetic 0.7001570 0.10627888 0.16974596 0.3670002 0.3992836
## Overt_Diabetic    0.5663265 0.49234393 0.55783295 0.1484869 0.6119734
```

So, Prior Probabilities of being in Normal category is 51%, being in Chemical Diabetic category is 25% and being in Overt Diabetic category is 23%.

Fitting the QDA model on test dataset:

```
qda_test_pred <- predict(qda.fit , newdata = test_data)
qda_test_pred_y = qda_test_pred$class
table(test_data$group ,qda_test_pred_y )
```

```
##                    qda_test_pred_y
##                     Normal Chemical_Diabetic Overt_Diabetic
##   Normal                20                 1              1
##   Chemical_Diabetic      0                 8              2
##   Overt_Diabetic         0                 0              9
```

```
mean(test_data$group != qda_test_pred_y)
```

```
## [1] 0.09756098
```

The Accuracy for QDA model is 97%

So, after fitting the model on both LDA and QDA, QDA turned out to have the better performance than LDA. This was expected as our data is not in multivariate normal form, so LDA won't work properly in this dataset.

Part C)

Given Dataset:

```
given_data = data.frame(relwt = c(1.86),glufast = c(184),glutest =
c(68),instest = c(122),sspg = c(544),group = c('x'))
y_lda_given_data = predict(lda.fit, newdata = given_data)
y_qda_given_data = predict(qda.fit, newdata = given_data)
```

```
y_lda_given_data
```

```
## $class
## [1] Overt_Diabetic
## Levels: Normal Chemical_Diabetic Overt_Diabetic
## 
## $posterior
##          Normal Chemical_Diabetic Overt_Diabetic
## 1 2.131158e-41                 0              1
## 
## $x
##          LD1      LD2
## 1 -41.58329 971.0468
```

```
y_qda_given_data
```

```
## $class
## [1] Overt_Diabetic
## Levels: Normal Chemical_Diabetic Overt_Diabetic
##
## $posterior
##   Normal Chemical_Diabetic Overt_Diabetic
## 1      0                 0              1
```

Both LDA and QDA predict that the data point will be of category ‘Overt Diabetes’.