EAS 503-Homework 3

Boobalaganesh Ezhilan
UB Number - 50288429
October 10, 2018

Problem 1

In the given problem we need to predict whether the given suburb is greater than or smaller than the median of crime rate by applying classification algorithms like **Logistic Regression**, **Linear Discriminant Analysis** and **K-nearest neighbor**

Knowing the data

Using the Boston data set which is in the ISLR package with 506 observation and 15 variables we need to split the data into train data and testing data. Here in this problem, I had split the data into 60 percent of training data and 40 percent of test data. Now, my test data will be having 333 rows and test data will have 173 observations.

Creating Response Variable

Based upon the median value of crime rate we will find the response for the given data. Which is **1** in case of the value of crime data is greater than the median value and **0** when crime data value is lesser than its median value. Generally, it is categorical variable with values 1 and 0.

Performing Logistic Regression

We will try to apply logistic regression over the training dataset and try to find the training error and test error. Just because the response variable is not of factor variable we will convert into factor variable and perform logistic regression.

When we look into the summary of the fit we will be getting significance stars .under the estimate in the second row is the coefficient associated with the variable listed to the left. The Standard error in the next column associated with these estimates. Next column would be normalized error value (i.e. dividing estimates with the standard error). By looking at the significance stars we can infer that nitrogen oxides concentration, index of accessibility to radial highways and the median value of owner-occupied homes in $1000s contributes the lot to the output response. Now we will calculate the Mean Square error for the logistic regression for both training data and test data set which is around 0.14 for the test data set

TrainingError Logistic Error	TestError Logistic Error
0.0631	0.14

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -41.008408   8.921259  -4.597 4.29e-06 ***
zn           -0.110221   0.049054  -2.247  0.02464 *
indus        -0.074398   0.057481  -1.294  0.19556
chas          0.931221   1.052474   0.885  0.37627
nox          58.398322  11.085827   5.268 1.38e-07 ***
rm           -0.920518   0.922684  -0.998  0.31845
age           0.025529   0.015977   1.598  0.11007
dis           0.948961   0.317410   2.990  0.00279 **
rad           0.855951   0.213321   4.013 6.01e-05 ***
tax          -0.006560   0.003082  -2.129  0.03327 *
ptratio       0.303154   0.149477   2.028  0.04255 *
black        -0.009905   0.005527  -1.792  0.07309 .
lstat         0.119948   0.064518   1.859  0.06301 .
medv          0.246520   0.093456   2.638  0.00834 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Applying Linear Discriminant Analysis

We will now try to apply a linear discriminant analysis model to the training data and try to find the training error and test error. Since the response variable is not of factor variable we will convert into factor variable and perform the linear discriminant analysis.

TrainingError LDA	TestError LDA
0.1231	0.1618

To perform linear discriminant analysis on factor variable given continuous input variable we need to convert the output into response variable. We see that the test error for both the LDA model and the linear regression model is more or less same.

Applying K-nearest neighbor

We will try to apply the Knn model to the training data and try to find the training error and test error. I had iterated the values of K from 1 to 10. The minimum test error is for K equals 2 and 3.

TestError Knn
0.06358

First Modelling Predictions

Comparing the above three methods Knn performs better than the other two methods. Usually, Logistic Regression will perform better when there is linear data. But in this case, there may be a chance that data distributed is non-linear or it not in Gaussian shape. Comparing the other two models the logistics model performs better than the LDA method. Finally, I think Knn performs better because of its non-parametric property

Model Name	Test Error
Logistic	0.14
Linear Discriminant	0.1618
Knn	0.06358

Second Modelling Predictions

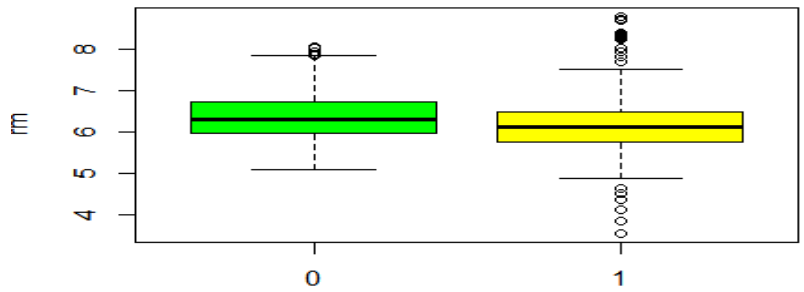
They had asked performs Three different models for different subsets of the data. Now I will remove the variables which the Logistic regression model find it as significant. (i.e) nitrogen oxides concentration, index of accessibility to radial highways and the median value of owner-occupied homes in $1000s. When I apply Logistic, Linear discriminant and Knn the test error is given by

