Understanding Diabetes with Data Mining

2023-07-19

OVERVIEW OF THE PROJECT

Let’s consider the Diabetes dataset (posted with assignment). For this, we shall assume the population prior probabilities are estimated using the relative frequencies of the classes in the data.

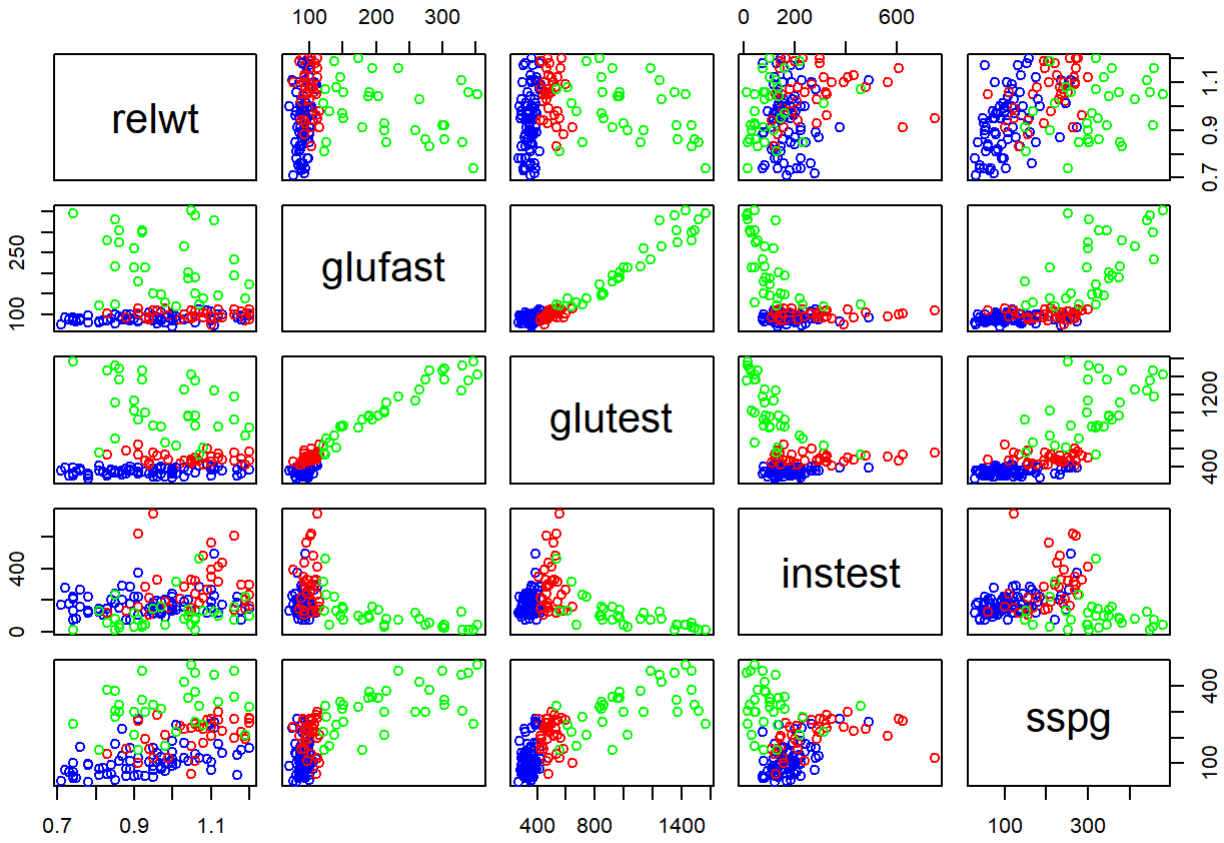
```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.3.3
```

```
setwd("C:/Users/devar/OneDrive/Documents")
load("Diabetes.RData")
colnames(Diabetes)
```

```
## [1] "relwt"   "glufast" "glutest" "instest" "sspg"    "group"
```

```
library(datasets)

cols <- Diabetes[, c("relwt", "glufast", "glutest", "instest", "sspg")]

colors <- c("blue", "red", "green")


Class_colours <- colors[Diabetes$group]
pairs(cols, col = Class_colours)
```



Here, we see that in the scatterplot between relwt vs. glufast, the class over_Diabetic has different covarriance with both Chemical_Diabetic and Normal as compared to the covariance between Chemical_Diabetic and Normal.

