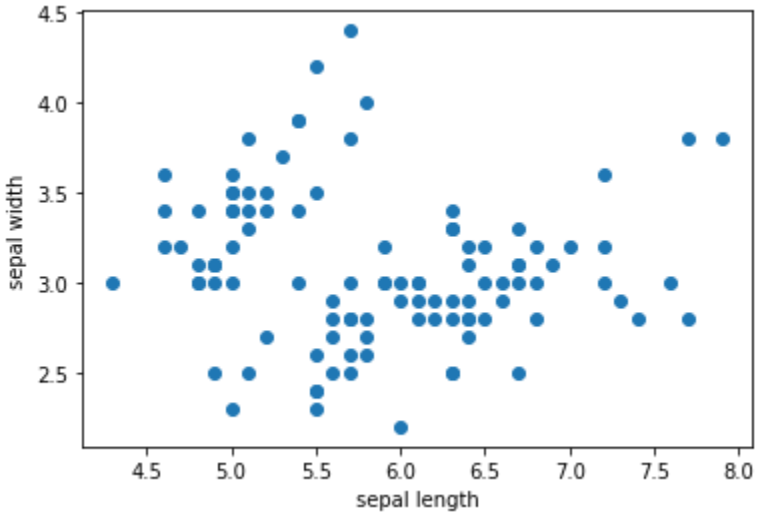
PATTERN RECOGNITION AND MACHINE LEARNING

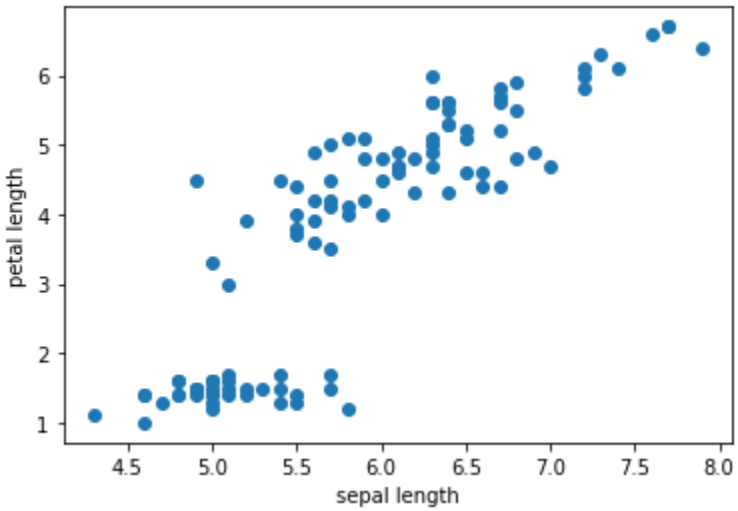
LAB 05
LAB REPORT

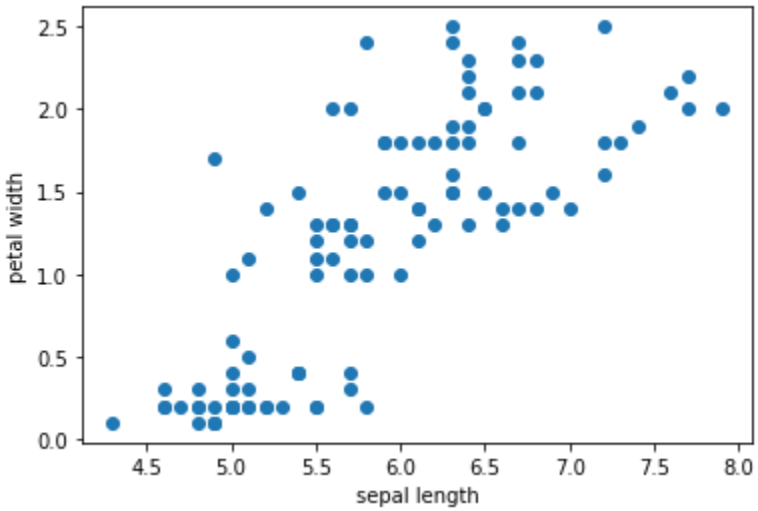
QUESTION 1.