Model Name	Test Error
Logistic	0.1791
Linear Discriminant	0.18497
Knn	0.06358

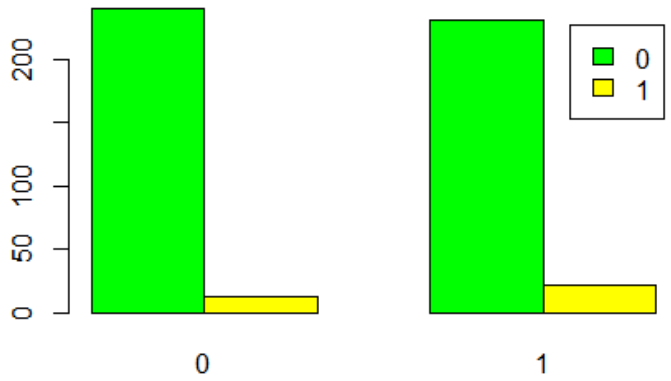
From the above table, we can see that the test error value of both Logistic and Linear discriminate value had started to increase. This is because of the reason that the removal of the data that contributes more to the crime rate of the suburbs. And Knn error remains the same or even decreased may be due to the fact that data get more non-linear so the knn performs better because of its Non-parametric nature.

Third Modelling Predictions

Now in the third modeling, I had removed the variable which I didn't contribute much to the prediction of the output response variable. So I removed some of the variables like an average number of rooms per dwelling, Charles River dummy variable and lower status of the population. I came to this conclusion based upon the box plot between the response variable and the other three variables and also from the summary of the significant values from the Linear Discriminant model.



The response between output Variable and rooms per dwelling



The response between output variable and Charles river variable

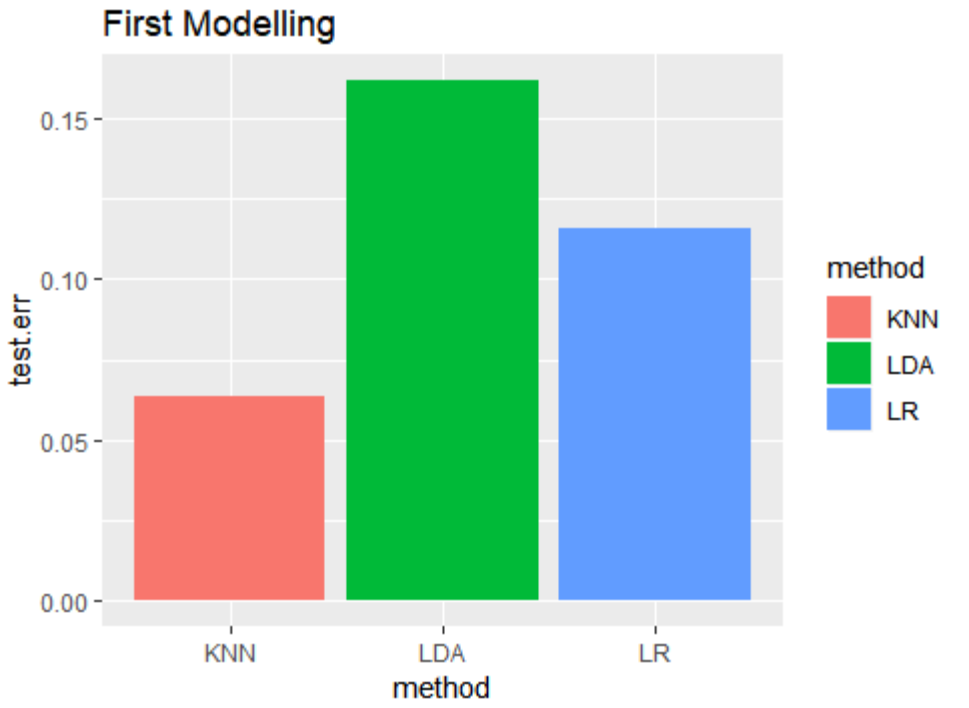
From the above boxplot and histogram, we can infer that both the variable is not contributing significant amount data for the output response.

