

Question 1.

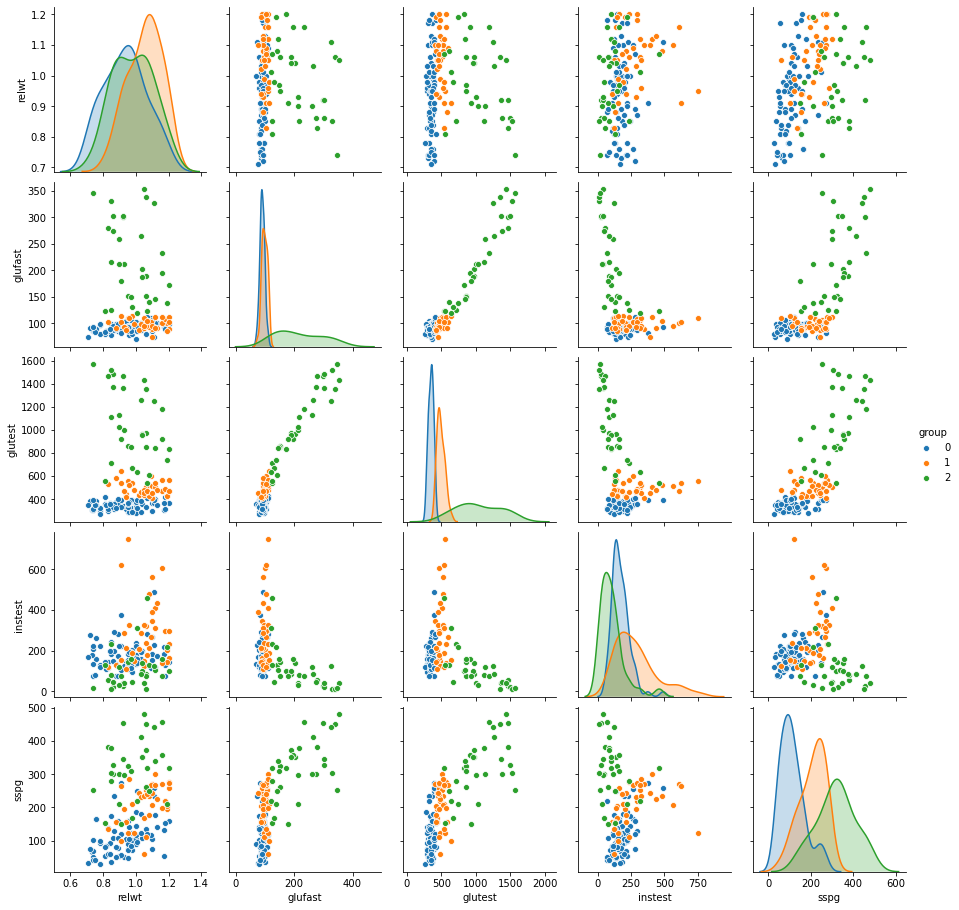
a)

Process:

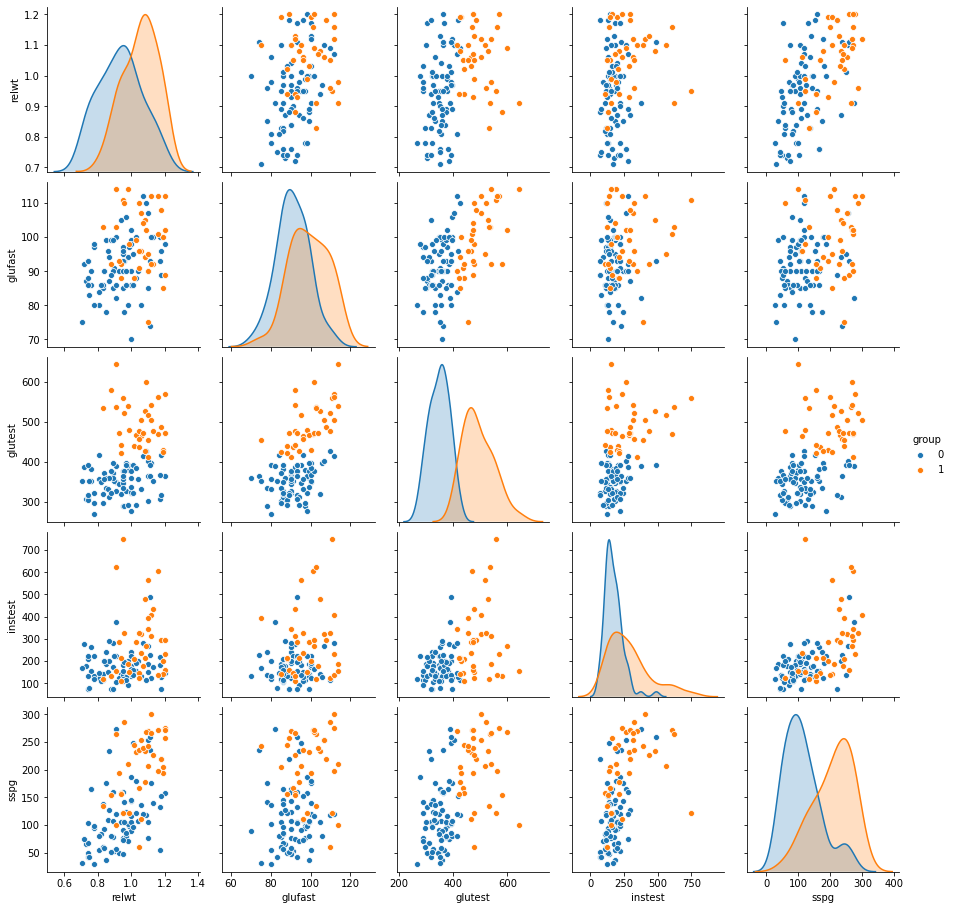
1. **Read data**: I read ‘Diabetes’ dataset.
2. **change string Y(group) categorical data into num value**: To verify the Y(group) values, I used the ‘unique’ function and changed it into the number by using the ‘map’ function.
3. **Plot the pairplot**: I plotted the pairplot with different colors representing the three different classes. Also, to confirm the covariance of three different classes at the same time, I plotted three sets of two different classes (ChemicalDiabetic - OvertDiabetic, Normal- -Overt\_Diabetic, Normal-ChemicalDiabetic).
4. **Show the covariance matrix:** To check the covariance value, I made three covariance matrices by three groups.

Outputs:

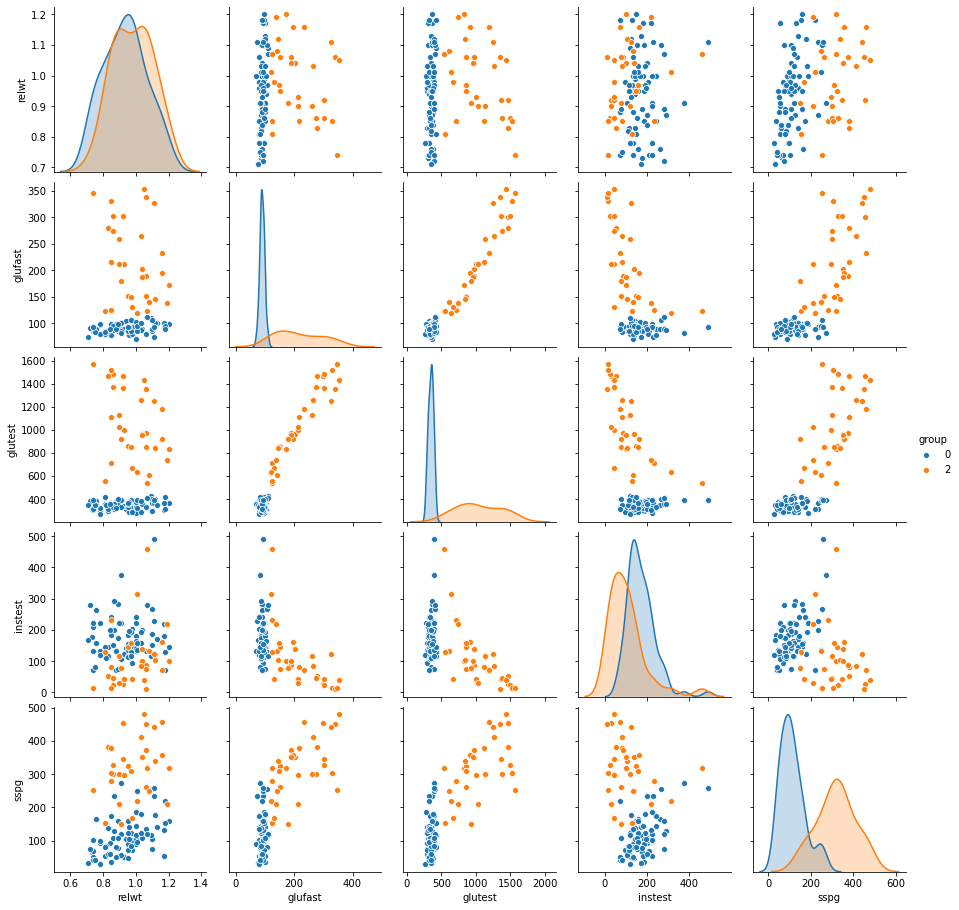
**Pair plot (Normal - Chemical\_Diabetic - overt\_Diabetic)**