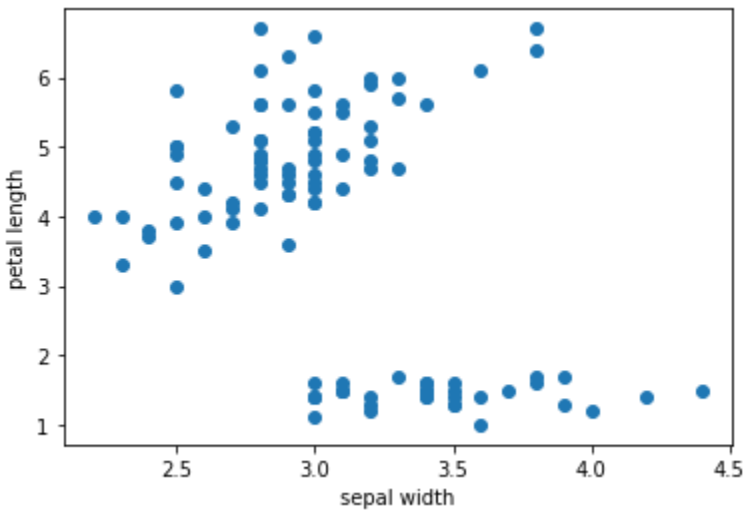
1)

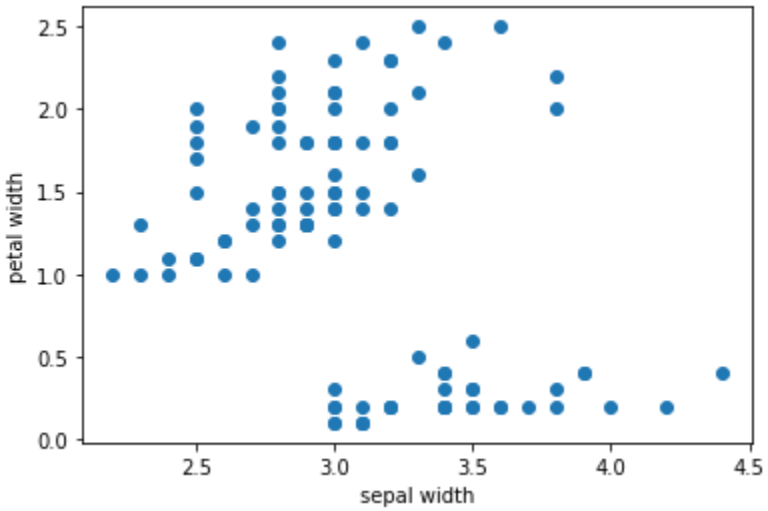
- The data was preprocessed by encoding the class column.
- Data was split into a 70:30 ratio and to ensure about equal occurrences of each class in the split, the stratify parameter in train_test_split was set to y.
- Plots for each pair were plotted:

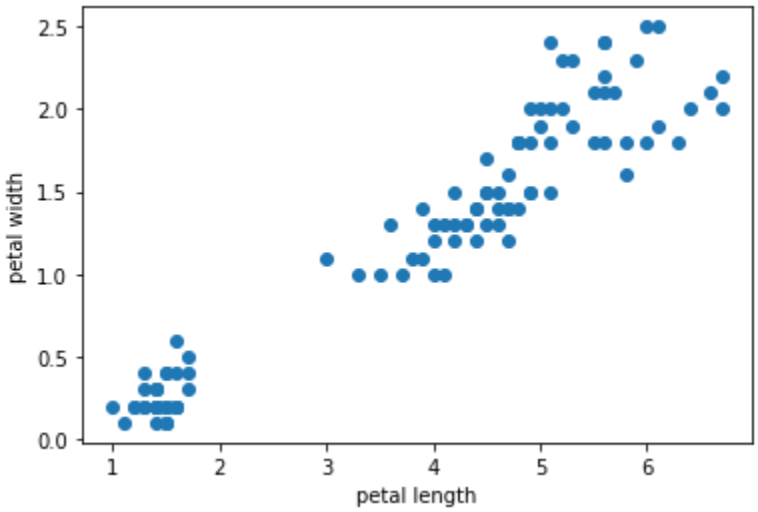



petal length
sepal length
1
2
3
4
5
6
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0


petal width
sepal length
0.0
0.5
1.0
1.5
2.0
2.5
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0


petal length
sepal width
1
2
3
4
5
6
2.5
3.0
3.5
4.0
4.5


petal width
sepal width
0.0
0.5
1.0
1.5
2.0
2.5
2.5
3.0
3.5
4.0
4.5


petal width
petal length
0.0
0.5
1.0
1.5
2.0
2.5
1
2
3
4
5
6

2)

