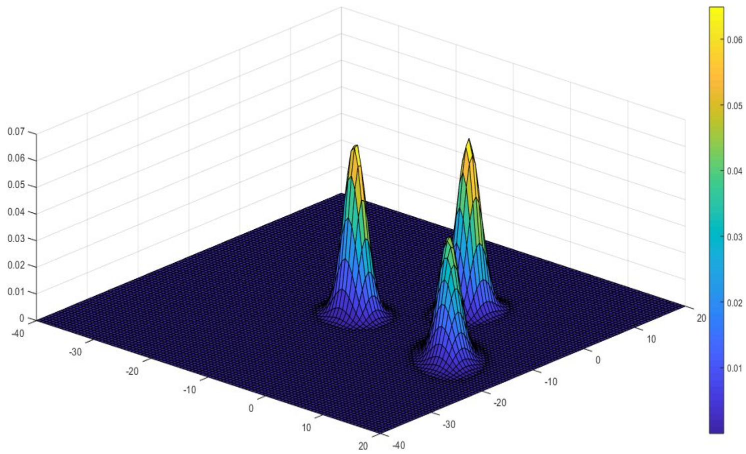
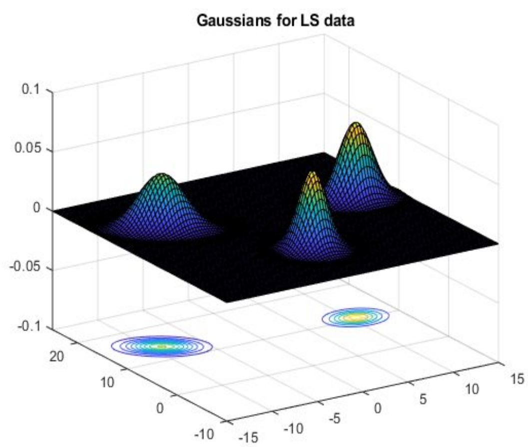
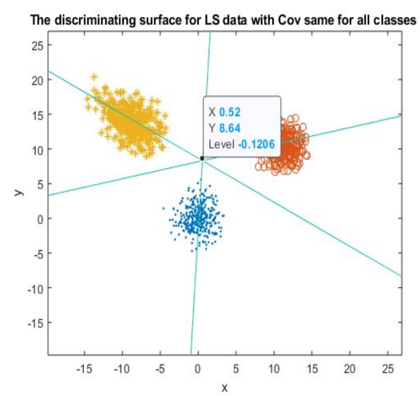
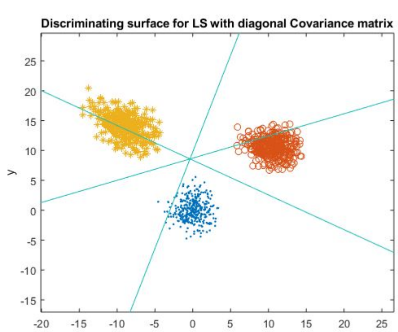
TEAM 14
ASSIGNMENT 2A