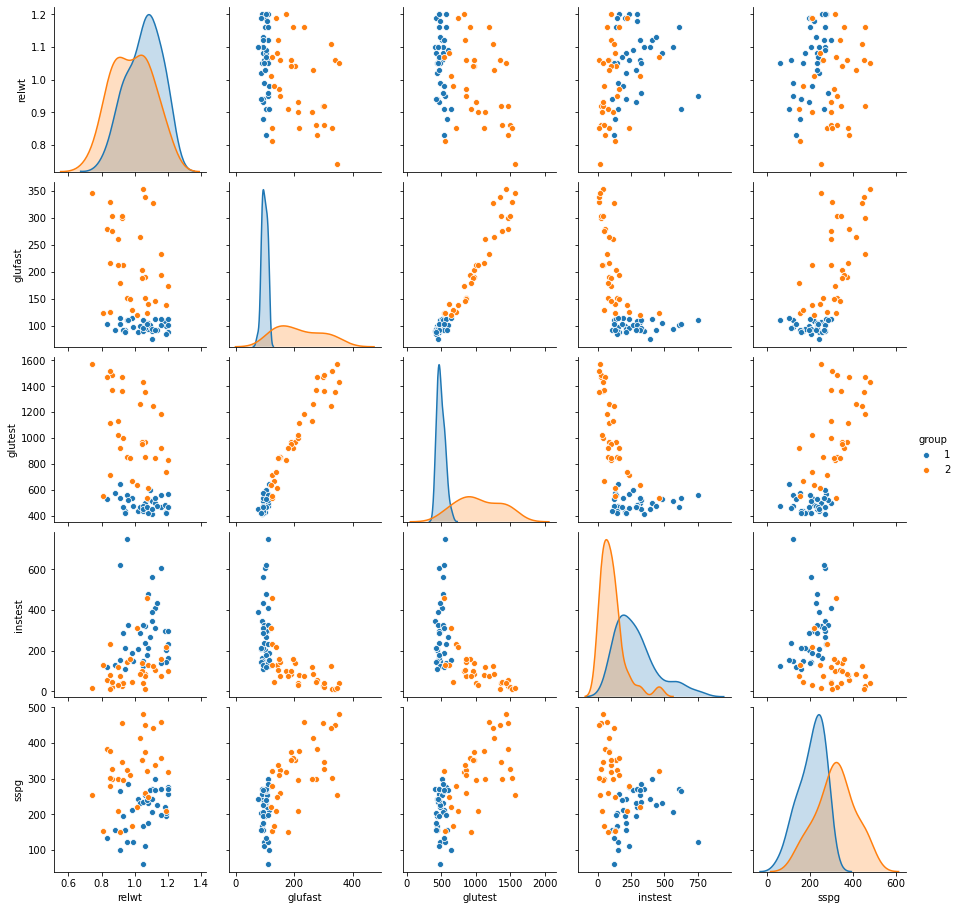
**Pair plot (Normal - Chemical\_Diabetic)**



