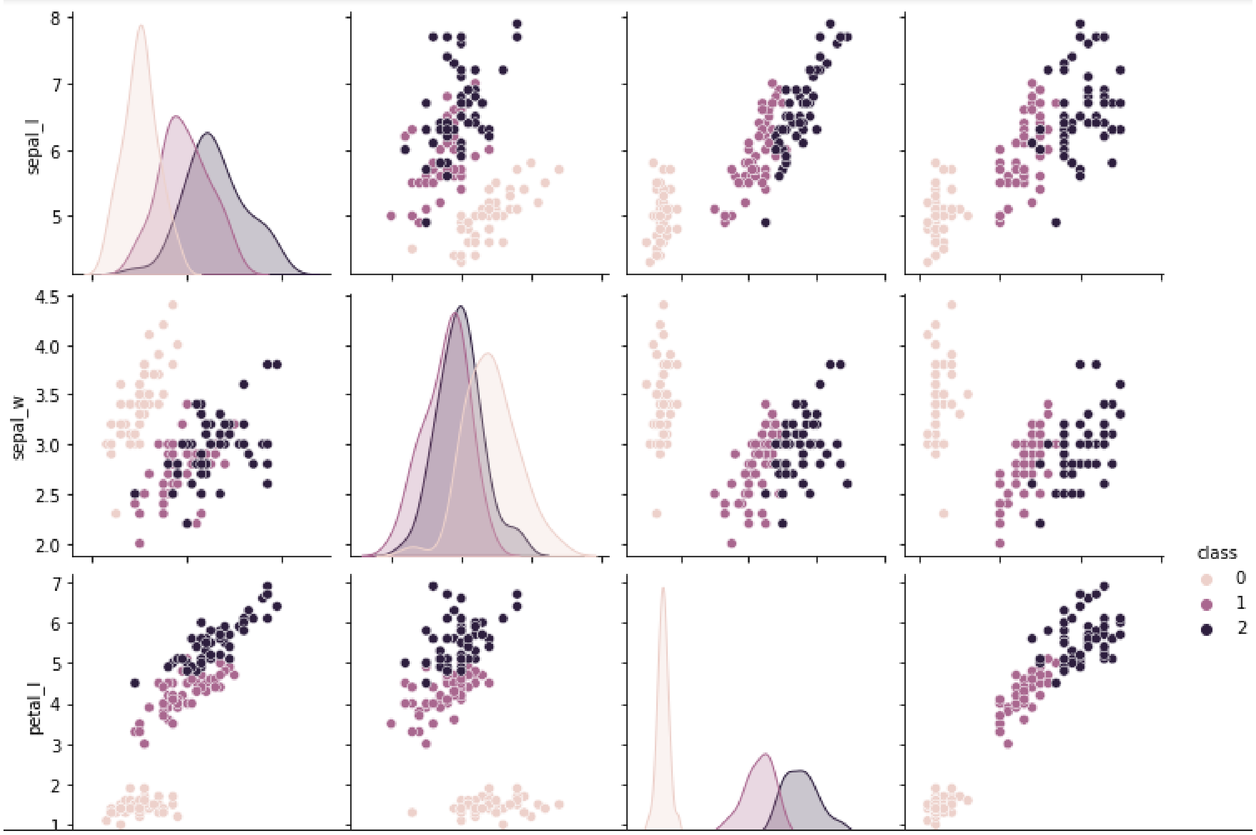
Pattern Recognition and Machine Learning
Report for : Lab 5

Name : Khushi Parikh
Roll No : B20EE029

Question 1:

- Data was imported and split into training and testing sets. We use the stratify argument to make sure all classes are present in the same number in the testing split.