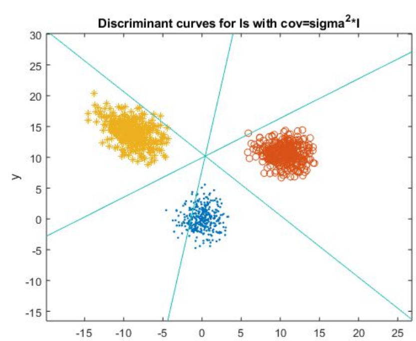
1.LINEARLY SEPERABLE DATA







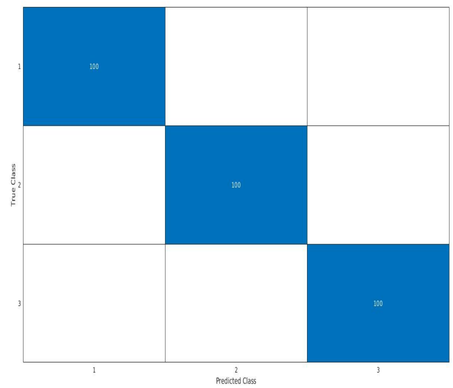


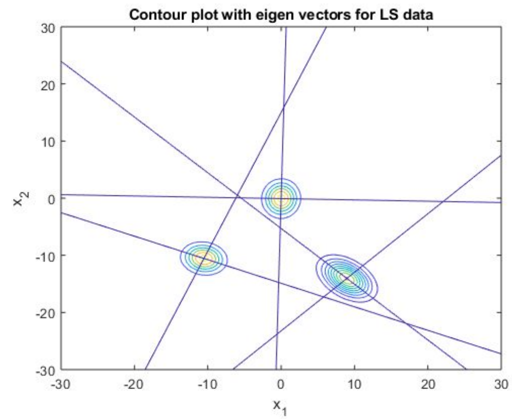


OBSERVATIONS: The given data for linearly seperable type is classified even when covariance matrix is assumed to be variance multiplied with identity matrix so the roc curve for all the caes is perfect so not mentioned here

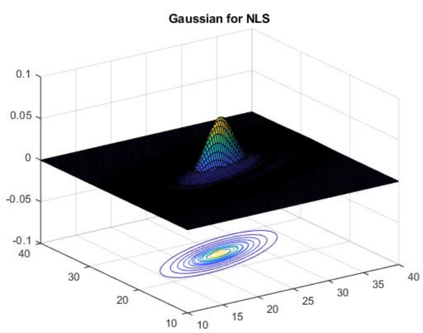
The given development data is classified 100% as shown in below confusion matrix.