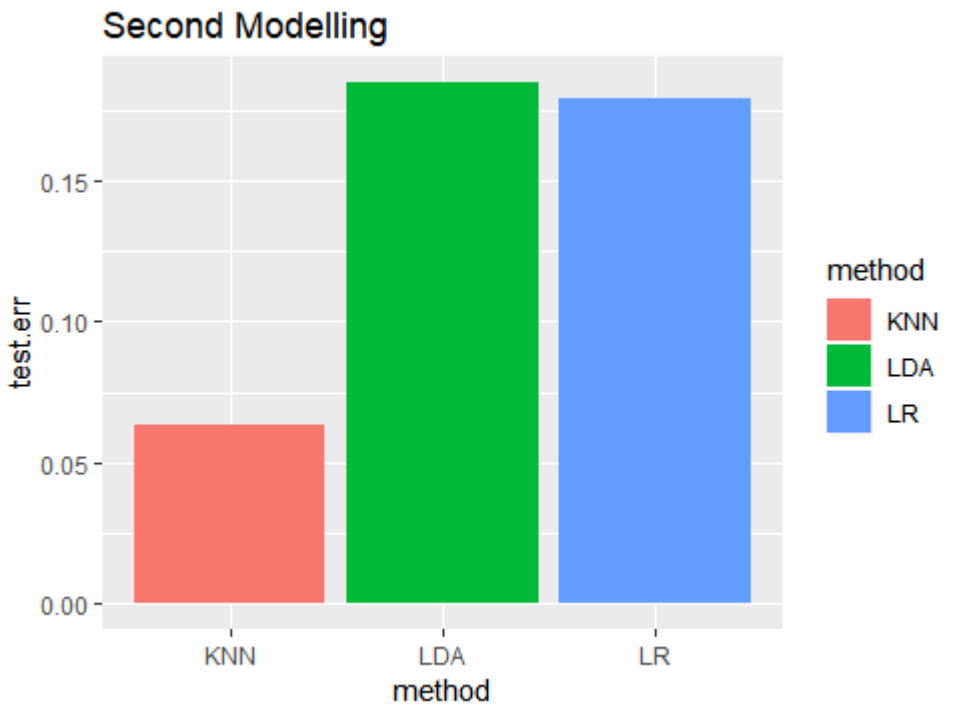
Model Name	Test Error
Logistic	**0.06963**
Linear Discriminant	**0.1271**
Knn	**0.0809**

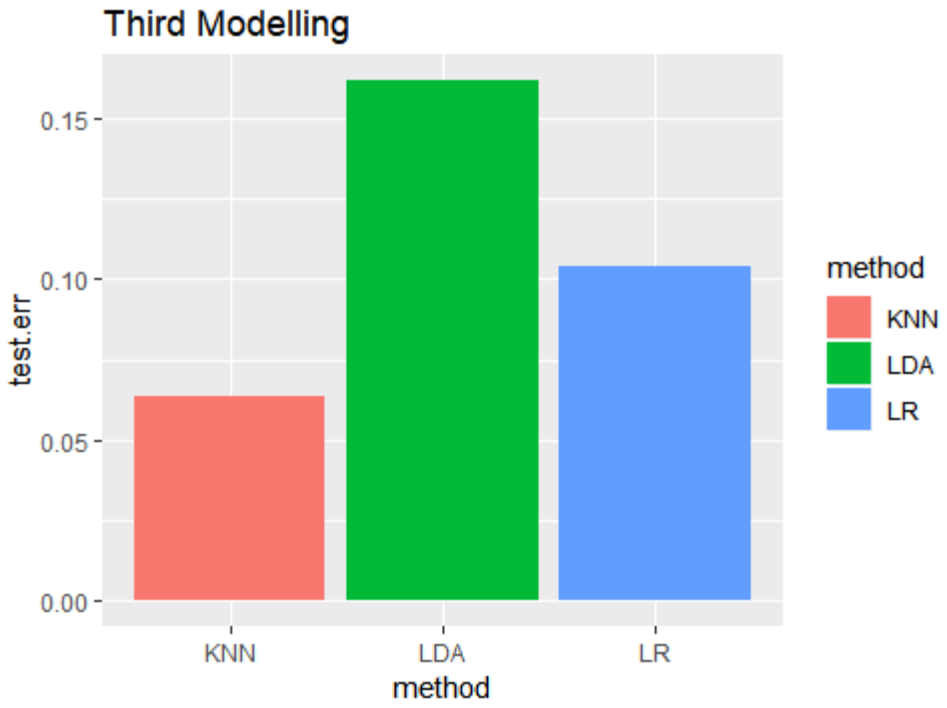
```
> thirdModelling
  method   test.err
1     LR 0.06936416
2    LDA 0.12716763
3    KNN 0.08092486
>
```

When we look into the test error values for different models Linear performs the best is because of the reason I removed the variables which are Non-linear to the output response. Now the logistic model performs better than the other models since the input variable is linear.