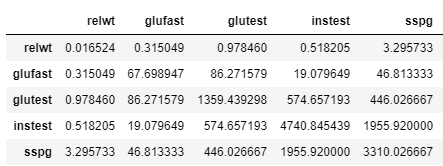
**Pair plot (Normal - overt\_Diabetic)**

****

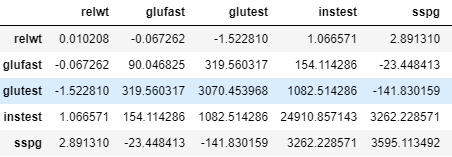
**Pair plot (Chemical\_Diabetic - overt\_Diabetic)**



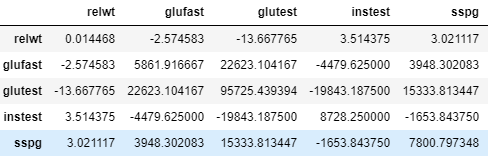
**Covariance matrix for Nomal**



**Covariance matrix for Chemical\_Diabetic**

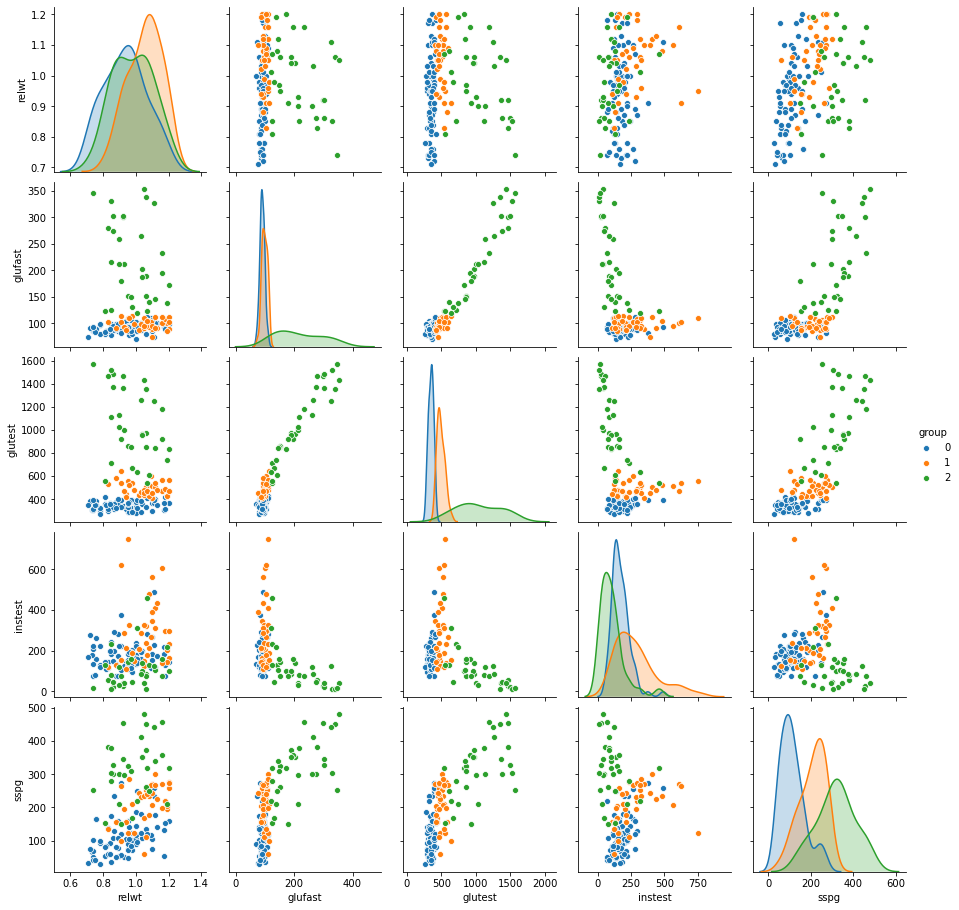


**Covariance matrix for Overt\_Diabetic**



Consideration:

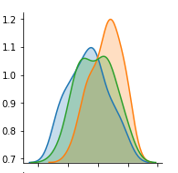
**Comparing each three group’s covariances with distribution.**



Let’s guess the similarity of the three group’s covariance with distribution by factors.

**Factor 1. Relwt (BLUE: Normal, Orange: Chemical\_Diabetic, Green: overt\_Diabetic)**

*\*three of them have multivariate normal\**



Here is the distribution graph for ‘Relwt’ by every three groups. Although **Normal** and **Overt\_Diabetic** have a more similar distribution than **Chemical\_Diabetic,** it is to say three of them has similar distribution.

**Covariance value**

*BLUE: Normal => 0.016524*

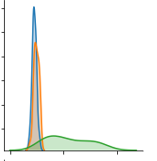
*Orange: Chemical\_Diabetic => 0.010208*

*Green: overt\_Diabetic => 0.014468*

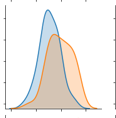
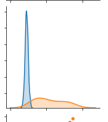
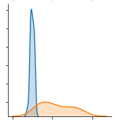
We can confirm this fact from the result of the exact values of the covariance.

**Factor 2. glufast (BLUE: Normal, Orange: Chemical\_Diabetic, Green: overt\_Diabetic)**

*\*Normal – chemical\_Diabetic relation has multivariate normal distribution\**



Here is the distribution graph for ‘glufast’ by each three groups. **Normal** and **Chemical \_Diabetic** has similar distribution while **overt \_Diabetic** has exactly different distribution.

1. 2. 3.

***1. Normal - Chemical\_Diabetic - overt\_Diabetic******2. Normal - overt\_Diabetic******3. Chemical\_Diabetic - overt\_Diabetic***

If we compare the three graphs above, it is clear that the **Normal** and **Chemical Diabetic** has a similar distribution. However, **Overt\_Diabetic** has a different distribution from the other two classes

**Covariance value**

*BLUE: Normal => 67.69*

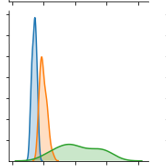
*Orange: Chemical\_Diabetic => 90.04*

*Green: overt\_Diabetic => 5861.91*

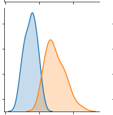
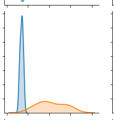
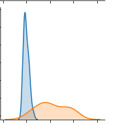
We can confirm this fact from the result of the exact values of the covariance.

**Factor 3. glutest (BLUE: Normal, Orange: Chemical\_Diabetic, Green: overt\_Diabetic)**

*\*Normal – chemical\_Diabetic relation has multivariate normal distribution\**



Here is the distribution graph for ‘glutest’ by three groups. **Normal** and **Chemical \_Diabetic** have similar distribution while **overt \_Diabetic** has an exactly different distribution.

1.2. 3. 

***1. Normal - Chemical\_Diabetic - overt\_Diabetic******2. Normal - overt\_Diabetic******3. Chemical\_Diabetic - overt\_Diabetic***

If we compare the three graphs above, it is clear that the **Normal** and **Chemical Diabetic** has a similar distribution. However, **Overt\_Diabetic** has a different distribution from the other two classes.

**Covariance value**

*BLUE: Normal => 1359*

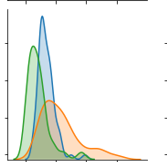
*Orange: Chemical\_Diabetic => 3070*

*Green: overt\_Diabetic => 95725*

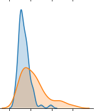
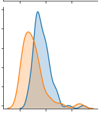
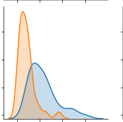
We can confirm this fact from the result of the exact values of the covariance.