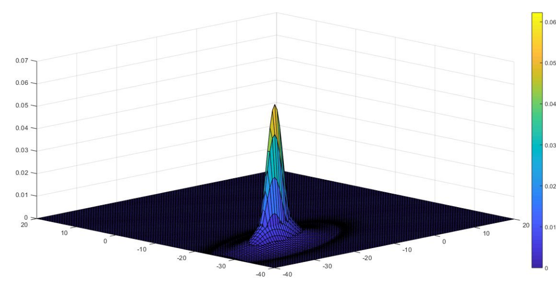
TEAM 14
ASSIGNMENT 2A

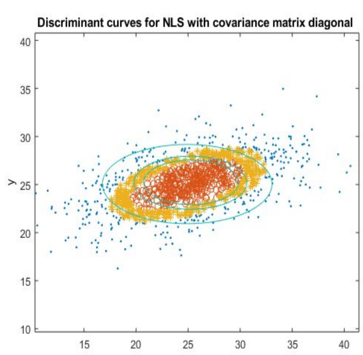


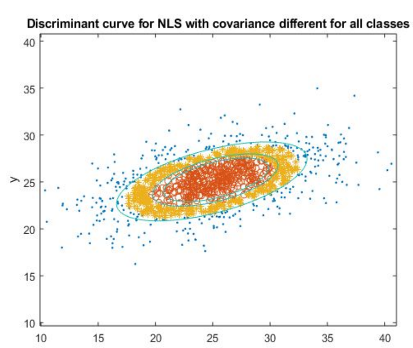


2.NON LINERALY SEPERABLE DATA



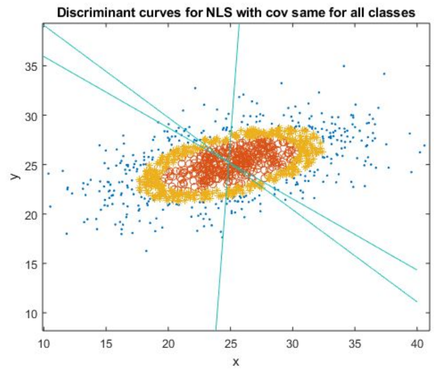


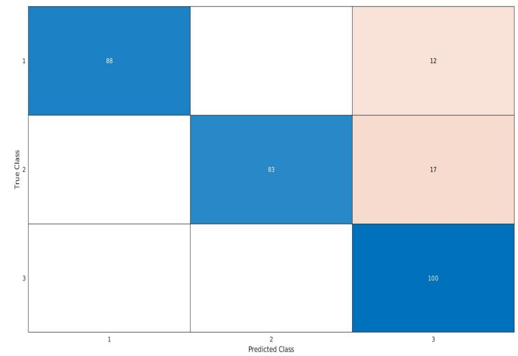


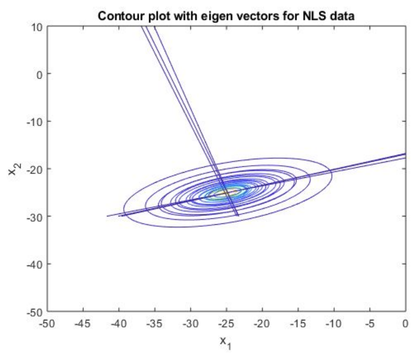


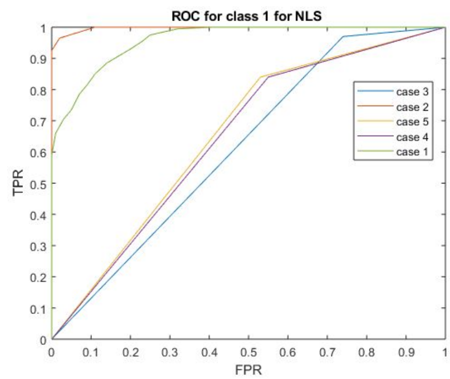
Observations:The NLS data shows the difference between covariance matrix selection in contrast to LS data .Linear discriminants can’t seperate NLS data.The non linear discriminant functions of case2 perform well when compared to other cases.
NLS confusion matrix shows that only one class has wrongly classified points.

