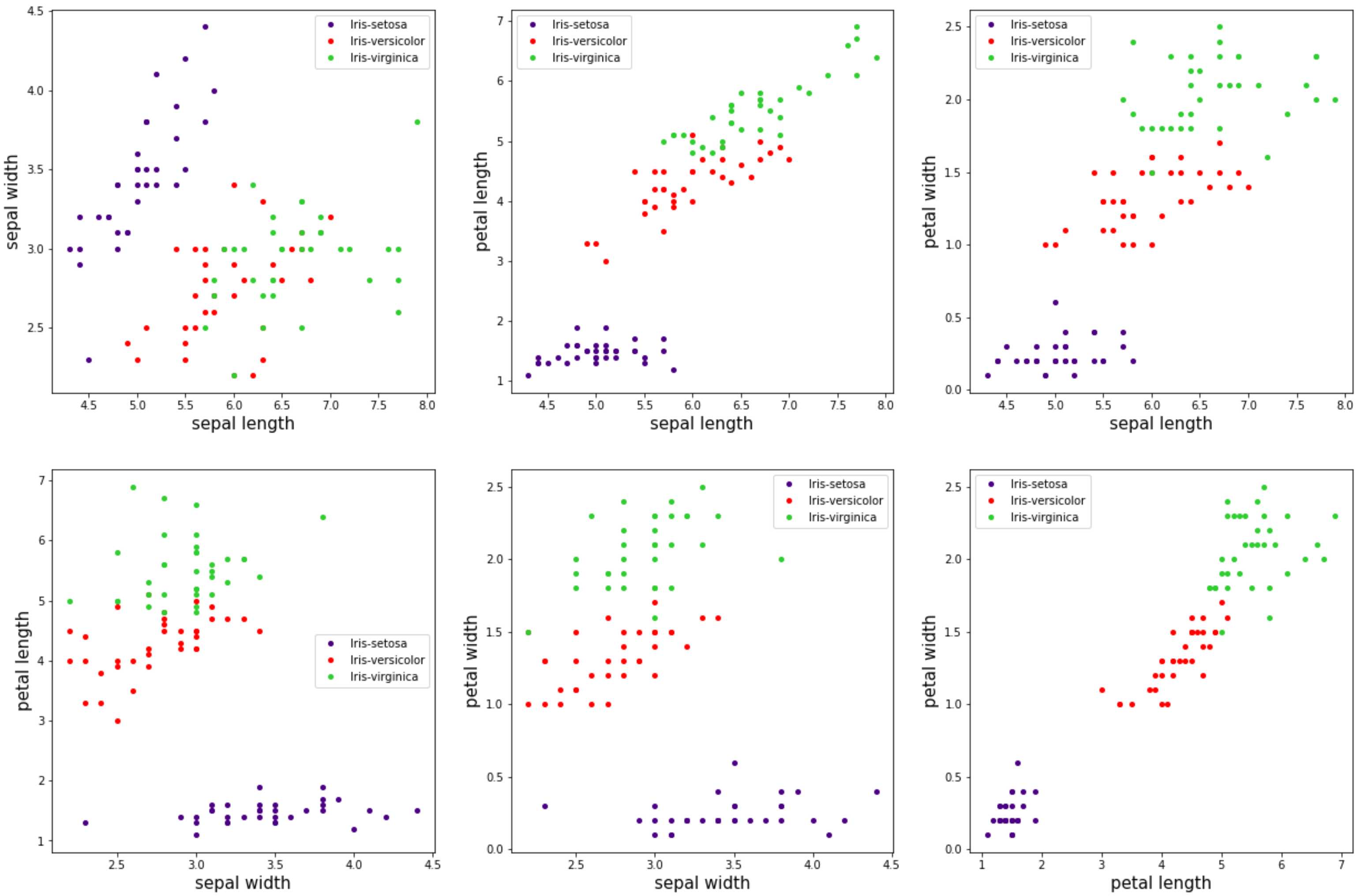
Maniya Yash Rajeshbhai
B20CS033

# ML Lab 5 Report

## Question 1 :

### 1. Preprocessing and Visualisation :

- There were no NULL values or unuseful data. So nothing to drop.
- Then plotted distributions of all the features pair-wise for analysing them.



## 2. Training QDA:

- After visualising the plots, selected 3 pairs which gave the most separation among the classes and used them for training the QDA model.