**Factor 4. instest (BLUE: Normal, Orange: Chemical\_Diabetic, Green: overt\_Diabetic)**

*\*Normal – overt\_Diabetic relation has multivariate normal distribution\**



Here is the distribution graph for ‘instest’ by three groups. **Normal** and **Overt\_Diabetic** have similar distribution compared with **Chemical\_Diabetic**. But differences in covariance among the three classes is not far as much as factor 2(glufast) and factor 3(glutest).

1.2. 3. 

***1. Normal - Chemical\_Diabetic - overt\_Diabetic******2. Normal - overt\_Diabetic******3. Chemical\_Diabetic - overt\_Diabetic***

If we compare the three graphs above, it is clear that the **Normal** and **overt Diabetic** has a similar distribution. However, **Factor 4(instest)** has more similar three distribution than **Factor 2** and **Factor 3.**

**Covariance value**

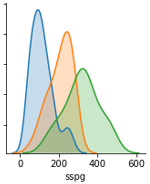
*BLUE: Normal => 4740*

*Orange: Chemical\_Diabetic => 24910*

*Green: overt\_Diabetic => 8728*

**Factor 5. sspg (BLUE: Normal, Orange: Chemical\_Diabetic, Green: overt\_Diabetic)**

*\*three of them have multivariate normal\**



Here is the distribution graph for ‘sspg’ by three groups. **Normal** and **Chemical \_Diabetic** have similar distribution while **overt \_Diabetic** has a little different distribution. Although **Normal**, **Overt\_Diabetic**, and **Chemical\_Diabetic** have not exacted similar distribution like **Factor 1(relwt)**, it is acceptable to say that 3 of them have a similar distribution in **Factor 5(sspg)** rather than the other factors.

**Covariance value**

*BLUE: Normal => 3310*

*Orange: Chemical\_Diabetic => 3595*

*Green: overt\_Diabetic => 7800*

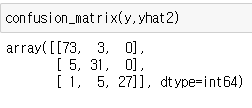
B)

Process:

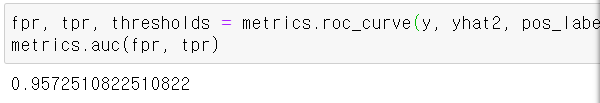
1. **Continue the process a**:
2. **Split Train and Test data**: I divide 30 percent of the data into the test data, and the rest of them became the train data.
3. **Fit the model**: With the train data, fit LDA and QDA model.
4. **Predict the test data and show the confusion matrix:** By using a fitted LDA and QDA model, predict the test data. With predictions, make two confusion matrices to

Outputs:

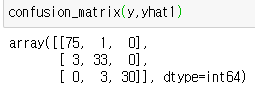
**Confusion matrix for the LDA model**



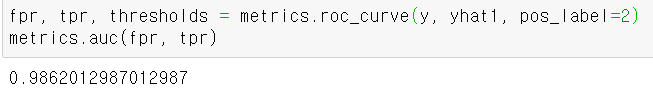
**AUC value for the LDA model**



**Confusion matrix for the QDA model**



**AUC value for the QDA model**



Consideration:

*THE QDA MODEL MADE BETTTER PREDICTION THAN THE LDA MODEL BECAUSE OF DISTRIBUTION FOR THE EACH GROUP OF DATA.*

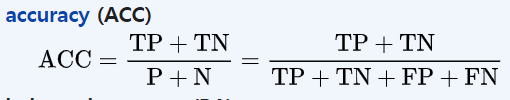
**Model accuracy**

**LDA = 131/145**

**QDA = 138/145**

The ACCURACY = **correct predictions** **/ (correct predictions + incorrect predictions)**

This accuracy also matches with the accuracy in the confusion matrix.