In the scatterplot between relwt vs instest, we see covariance for each class is different then in the the above mentioned plot.

In the scatterplot between relwt vs sspg we see a somewhat positive covariance between all the three classes, most prominently between Normal and Chemical_Diabetic as the classes moves together in the same direction.

In the scatterpolot between glufast vs. glutest, we see a linear plot of covariance and is positive covariance among all the three classes as they all move in the same direction.

In the scatterplot between glufast vs. instest, we see no positive covariance as all the classes are although tightly packed but they are not moving along the same direction.

In the scatterplot between glufast vs. sspg, we see no positive covariance as all the classes are although tightly packed but they are not moving along the same direction.

In the scatterplot between glutest vs. instest, we see no positive covariance as all the classes are although tightly packed but they are not moving along the same direction.

In the scatterplot between glufast vs. sspg, we see no positive covariance as all the classes are although tightly packed but they are not moving along the same direction.

In the scatterplot between istest vs. sspg, we see taht there is a positive covariance between Chemical_Diabetic and Normal.

Now, overall we see similar covariance matrices in the scatterplots of glufast vs. instest and glutest vs. instest. Also, the same is observerd in the scatterplots of glufast vs. sspg and glutest vs. sspg. Rest all the scatterplots shows different covariance matrices from each other.

Classes have different covariance matrices so We do not see any evidence of multivariatte normal as we do not see them forming any elliptical symmetric shape. This is becuase different classes have different covariance matrices. We also do see a lot of outliers which distor the normality assessment.

Now, I will apply linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) and then compare, how does the performance of QDA compare to that of LDA in this case.