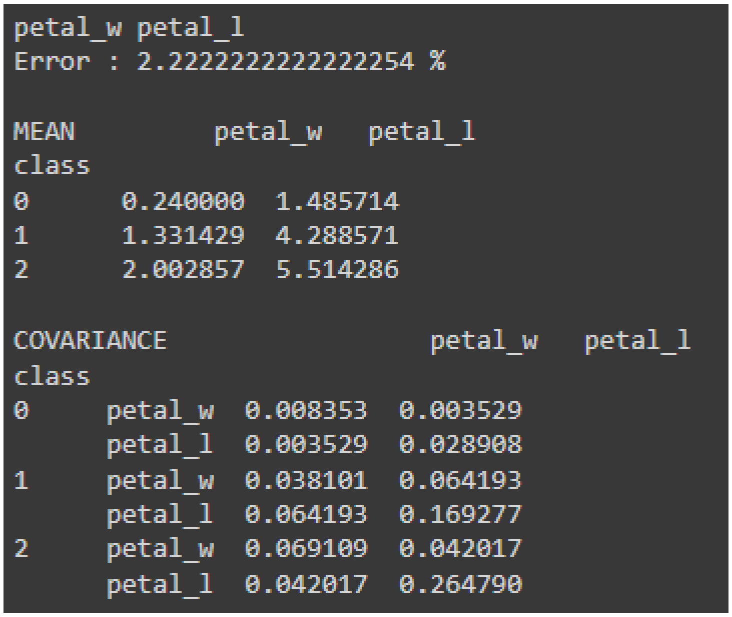
- From this we can see that the data is visible in a distinct manner when petal_w is taken with the other three features.

-

```
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis
```

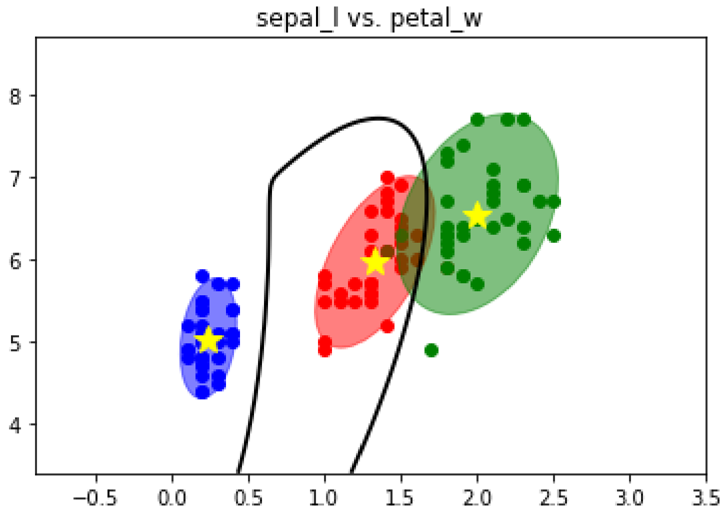
  was imported and models were trained.

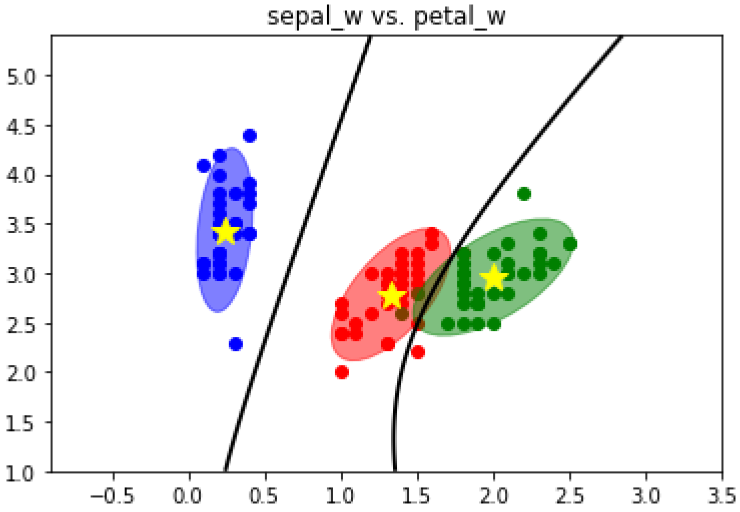
- The following values were obtained for petal_w and petal_l