**(correct predictions + incorrect predictions)** indicate total rows which is same with (TP+TN) + (FP+FN)

**correct predictions** indicate TP + TN.

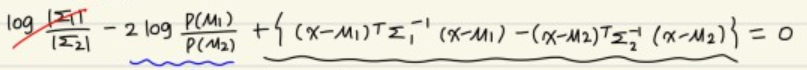
We can see that the QDA model has better accuracy in test prediction than the LDA model. We can find the reason for this from **Question a**.

When we conduct discriminant analysis in the data set, distribution for each group data is the most important factor. **Here are the reasons for this**.

* **The two group of data have same distribution (Σ1 = Σ2)**

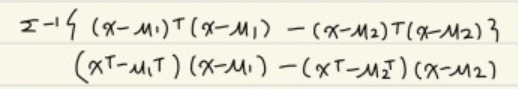
Σ1 = Σ2 = pooled variance = ((n1-1) Σ1 +(n2-1) Σ2)/ (n1 + n2 -2) = Σ

Apply it into discriminant function

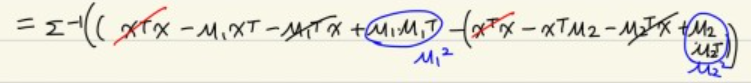


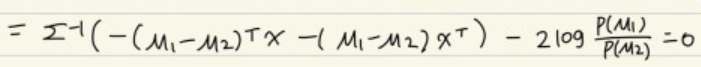
Blue line is a constant.

Rearrange the equation for the Black line with Σ-1 (notice that Σ1 = Σ**2**)



Multiply each equation and make it simpler.





The equation becomes like the picture above.

This arranged equation became the linear function form as like **AX + B = 0.**

This is because the two groups have in pooled variance **(Σ1 = Σ2)**

**In this case we can see the division boundary became linear.**

* **The two group of data have different distribution (Σ1 =! Σ2)**

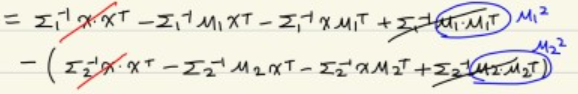
Let’s do the same process but with two different distribution.



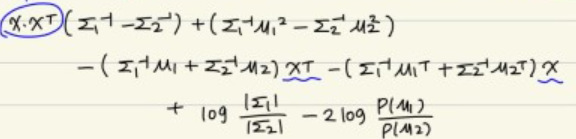
First two log equations are constants.



Except two constants, the equation can be changed like the picture above.



Multiply each equation term than we can see the



Then we can rearrange the equation by X\*XT, XT, and X

This arranged equation became the quadratic function form as like **AX^2 + BX +C = 0.**

This is because the two groups have different distribution **(Σ1 =! Σ2)**

**In this case, we can see the decision boundary became a quadratic function form.**

For these two reasons, we can guess the QDA model which used the polynomial decision boundary for the classification can get better accuracy than the LDA model which uses the Linear decision boundary for the classification in the dataset.

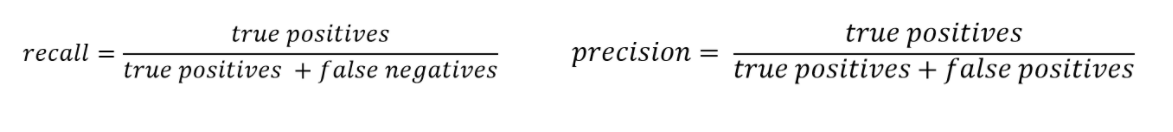
We can see the three groups in our dataset has different distribution in our **Question a.**

Therefore, we can see the result that => **LDA = 131/145 < QDA = 138/145**

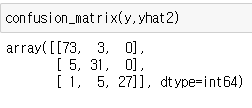
**Let’s compare two predicted data from the two models by the Confusion matrix.**

*Except for the value for Recall(****Overt\_Diabetic****), all the recall and precision values for three groups indicate that the QDA model has better performance than the LDA.*

*Also, if we compare two* ***AUC values*** *for the predicted data from the LDA and the QDA model, the QDA model has better performance than the LDA.(****LDA:0.9572.. < QDA:0.9826..****)*

When We calculate fail rates by 3 sets of each two groups among three for the **LDA and QDA** model, the QDA got lower rates in all three cases.