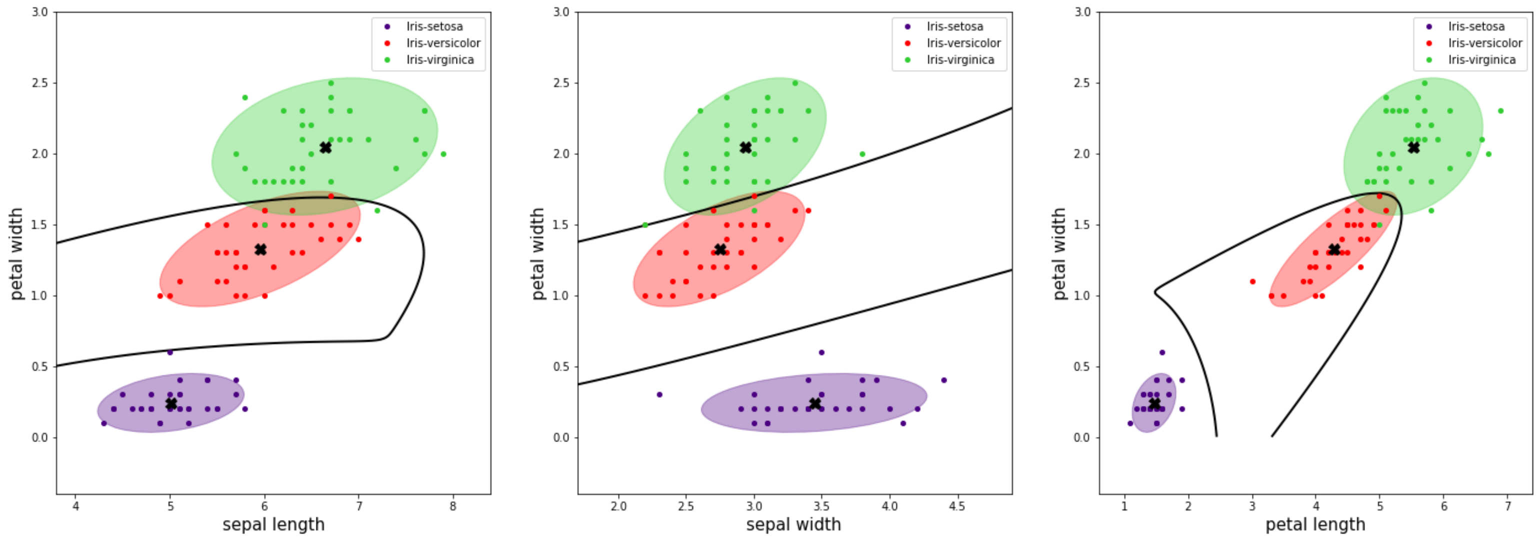
## 3. Mean and Covariance Matrix for QDAs :

- Calculated Mean and Covariance matrix for the selected pairs according to the classes.
- One of the 3 pairs is shown here :

```
For features [sepal length, petal width] :

Mean :

[[5.01428571 0.24      ]
 [5.96       1.32571429]
 [6.64571429 2.04857143]]

Covariance of Distributions :

[[[0.15008403 0.01147059]
  [0.01147059 0.01070588]]

 [[0.28129412 0.06311765]
  [0.06311765 0.04137815]]

 [[0.35667227 0.03359664]
  [0.03359664 0.05845378]]]
--------------------------------------------
```

## 4. Decision Boundary:

- From the obtained means, covariance matrices, I plotted the elliptical distributions and then using the predicted probabilities from QDA models, plotted the decision boundary.




## 5. Class Predictions :

- Using the trained QDA models, predicted the classes for the testing data.
- Found that petal length and petal width pair gives maximum accuracy and minimum error rate. ( err_rate = 1-accuracy_score )