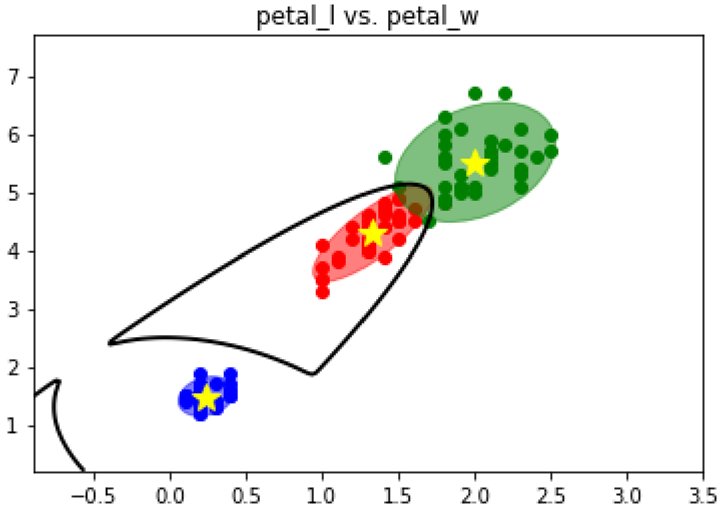


```
petal_w petal_l
Error : 2.2222222222222254 %

MEAN          petal_w   petal_l
class
0      0.240000  1.485714
1      1.331429  4.288571
2      2.002857  5.514286

COVARIANCE                   petal_w   petal_l
class
0     petal_w  0.008353  0.003529
      petal_l  0.003529  0.028908
1     petal_w  0.038101  0.064193
      petal_l  0.064193  0.169277
2     petal_w  0.069109  0.042017
      petal_l  0.042017  0.264790
```

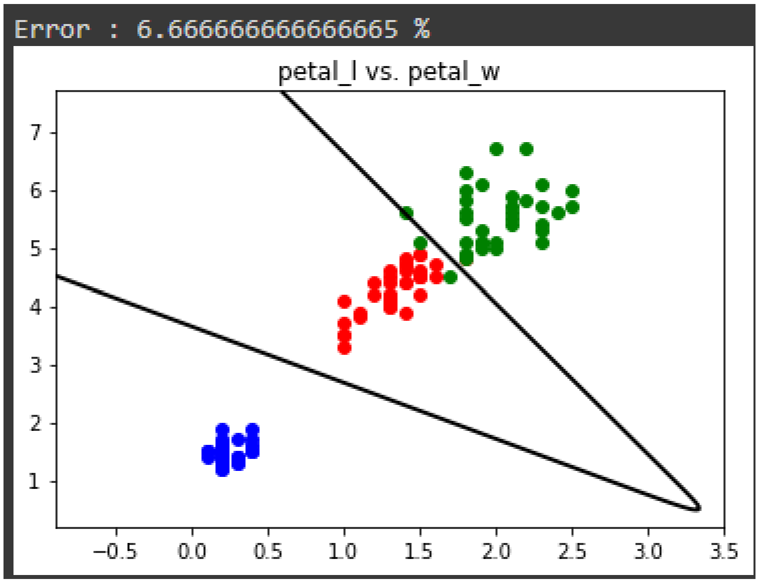
- Decision boundaries were as follows -





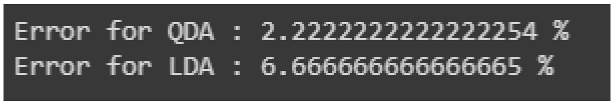


- The minimum error was given by petal_w and petal_l. The LDA model was trained on these features.



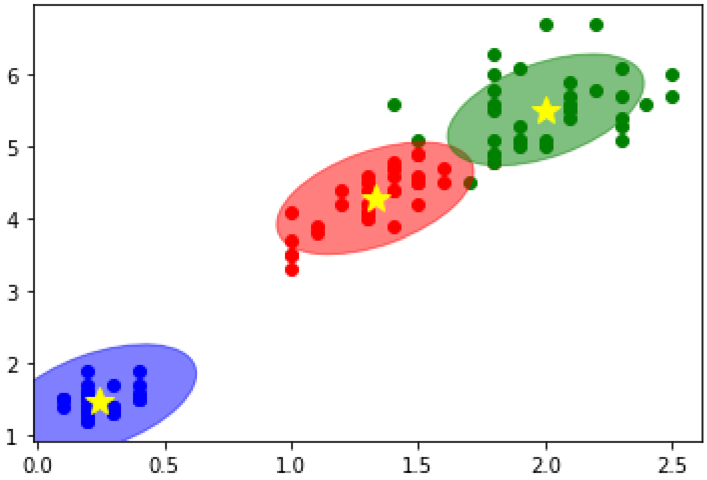
```
Error : 6.666666666666665 %
```

- Error obtained for the two models for parameters petal_l and petal_w :



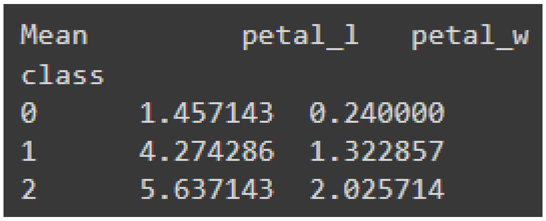
```
Error for QDA : 2.2222222222222254 %
Error for LDA : 6.666666666666665 %
```