**Confusion matrix for the LDA model**



**In the test confusion matrix for LDA there are three values**

**'Normal':0, 'Chemical\_Diabetic':1, 'Overt\_Diabetic':2**

Precision (**Normal**) = 73 / 73 + 3 + 0 = 0.96..

Precision (**Chemical\_Diabetic**) = 31 / 5+ 31 + 0 = 0.86..

Precision (**Overt\_Diabetic**) = 27 / 1+ 5 + 27 = 0.81..

Recall(**Normal**) = 73 / 73 + 5 + 1 = 0.92..

Recall(**Chemical\_Diabetic**) = 31 / 3 + 31 + 5 = 0.79

Recall(**Overt\_Diabetic**) = 27 / 0+ 0 + 27 = 1

We can also calculate fail rates by every two groups for the **LDA** model. This result will show that what decision boundary is hard for the model to classify two groups.

To classify **'Chemical\_Diabetic':1** and **'Overt\_Diabetic':2** is harder than others.

**'Normal':0, 'Chemical\_Diabetic':1**

(3+5)/(73+31+3+5) = 0.071..

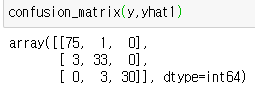
**'Chemical\_Diabetic':1, 'Overt\_Diabetic':2**

5/(31+27+5) =0.079..

**'Normal':0, 'Overt\_Diabetic':2**

1/(73+27+1) = 0.009

**Confusion matrix for the QDA model**



**In the test confusion matrix for QDA there are three values**

**'Normal':0, 'Chemical\_Diabetic':1, 'Overt\_Diabetic':2**

Precision (**Normal**) = 75 / 75 + 1 + 0 = 0.98..

Precision (**Chemical\_Diabetic**) = 33 / 3+ 33 + 0 = 0.91..

Precision (**Overt\_Diabetic**) = 30 / 0+ 3 + 30 = 0.90..

Recall(**Normal**) = 75 / 75 + 3 + 0 = 0.96..

Recall(**Chemical\_Diabetic**) = 33 / 3 + 33 + 0 = 0.91

Recall(**Overt\_Diabetic**) = 30 / 0+ 3 + 30 = 90..

We can also calculate fail rates by every two groups for the **QDA** model. This result will show that what decision boundary is hard for the model to classify two groups.

To classify **'Chemical\_Diabetic':1** and **'Overt\_Diabetic':2** is harder than others.

**'Normal':0, 'Chemical\_Diabetic':1**

(3+1)/(75+33+3+1) = 0.035..

**'Chemical\_Diabetic':1, 'Overt\_Diabetic':2**

3/(33+30) =0.047..

**'Normal':0, 'Overt\_Diabetic':2**

0

C)

Process:

1. **Continue the process a and b**:
2. **Make the individual datafram**: I made a datafram with [1.86,184,68,122,544] which is reflected the condition for the question.
3. **Predict the data with the LDA model and QDA model:** With the LDA and QDA model we made previous, predict the datafame, and compare two results.

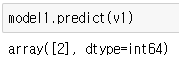
Outputs:

**Result with the LDA model**



* Normal

**Result with the QDA model**



* Overt\_Diabetic

Consideration:

As I mentioned in question A, the data are not multivariate normal. Also, I mentioned in Question B that when the three class of distributions are different from to each other, **the decision boundary becames a quadratic function form.** In this case, we can guess the QDA model which uses the polynomial decision boundary for the classification can get a better prediction than the LDA model which uses the Linear decision boundary for the classification in the dataset.

Reference:

#http://lijiancheng0614.github.io/scikit-learn/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html#sklearn.decomposition.LatentDirichletAllocation.fit\_transform

#https://blog.naver.com/jaehong7719/221926671654

#https://blog.naver.com/powerparan/221867153428

#\*\*\*https://blog.naver.com/sanghan1990/221126257295\*\*\*