The below is the histogram for the different error for the first, second and third modeling


First Modelling
0.15
0.10
0.05
0.00
test.err
KNN
LDA
LR
method
method
KNN
LDA
LR


Second Modelling
0.15
0.10
0.05
0.00
test.err
KNN
LDA
LR
method
method
KNN
LDA
LR



Problem 2

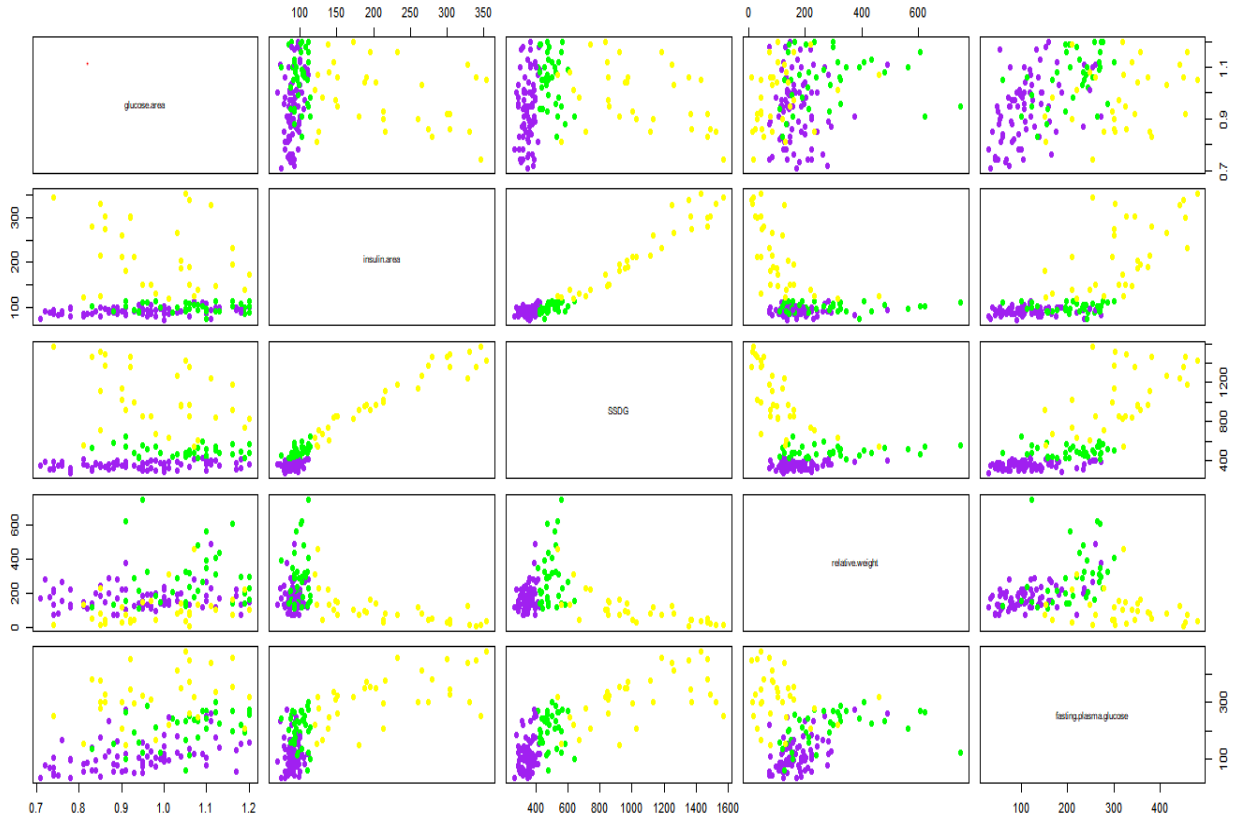
The problem is to apply the Linear and Quadratic Discriminant analysis in the diabetes dataset and predict whether the particular row comes under which class number under LDA and QDA and also we need to find whether the class variable has different covariance matrices and they follow multivariate normal

Knowing the data

Using the Diabetes dataset which is in the MMST package with 145 observation and 7 variables we need to split the data into train data and testing data. Here in this problem, I had split the data into 60 percent of training data and 40 percent of test data. Now, my test data will be having 95 rows and test data will have 50 observations.