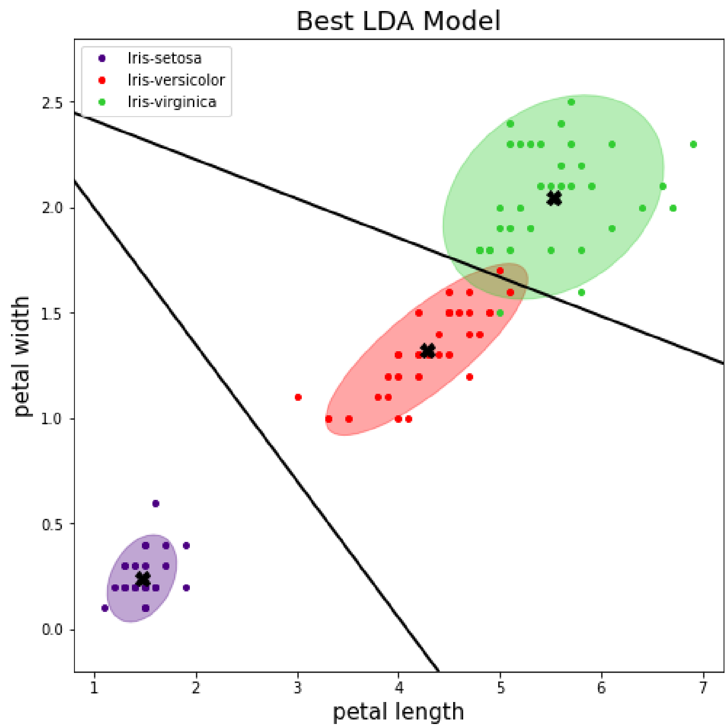
```
We got following Error Rates for the following pairs :

For [sepal length , petal width] : 0.0667
For [sepal width , petal width] : 0.0667
For [petal length , petal width] : 0.0444

Hence the best model is obtained from the pair [petal width , petal length]
Minimum Error Rate : 0.0444
Maximum Accuracy : 95.56 %
```

## 6. LDA model :

- Trained a LDA model for the best pair using training data.

## 7. Decision boundary by LDA :

- Predicted probabilities by the LDA and used them for plotting the decision boundaries for all the classes.

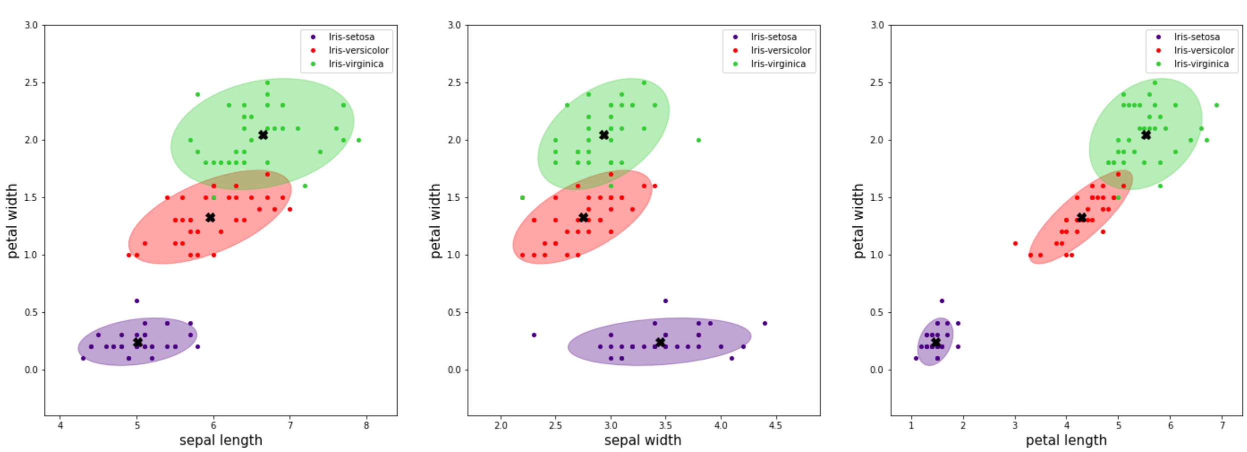



## 8. Error Rate for LDA :

- Predicted the classes of the testing data and calculated error rates of the same. Found that LDA performs worse than QDA as expected. ( LDA underfits the given dataset )

```
For best LDA Model, following are the metrics :
Error Rate : 0.0889
Accuracy score : 91.11 %
```