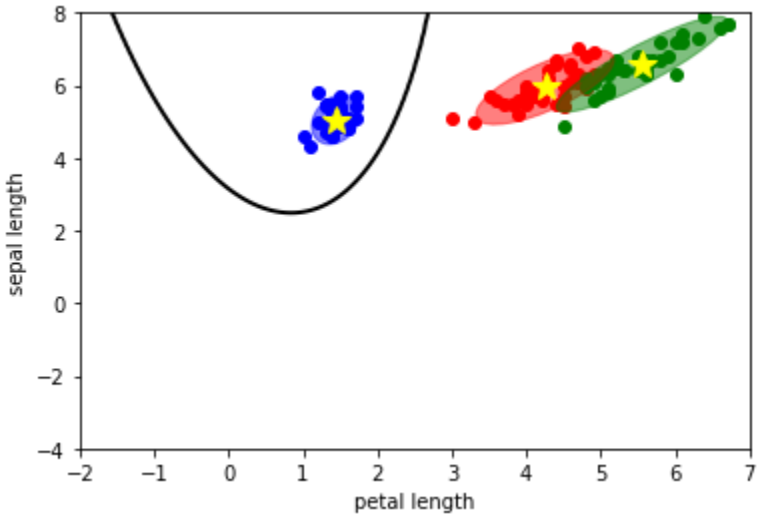
- The three pairs for QD analysis were chosen to be
- `# petal length vs sepal length`
- `# petal length vs sepal width`
- `# petal width vs petal length`
- Models were trained accordingly.

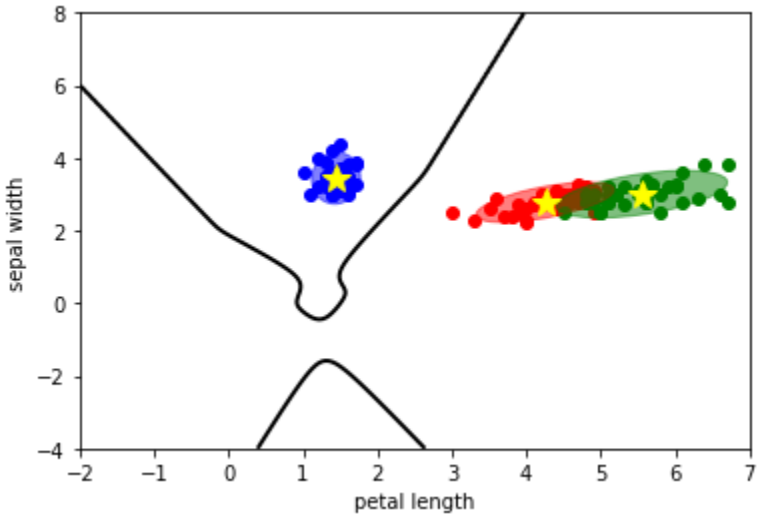
3)

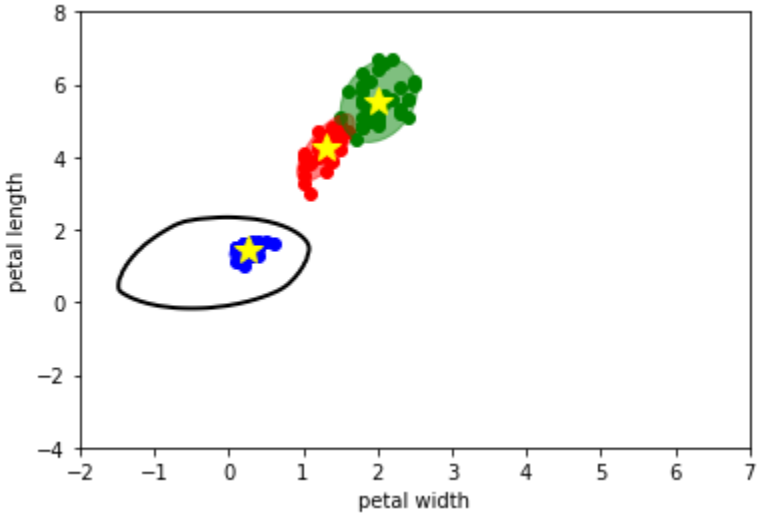
- .covariance_ and .means_ were used to report corresponding features.