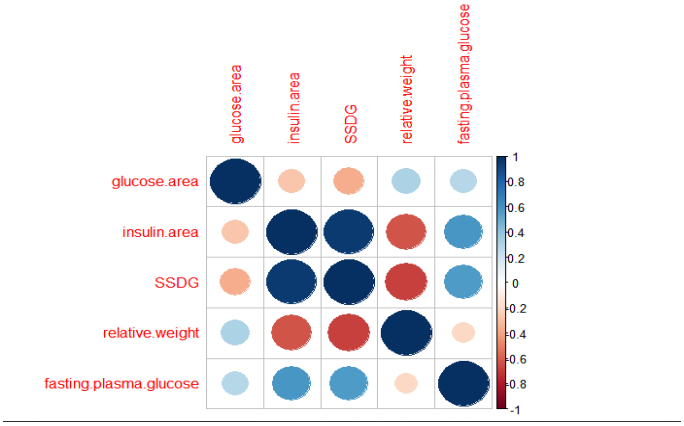
Plotting to scatter plot

A) I had plotted pairwise scatter plot for five variables and three different colors for three different class variable. Yellow represents class 1, green represent class 2 and purple represents class 3 data points respectively. From the pairwise scatter plot we can see that the data of each class has its own covariance distribution.

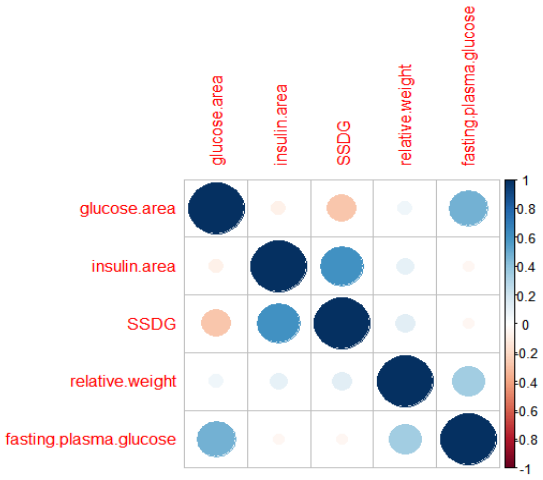


Pairwise scatter plot for six different variables

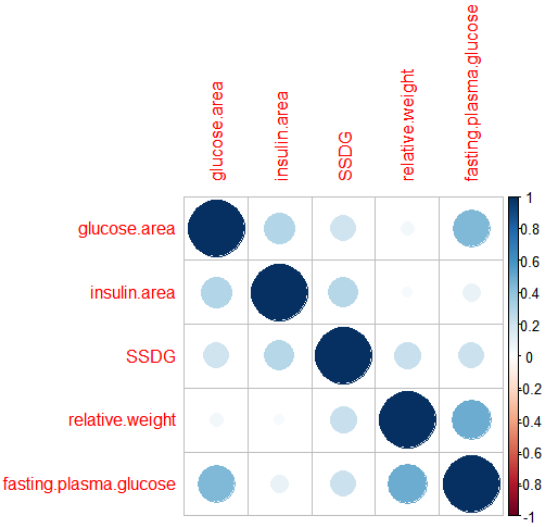
In the above figure when we look into 3rd row 2nd column we can see that purple and green and purple are not scattered but the yellow is scattered along the linear line from this inference we can say that each class has its own covariance distribution. To confirm this I had plotted correlation plot for three different classes we can see that covariance are different . Now I will try to find whether the correlation matrix for the different class matches with above scatter plot



Correlation Matrix for Class 1



Correlation Matrix for Class 2



Correlation Matrix for Class 3

	glucose.area	insulin.area	SSDG	relative.weight	fasting.plasma.glucose
glucose.area	1.00	-0.28	-0.37	0.31	0.28
insulin.area	-0.28	1.00	0.96	-0.63	0.58
SSDG	-0.37	0.96	1.00	-0.69	0.56
relative.weight	0.31	-0.63	-0.69	1.00	-0.20
fasting.plasma.glucose	0.28	0.58	0.56	-0.20	1.00

Correlation matrix for class number 1

```
> Pima.corr_two
                      glucose.area insulin.area  SSDG relative.weight fasting.plasma.glucose
glucose.area                 1.00        -0.07 -0.27            0.07                  0.48
insulin.area                -0.07         1.00  0.61            0.10                 -0.04
SSDG                        -0.27         0.61  1.00            0.12                 -0.04
relative.weight              0.07         0.10  0.12            1.00                  0.34
fasting.plasma.glucose       0.48        -0.04 -0.04            0.34                  1.00
```