## 9. Visualisation of Distributions :

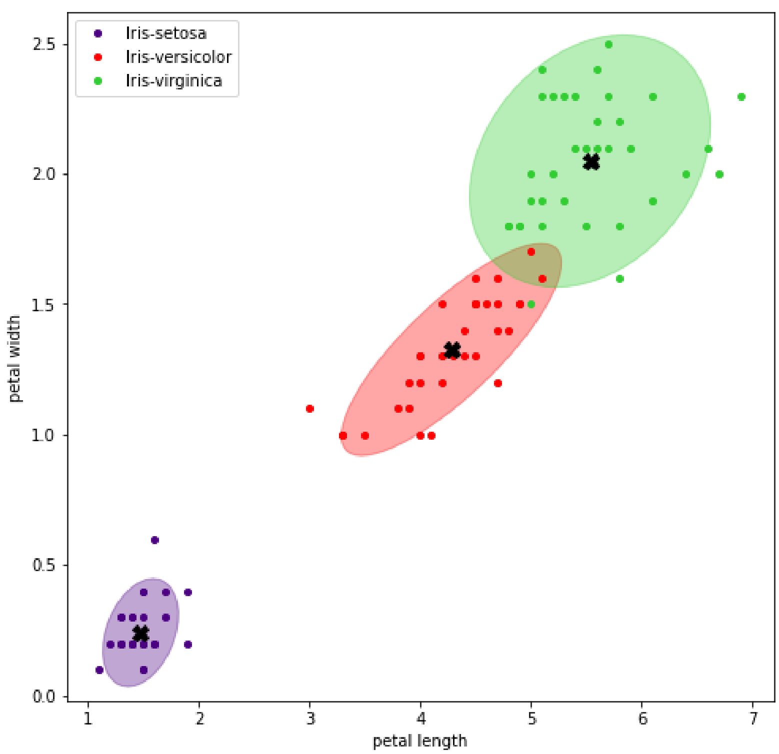


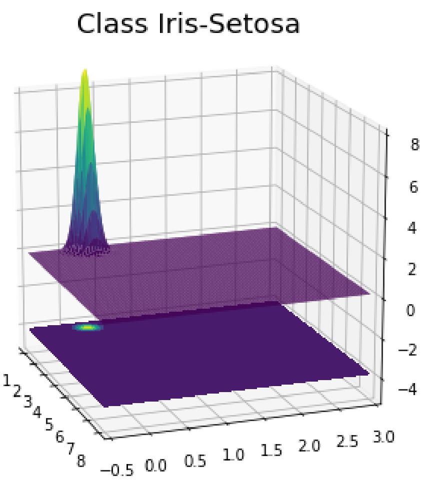
# Question 2 :

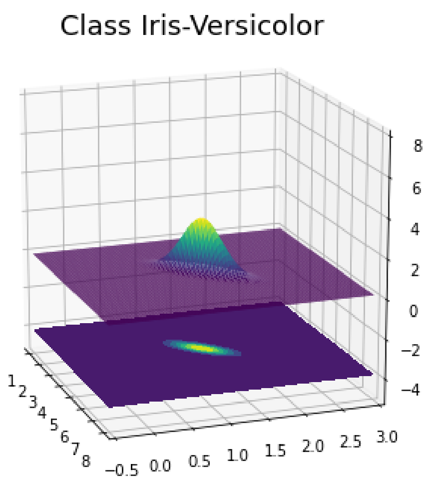
## A. Visualising Gaussian Distributions [ petal length, petal width ] :

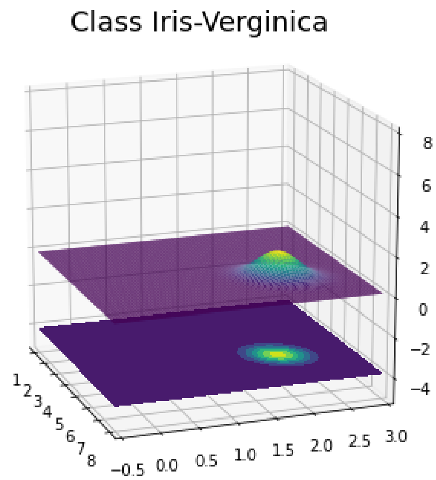
- Calculated Mean and Covariance Matrices and used them to plot ellipses for the class distributions.
- Then plotted gaussian distribution for all the classes in 3 dimensions for better visualisation.

```
Sample Mean for Petal Length and Petal Width :
[[1.47142857 0.24      ]
 [4.28       1.32571429]
 [5.53142857 2.04857143]]

Sample Covariance Matrix for Petal Length and Petal Width :
[[[0.02915966 0.00588235]
  [0.00588235 0.01070588]]

 [[0.24752941 0.08258824]
  [0.08258824 0.04137815]]

 [[0.2957479  0.03695798]
  [0.03695798 0.05845378]]]
```






## B. Compute Likelihood :

- Made a function compute_likelihood() for calculating the likelihoods keeping multivariate distribution into account.