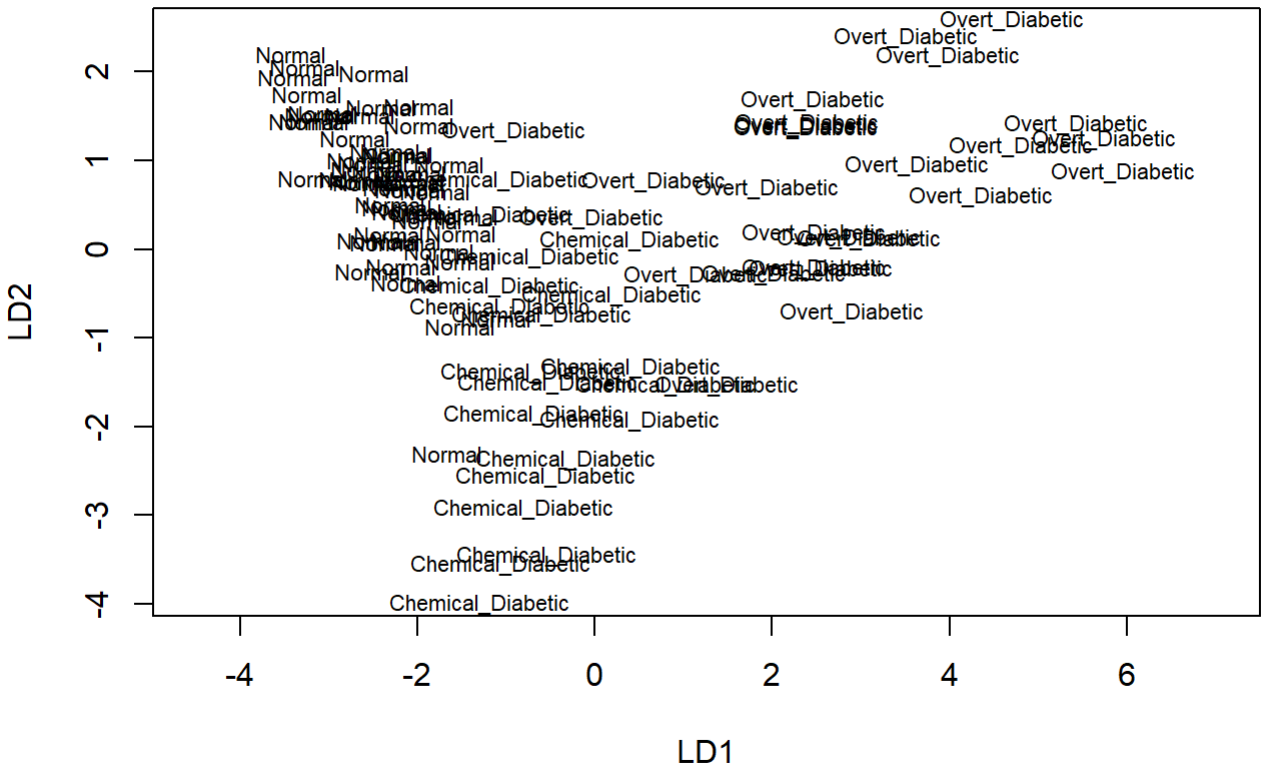
```
library(klaR)
```

```
## Warning: package 'klaR' was built under R version 4.3.3
```

```
library(MASS)
load("Diabetes.RData")

setwd("C:/Users/devar/OneDrive/Documents")
load("Diabetes.RData")

#Now to create test and training set

set.seed(123)
indis <- sample(1:nrow(Diabetes), round(2/3*nrow(Diabetes)), replace = FALSE)

diabetes_train <- Diabetes[indis, ]
diabetes_test <- Diabetes[-indis, ]

######################
#APPLYING LDA
######################
lda.fit <- lda(group~ ., data = diabetes_train)
lda.fit
```

```
## Call:
## lda(group ~ ., data = diabetes_train)
##
## Prior probabilities of groups:
##            Normal Chemical_Diabetic    Overt_Diabetic
##         0.5257732         0.2061856         0.2680412
##
## Group means:
##                        relwt   glufast   glutest   instest     sspg
## Normal             0.9333333  90.70588  349.9608 183.94118 122.9608
## Chemical_Diabetic 1.0770000 101.60000  507.1000 332.45000 220.3500
## Overt_Diabetic     0.9742308 220.34615 1055.1538  87.57692 315.7308
##
## Coefficients of linear discriminants:
##                   LD1           LD2
## relwt     1.740195087 -4.3104727464
## glufast  -0.029474300  0.0254809143
## glutest   0.011745960 -0.0055612909
## instest  -0.001458224 -0.0071136677
## sspg      0.002760410  0.0007069769
##
## Proportion of trace:
##    LD1    LD2
## 0.8573 0.1427
```

```
plot(lda.fit)
```



```
# make predictions for the test and training.
test_pred <- predict(lda.fit, newdata = diabetes_test)


train_pred <- predict(lda.fit, newdata = diabetes_train)


#getting the error rates
train_error <- (1/length(diabetes_train$group))*length(which(diabetes_train$group != train_pred$class))
test_error <- (1/length(diabetes_test$group))*length(which(diabetes_test$group != test_pred$class))
train_error
```

```
## [1] 0.09278351
```

```
test_error
```

```
## [1] 0.1458333
```

```
########################
#Applying QDA
########################

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

```
## Call:
## qda(group ~ ., data = diabetes_train)
##
## Prior probabilities of groups:
##            Normal Chemical_Diabetic    Overt_Diabetic
##         0.5257732         0.2061856         0.2680412
##
## Group means:
##                        relwt   glufast   glutest   instest     sspg
## Normal             0.9333333  90.70588  349.9608 183.94118 122.9608
## Chemical_Diabetic 1.0770000 101.60000  507.1000 332.45000 220.3500
## Overt_Diabetic    0.9742308 220.34615 1055.1538  87.57692 315.7308
```

```
train_pred <- predict(qda.fit, newdata = diabetes_train)
test_pred <- predict(qda.fit, newdata = diabetes_test)

ytrain <- train_pred$class
ytest <- test_pred$class

y_true_train <- diabetes_train$group
y_true_test <- diabetes_test$group

train_err <- (1/length(ytrain))*length(which(y_true_train != ytrain))
test_err <- (1/length(ytest))*length(which(y_true_test != ytest))

train_err
```

```
## [1] 0.03092784
```

```
test_err
```

```
## [1] 0.08333333
```

Here, we see that both LDA and QDA performs similarly upon comparing each of their training errors with their respective test errors. This is because the difference between the test and training error of both QDA and LDA are same, QDA being slightly better which can neglected.

Now, let's consider an individual has (glucose test/intolerence = 68, insulin test =122, SSPG = 544. Relative weight = 1.86, fasting plasma glucose = 184). We will find out to which class does LDA assign this individual and to which class does QDA.