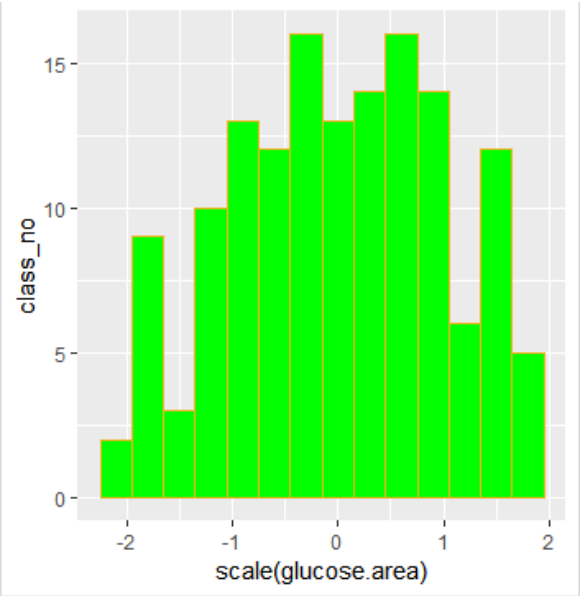
Correlation matrix for class number 2

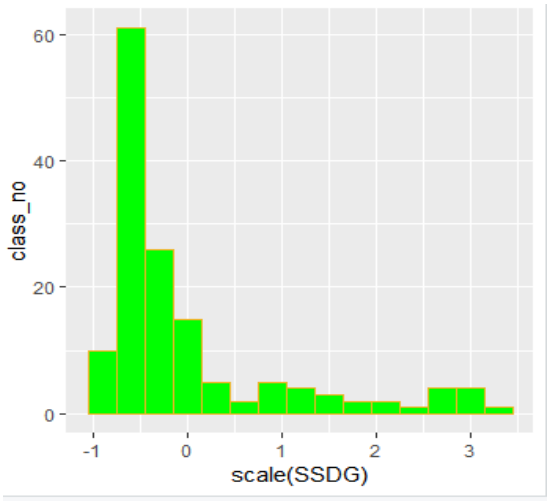
```
> Pima.corr_three
                      glucose.area insulin.area SSDG relative.weight fasting.plasma.glucose
glucose.area                 1.00         0.30 0.21            0.06                  0.45
insulin.area                 0.30         1.00 0.28            0.03                  0.10
SSDG                         0.21         0.28 1.00            0.23                  0.21
relative.weight              0.06         0.03 0.23            1.00                  0.49
fasting.plasma.glucose       0.45         0.10 0.21            0.49                  1.00
```

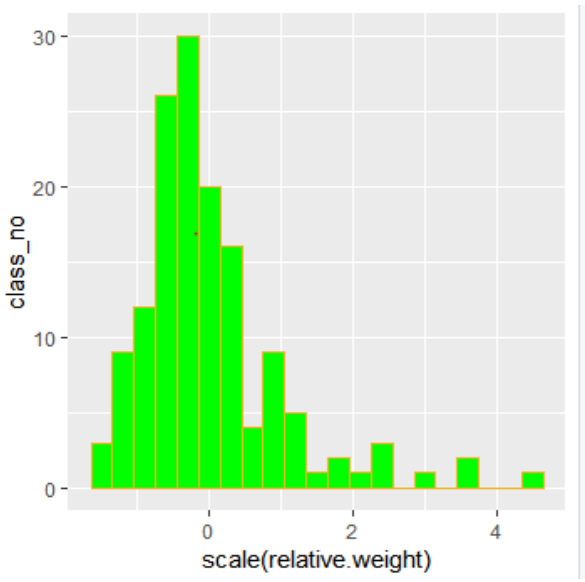
Correlation matrix for class number 3

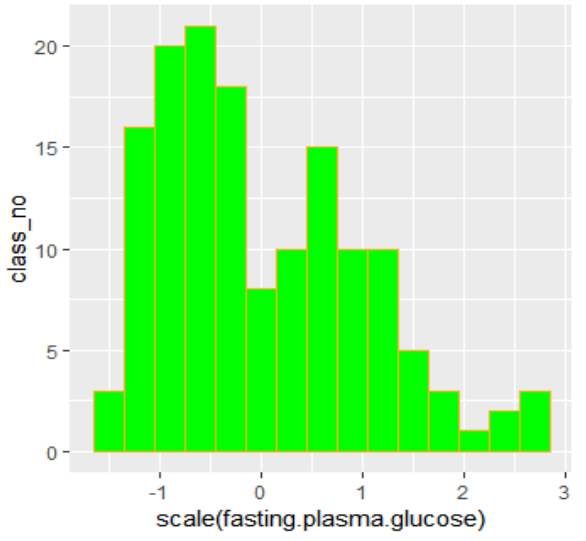
From the correlation matrix, we can see that it is different for different classes. Thus the constructed covariance matrix for data for class and results does support the above observation. Data of each class has its own covariance distribution.