- Gaussian Distributions :

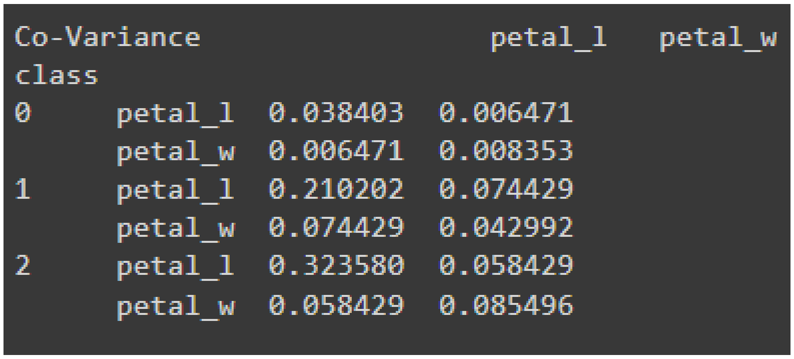


Question 2:

- The iris dataset was imported and two rows were kept according to the question. Mean and covariance matrix for whole data was calculated and the results were -

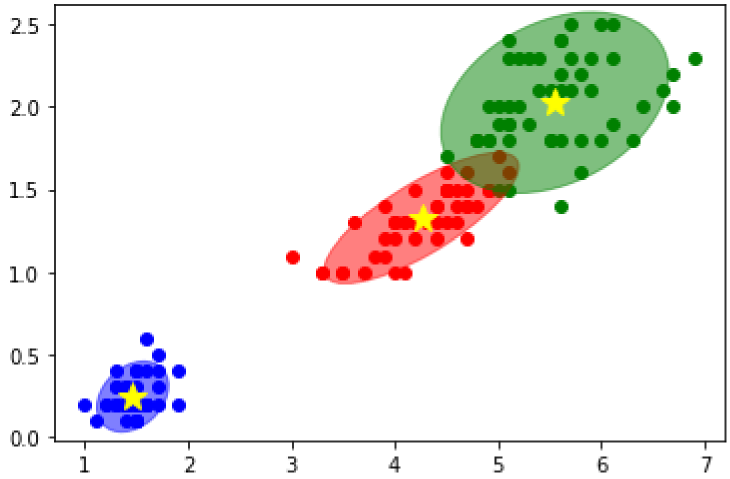


```
Mean          petal_l   petal_w
class
0       1.457143  0.240000
1       4.274286  1.322857
2       5.637143  2.025714
```

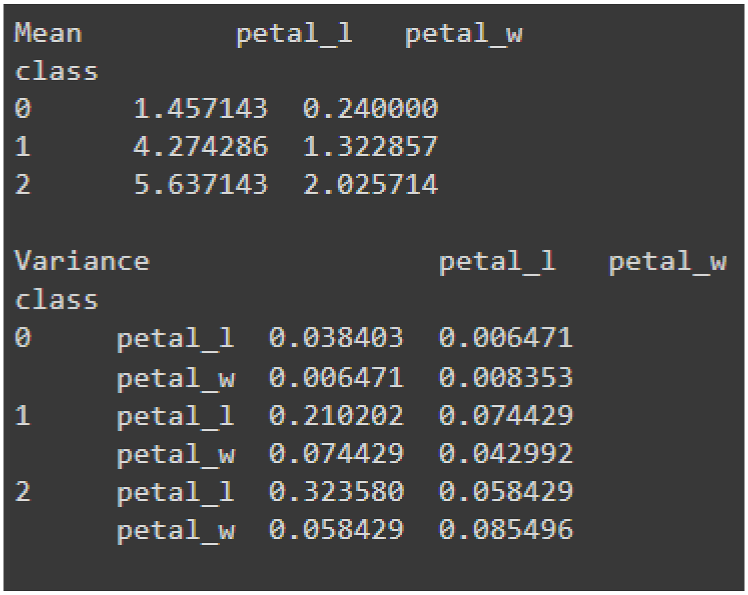


```
Co-Variance                     petal_l   petal_w
class
0      petal_l  0.038403  0.006471
       petal_w  0.006471  0.008353
1      petal_l  0.210202  0.074429
       petal_w  0.074429  0.042992
2      petal_l  0.323580  0.058429
       petal_w  0.058429  0.085496
```