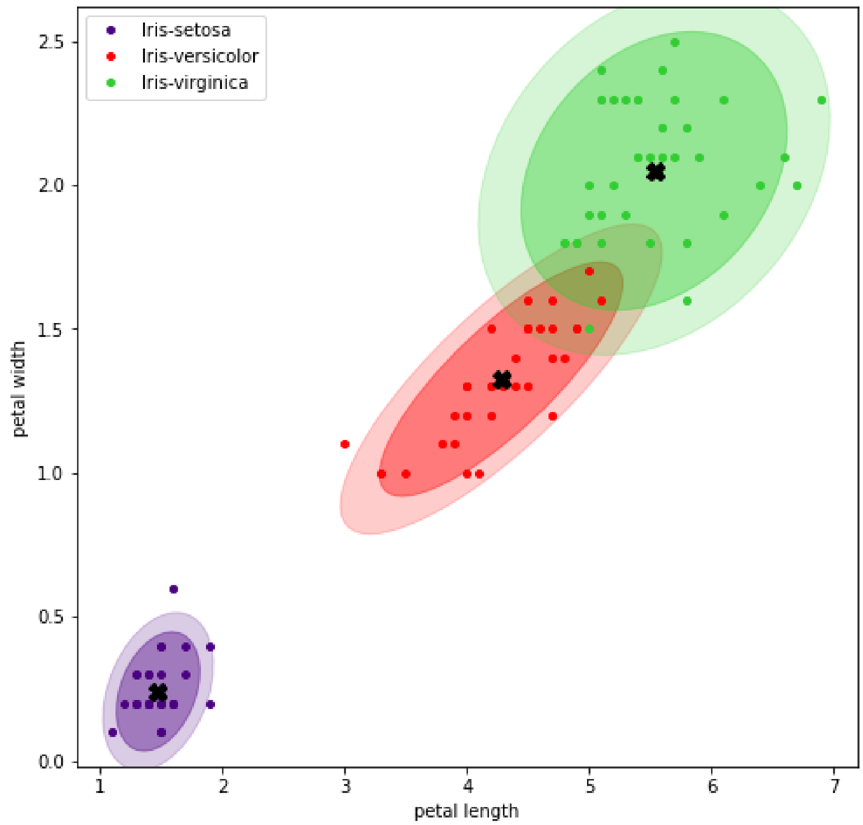
$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

## C. Maximum Likelihood Estimation :

Performed MLE by calculating the class-wise mean and covariance matrix and using them while calculating the likelihoods.

## D. Visualising the Distributions of ( A and C ):

From mean and covariances obtained in part A and C, plotted ellipses for all the classes in the same plot as said. The ellipse represents the contour of the distribution at a certain height.




## E. Making Predictions :

Predicted the classes for the testing data points and compared the accuracy score with QDA's accuracy.

We find that QDA performs better than the Naive Bayesian model.

```
Accuracy obtained by Naive Bayes Classifier :
Accuracy : 91.11 %

Accuracy obtained by QDA :
Accuracy : 95.56 %
```

## Question 3 :

1. **Likelihood Calculation :**
    - Calculated the likelihoods for all the data points in the training dataset.

2. **Laplace Smoothing :**
    - For the probability to remain non-zero for the certain dataset entries, we use laplace smoothing to make the probability a small positive value.

$$p_\lambda(C_k) = p_\lambda(Y = C_k) = \frac{\sum_{i=1}^{N} I(y_i = C_k) + \lambda}{N + K\lambda}$$

3. **Naive Bayes Classification :**
    - Classified the testing data to the corresponding classes and measured the accuracy scores.
    - Also compared with the QDA accuracy, and we see that QDA performs better than the Naive Bayes Classifier.

```
Accuracy obtained by Naive Bayes Classifier :
Accuracy : 91.67 %

Accuracy obtained by QDA :
Accuracy : 95.56 %
```

*End of the Report !*