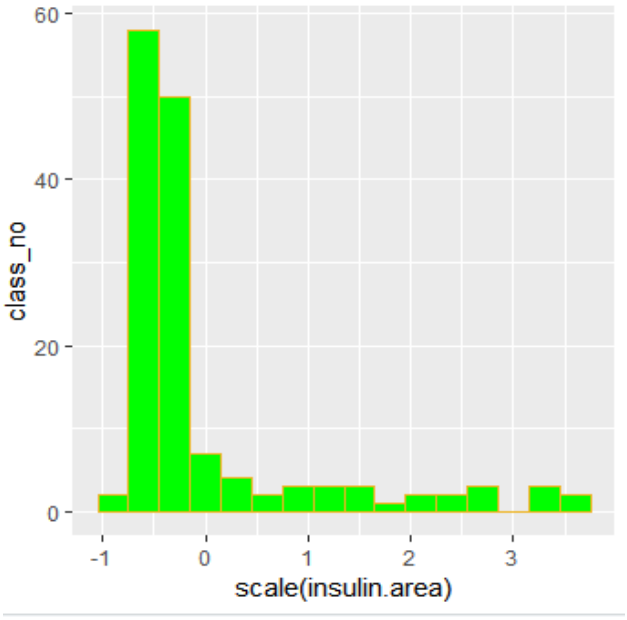
In order to find whether the data set is multivariate normal, we need to plot the histogram for a variable in the dataset and check whether it follows a Gaussian distribution.




60
40
20
0
class_no
-1
0
1
2
3
scale(SSDG)


30
20
10
0
class_no
0
2
4
scale(relative.weight)


20
15
10
5
0
class_no
-1
0
1
2
3
scale(fasting.plasma.glucose)



From the above histogram, we can see that there is no variable is following Gaussian distribution.

One definition is that a random vector is said to be *k*-variate normally distributed if every linear combination of its *k* components has a univariate normal distribution

To confirm this I will be running Mardia's MVN test to check whether there is multivariate normal in the data set. From this, we can calculate the sample skewness and kurtosis to check if they are Gaussian distributed. A Gaussian distributed data set will have skewness = 0 and kurtosis = 3.When we run the MVN command for class 2 it is multivariate normal and for class 1 and class 2 only one of test is passing so as the whole the class variable is not multivariate normal when we combine all the three classes. so I can say only one class 2 it is properly multivariate and for class 1 and class 2 it is not properly multivariate

```
> result <- mvn(data = pima, mvnTest = "mardia")
> result$multivariateNormality
             Test         Statistic              p value Result
1 Mardia Skewness 528.466144798077   1.244386084343e-65     NO
2 Mardia Kurtosis 8.02044619307852 1.11022302462516e-15     NO
3             MVN              <NA>                 <NA>     NO
```

Mardia's MVN test

Test	Statistic	p value	Result
Mardia Skewness	72.1130242128528	0.000223018134108012	NO
Mardia Kurtosis	0.759602021039537	0.447492510707829	YES
MVN	<NA>	<NA>	NO

MVN test for class 1

Test	Statistic	p value	Result
Mardia Skewness	72.530319596442	0.0678591919038945	YES
Mardia Kurtosis	-0.487189189227454	0.626124265808599	YES
MVN	<NA>	<NA>	YES

MVN test for class 2

Test	Statistic	p value	Result
Mardia Skewness	99.0372704053539	0.000345922647527644	NO
Mardia Kurtosis	1.68621376684456	0.0917546332080277	YES
MVN	<NA>	<NA>	NO

MVN test for class 3

Thus the test confirms there is no multivariate normal in the variables of the dataset.