- The data of the whole dataset was plotted using mean and covariance -



- Mean and covariance of the training set was calculated -

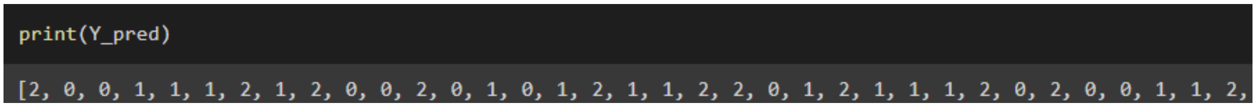


```
Mean          petal_l   petal_w
class
0      1.457143  0.240000
1      4.274286  1.322857
2      5.637143  2.025714

Variance                 petal_l   petal_w
class
0     petal_l  0.038403  0.006471
      petal_w  0.006471  0.008353
1     petal_l  0.210202  0.074429
      petal_w  0.074429  0.042992
2     petal_l  0.323580  0.058429
      petal_w  0.058429  0.085496
```

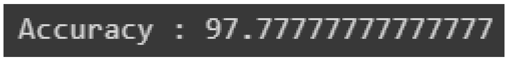
- The multivariate normal density was calculated using the following formula -

$$N(x;\mu,\sigma^2) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu)\right]$$

- Here, x refers to each row which consists of two columns - petal_length and petal_width. For each row of the test data, the likelihood of each of the three labels is calculated and multiplied with the prior. This results in 3 values as there are 3 classes. The index with the maximum probability is chosen as the value for the label. Tha accuracy was then measured -

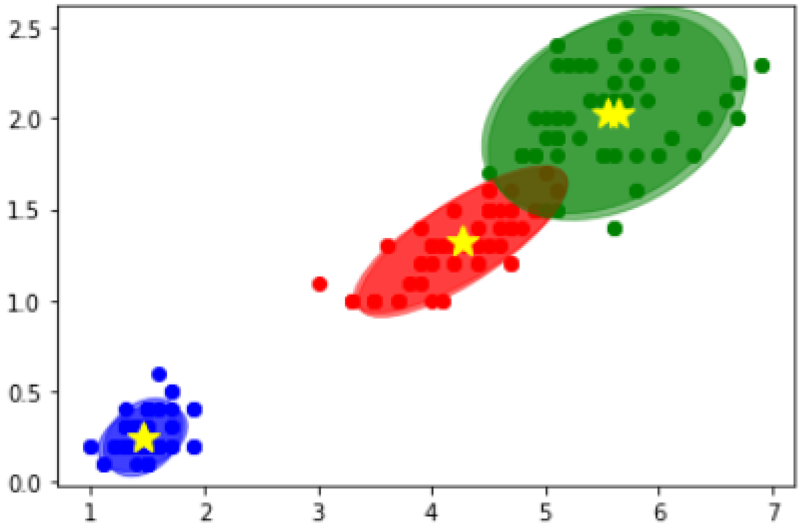


```
print(Y_pred)

[2, 0, 0, 1, 1, 1, 2, 1, 2, 0, 0, 2, 0, 1, 0, 1, 2, 1, 1, 2, 2, 0, 1, 2, 1, 1, 1, 2, 0, 2, 0, 0, 1, 1, 2,
```



```
Accuracy : 97.77777777777777
```

- Plot for distribution of same classes for training and complete data -



Question 3:

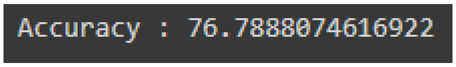
- The label column of the output was added to the training dataset. The training dataset was first divided by label and then each of them were divided by the termId. The count of each class was computed and stored in an array.
- A numpy array of size (20,16688) was initialized to zeros in order to compute the likelihood and was named ll. The value ll[i][j] is the probability(label=i / termId=j).
- The likelihood array contained zeros and hence laplace smoothing was used. The formula for laplace smoothing is -

$$p_{i,\ \alpha\text{-smoothed}} = \frac{x_i + \alpha}{N + \alpha d},$$

- In order to find the label of each of the output data, we have to calculate the MLE.

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \overbrace{P(d|c)}^{\text{likelihood}} \overbrace{P(c)}^{\text{prior}}$$

- For each docID, we calculate the likelihood * prior of each class and calculate its maximum value. This is assigned as the output label of that docId in the test dataset.
- Final Accuracy :



```
Accuracy : 76.7888074616922
```