```
#Readings of the individual
readings <- data.frame(
  glutest = 68,
  instest = 122,
  sspg = 544,
  relwt = 1.86,
  glufast = 184
)

# Predict the class using LDA
lda_pred <- predict(lda.fit, newdata = readings)$class

# Predict the class using QDA
qda_pred <- predict(qda.fit, newdata = readings)$class

# Print the predictions
print("LDA predicts the individual to be in the below class:")
```

```
## [1] "LDA predicts the individual to be in the below class:"
```

```
print(lda_pred)
```

```
## [1] Normal
## Levels: Normal Chemical_Diabetic Overt_Diabetic
```

```
print("QDA predicts the individual to be in the below class:")
```

```
## [1] "QDA predicts the individual to be in the below class:"
```

```
print(qda_pred)
```

```
## [1] Overt_Diabetic
## Levels: Normal Chemical_Diabetic Overt_Diabetic
```

Data Exploration and Visualization

I began by loading the Diabetes dataset and selecting specific variables for analysis, including relative weight, fasting plasma glucose, glucose test results, insulin test results, and SSPG (steady-state plasma glucose). To visualize the relationships between these variables and the diabetes classes, I created scatterplots using the pairs function, with different colors representing the classes: Normal, Chemical_Diabetic, and Overt_Diabetic.

Through these scatterplots:

I observed varying covariance patterns between different pairs of variables across the three classes. For example, the scatterplot between relwt (relative weight) and glufast (fasting plasma glucose) showed that the covariance for the Overt_Diabetic class differed from the other two classes. Similarly, other scatterplots revealed different covariance structures, with some showing positive covariance, especially between the Normal and Chemical_Diabetic classes, while others showed no positive covariance. Overall, the scatterplots did not exhibit the elliptical symmetric shapes typical of multivariate normal distributions, largely due to different covariance matrices across the classes and the presence of outliers. This suggested that the assumption of multivariate normality may not hold for this dataset.

Applying Linear Discriminant Analysis (LDA

Next, I applied Linear Discriminant Analysis (LDA) to classify the diabetes classes. First, I split the dataset into training and testing sets, with two-thirds of the data used for training and the remaining third for testing. I then fit an LDA model to the training data, which I visualized using a plot.

I predicted the classes for both the training and testing datasets and calculated the error rates to evaluate the model's performance. The training error and testing error were calculated as the proportion of misclassified observations in each set.

Applying Quadratic Discriminant Analysis (QDA)

Following LDA, I applied Quadratic Discriminant Analysis (QDA) using the same training and testing datasets. Similar to LDA, I made predictions for both datasets and calculated the respective error rates.

Comparison of LDA and QDA

When comparing the performance of LDA and QDA, I found that both models performed similarly, with their training and testing error rates being quite close. Although QDA showed a slightly better performance, the difference was minimal and could be considered negligible. This similarity in performance suggests that both LDA and QDA are viable methods for classifying the diabetes classes in this dataset, despite the different covariance structures observed in the exploratory analysis.

Class Prediction for a New Individual

Finally, I tested how LDA and QDA would classify a new individual with specific measurements:

Glucose test result: 68 Insulin test result: 122 SSPG: 544 Relative weight: 1.86 Fasting plasma glucose: 184 Using the fitted LDA and QDA models, I predicted the class for this individual:

LDA predicted that the individual belonged to the Normal class. QDA predicted that the individual belonged to the Overt_Diabetic class. The differing predictions highlight how LDA and QDA can sometimes yield different classifications, particularly in datasets where the classes have distinct covariance structures.