Performing LDA and QDA

B) When we apply the LDA and QDA into dataset the test error and training error is calculated. From the result obtained we can infer that QDA performs better than the LDA

Models	Test Error
LDA	0.26
QDA	0.24

Quadratic discriminant analysis (QDA) is closely related to linear discriminant analysis (LDA), where it is assumed that the measurements from each class are normally distributed. Unlike LDA however, in QDA there is no assumption that the covariance of each of the classes is identical . Because of this reason since there is no identical covariance in data set QDA is performing better compared to LDA .

This is because of the reason that data have the same covariance matrix for all the classes, which is one of the assumptions for LDA to work correctly and QDA does not need to have a

common covariance matrix. From the above observation, we can infer from both pairwise scatter plot and covariance matrix class variables have different covariance. And as QDA performs better than LDA, I can imagine that the non-linear decision boundary helps this decision. So the non-parametric method presents the best results.

C) For the given data the individual has class 2 for LDA. And the individual has class 3 for QDA. For the different levels, the probability is the maximum for Level 3 in the QDA model.

```
$`class`
[1] 2
Levels: 1 2 3

$posterior
          1         2             3
1 0.4784871 0.5215104 2.497132e-06
```

Prediction for the LDA model

```
$`class`
[1] 2
Levels: 1 2 3

$posterior
          1         2             3
1 0.4784871 0.5215104 2.497132e-06
```

Prediction for QDA model

3 a). We need to prove that sum of the posterior probabilities equals to 1

$$\sum_{k=1}^{k} P_r\left(G = k \mid X = x\right) = 1$$

Under the assumptions, the posterior probabilities are given by
Consider first equation

$$P_r(G = k \mid X = x) = \frac{e^{(\beta_{k0} + \beta_k^T x)}}{1 + \sum_{l=1}^{k-1} e^{(\beta_{l0} + \beta_l^T x)}} \; \forall\, k = 1,2, \ldots . k-1$$

Now consider second equation

$$P_r(G = K|\, X = x) = \frac{1}{1 + \sum_{K=1}^{K-1} e^{(\beta_{l0} + \beta_l^T x)}}$$

Now consider third equation

$$\log \frac{P_r(G = k \mid X = x)}{P_r(G = K \mid X = x)} = \beta_{k0} + \beta_k^T x$$

Now converting log function into exponential

$$e^{(\beta_{k0} + \beta_k^T x)} = \frac{P_r(G = k \mid X = x)}{P_r(G = K|\, X = x)}$$

Now subs equation three into equation 1,

$$P_r(G = k \mid X = x) = \frac{\frac{P_r(G = k \mid X = x)}{P_r(G = K \mid X = x)}}{\frac{1 + \sum_{l=1}^{k-1} P_r(G = l|\, X = x)}{P_r(G = K \mid X = x)}}$$

On cancelling the denominators we get,

$$P_r(G = k \mid X = x) = \frac{P_r(G = k \mid X = x)}{\sum_{k=1}^{k} P_r(G = k \mid X = x)}$$

$$P_r(G = K \mid X = x) = \frac{P_r(G = K \mid X = x)}{\sum_{k=1}^{k} P_r(G = k \mid X = x)}$$

$$\sum_{k=1}^{k} P_r(G = k \mid X = x) = \sum_{k=1}^{k-1} P_r(G = k \mid X = x) + P_r(G = K \mid X = x)$$

$$\sum_{k=1}^{k} P_r(G = k \mid X = x) = \frac{\sum_{k=1}^{k-1} P_r(G = k \mid X = x)}{\sum_{k=1}^{k} P_r(G = k \mid X = x)} + \frac{P_r(G = K \mid X = x)}{\sum_{k=1}^{k} P_r(G = k \mid X = x)}$$

Hence we get, Both the numerator and denominator are the same
Thus the assumptions are satisfied

$$\frac{\sum_{k=1}^{k} P_r(G = k \mid X = x)}{\sum_{k=1}^{k} P_r(G = k \mid X = x)} = 1$$

3 b)

$$1 - p(X) = 1 - \frac{e^{(\beta_0 + \beta_1 X)}}{1 + e^{(\beta_0 + \beta_1 X)}} = \frac{1}{1 + e^{(\beta_0 + \beta_1 X)}}$$

$$\frac{1}{1-p(X)} = 1 + e^{(\beta_0 + \beta_1 X)}$$

$$p(X) * \frac{1}{1-p(X)} = \frac{e^{(\beta_0 + \beta_1 X)}}{1 + e^{(\beta_0 + \beta_1 X)}} \left(1 + e^{(\beta_0 + \beta_1 X)}\right),$$

$$\frac{p(X)}{1-P(X)} = e^{(\beta_0 + \beta_1 X)}$$