4)

- Decision boundary for the QD analysis was plotted for the 3 trained models with the help of contour map:







5)

- Error rate was calculated for each model concerning the 3 selected pairs.

```
petal length vs sepal length

Predictions: [0 2 0 1 1 2 0 1 2 1 2 0 1 1 1 2 1 1 1 0 1 2 2 0 0 1 0
0 2 2 2 0 2 1 2 0 0

 0 0 1 2 0 1 2 1]
```

Error rate: 0.06666666666666665

petal length vs sepal width

Predictions: [0 2 0 1 1 2 0 2 2 1 2 0 1 1 1 2 2 2 1 0 2 2 2 0 0 1 0 0 2 2 2 0 2 1 2 0 0

0 0 1 2 0 1 2 1]

Error rate: 0.06666666666666665

petal width vs petal length

Predictions: [0 2 0 1 1 2 0 1 2 1 1 0 1 1 1 2 2 2 1 0 2 2 2 0 0 1 0 0 2 2 2 0 2 1 2 0 0

0 0 1 1 0 1 2 1]

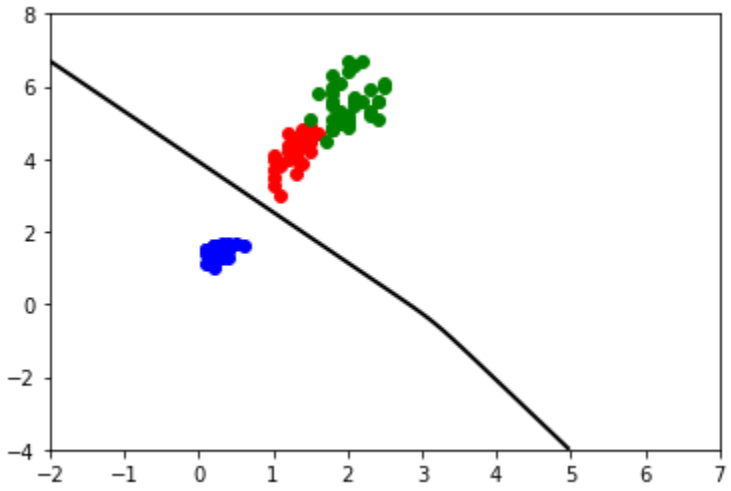
Error rate: 0.0444444444444444

6)

- The pair with best results was Petal length vs Petal width
- LDA model was trained on the same.

7)

- Similar contour technique was used to plot the decision boundary for the trained LDA model:

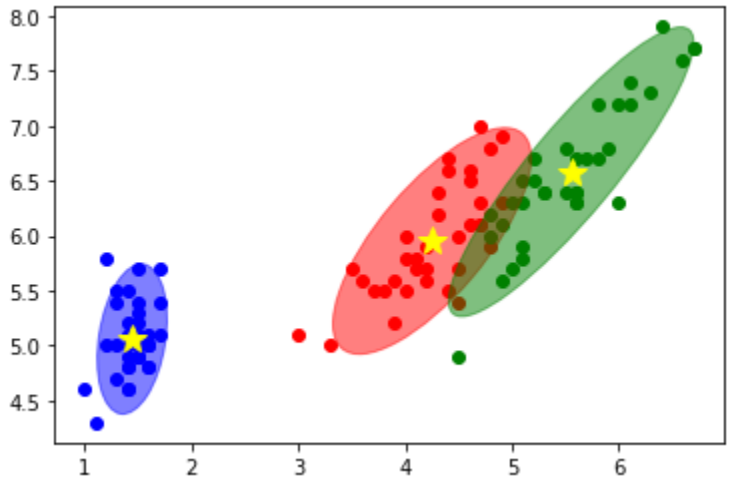


8)

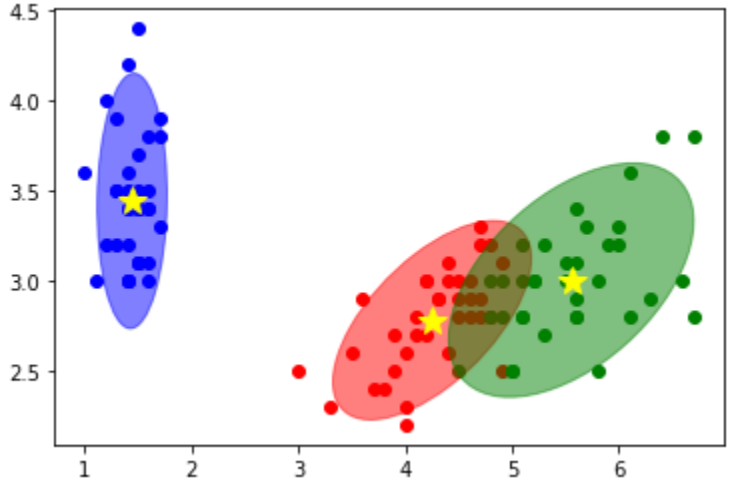
- `Error rate for LDA model: 0.06666666666666665`
- `Error rate for QDA model: 0.0444444444444444`
- It is clear that the QDA model performed better.
- QDA has more flexibility for the covariance matrix than LDA.
- QDA tends to fit data better than LDA.
- QDA assumes a separate covariance matrix for every class and hence provides more parameters.

9)

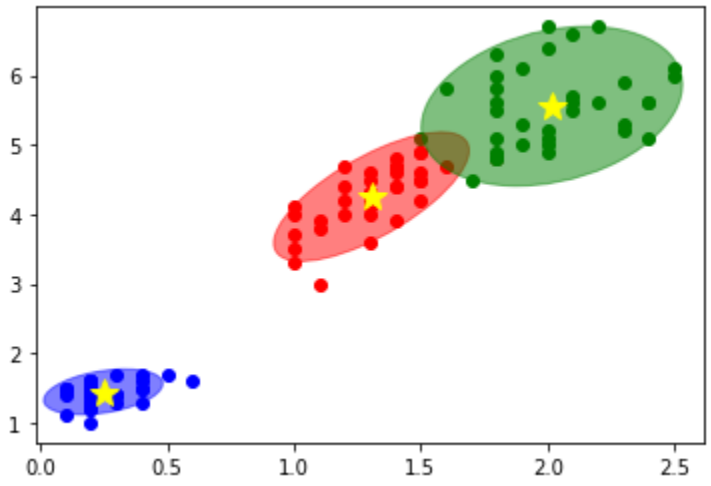
Plots for QDA:



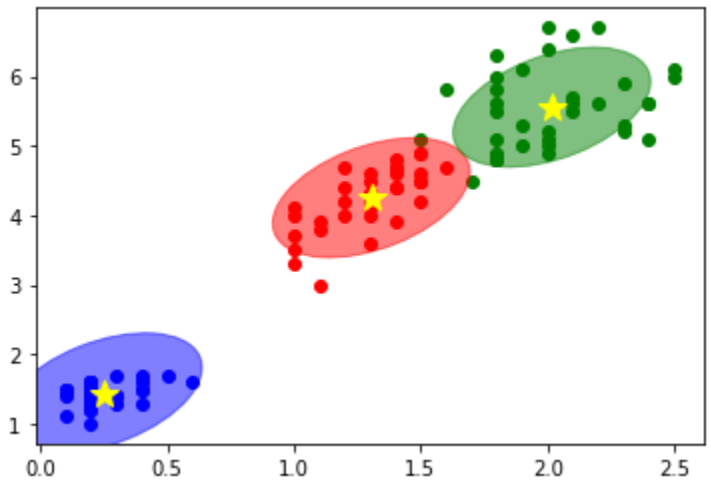
Plots for QDA:



Plots for QDA:



Plot for LDA

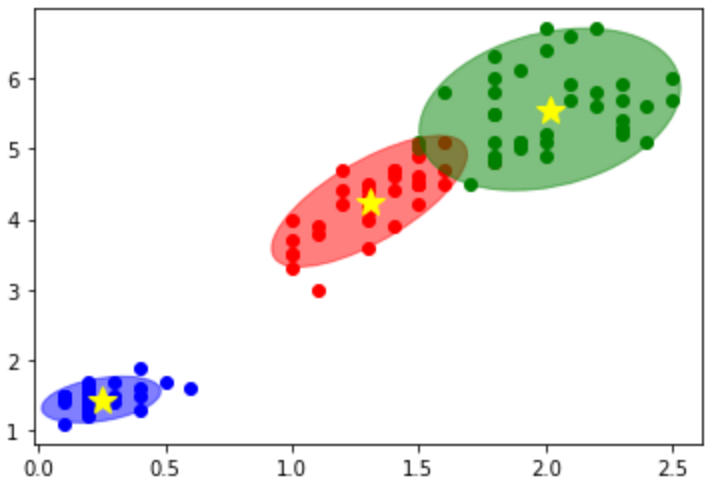


QUESTION 2.

1)

- Sample mean and sample covariance matrix were calculated for each class:

```
List of means: [[1.464, 0.2439999999999999], [4.26,
1.3259999999999998], [5.552, 2.026]]

List of Covariance Matrices:

 [              petal length  petal width

petal length       0.030106     0.005698

petal width        0.005698     0.011494,

                 petal length  petal width

petal length       0.220816     0.073102

petal width        0.073102     0.039106,

                petal length  petal width

petal length       0.304588     0.048824

petal width        0.048824     0.075433]
```



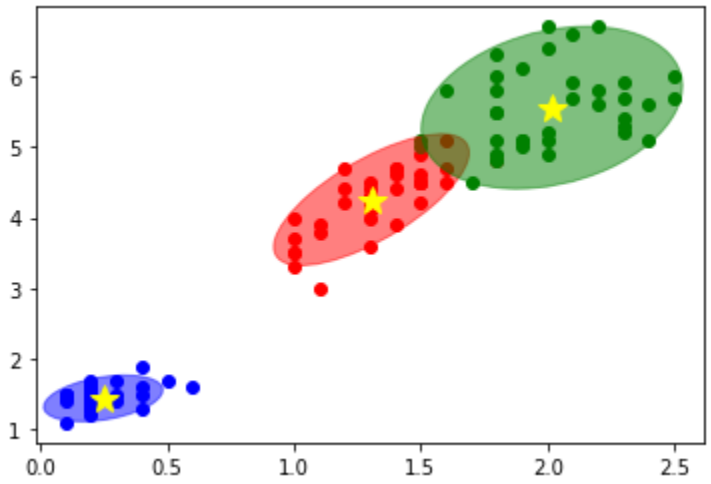
2)

- compute_likelihood function was created to calculate likelihood for given input in accordance with means and covariance matrices obtained.

3)

- Bayes Classifier was made using the training data by performing maximum likelihood estimation.

4)



5)

QDA Accuracy petal width petal length : 100.0 %

Gaussian Likelihood accuracy 91.11111111111111 %

Covariance Matrix of class

0	petal length	petal width
petal length	0.030106	0.005698
petal width	0.005698	0.011494

Means of class

0 [1.464, 0.2439999999999999]

Covariance Matrix of class

1	petal length	petal width
petal length	0.220816	0.073102
petal width	0.073102	0.039106

Means of class

1 [4.26, 1.3259999999999998]

Covariance Matrix of class

2	petal length	petal width
petal length	0.304588	0.048824
petal width	0.048824	0.075433

Means of class

2 [5.552, 2.026]

QUESTION 3.

1)

- Datasets and Label sets were loaded and stored.
- Labels were mapped to docId using mapper with zip and dict.
- The dataset was grouped according to the label of each document.
- Likelihood of the entire train data set was found out.

2)

- Numerous NaN values were observed due to non occurrence of several words.
- This will result in 0 probability in likelihood estimation and hence Laplace Smoothing was applied.
- This assumes all the seen as well as unseen data as one more than the number of times it occurred.

3)

- Naive Bayes Classifier was used through sklearn library.
- The accuracy was found to be `88.44952958752896 %`.