TEAM 14
ASSIGNMENT 2A

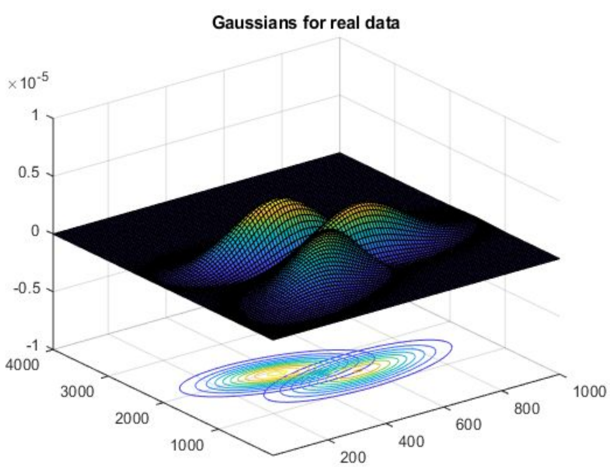


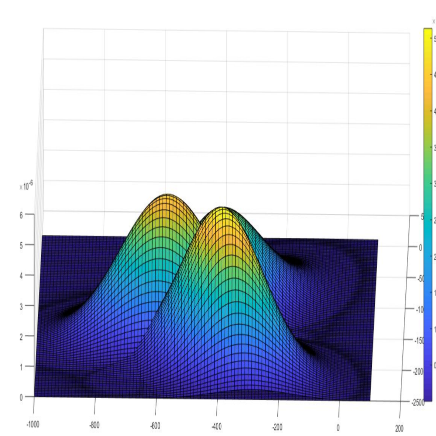




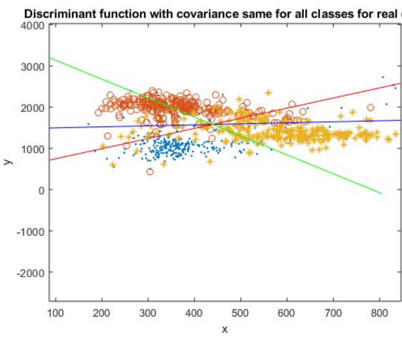


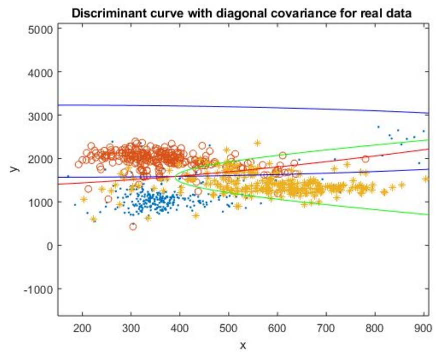
3.REAL DATA

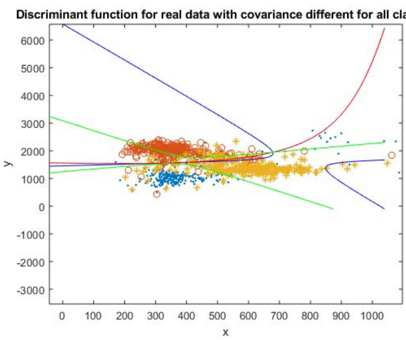


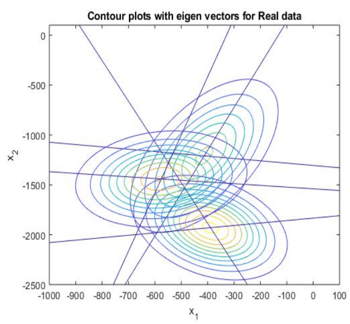


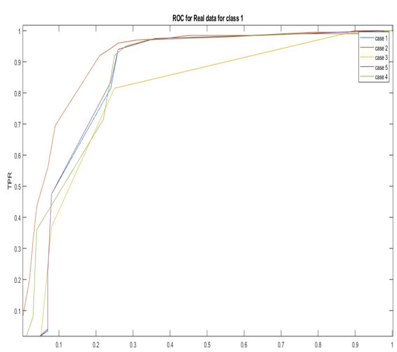
TEAM 14
ASSIGNMENT 2A

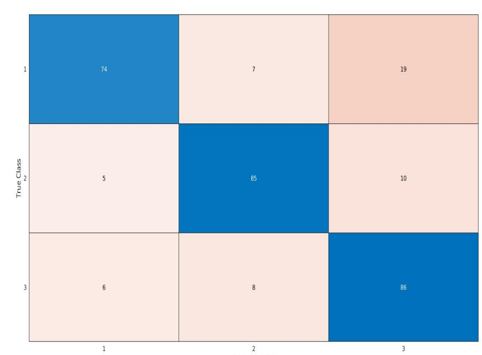












Observations:
The Difference between 5 classes is seen distinctively in real data.
The confusion matrix shows that the classification is not 100 percent in any of the classes.
The roc curve shows that the case two ie non-diagonal different covariance gives you the best classifier when compared to other cases.
From the eigenvector graph we know that the data is distributed in that way
Linear perform midway when compared to nonlinearly seperable data

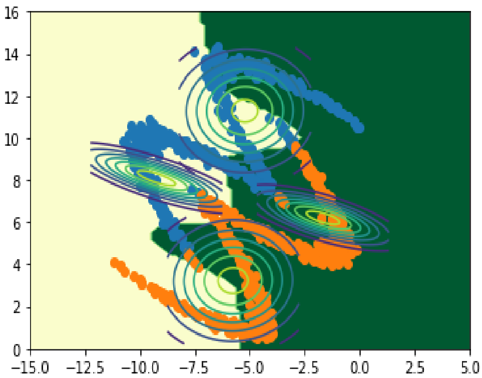
Assignment 2B

1. Synthetic data

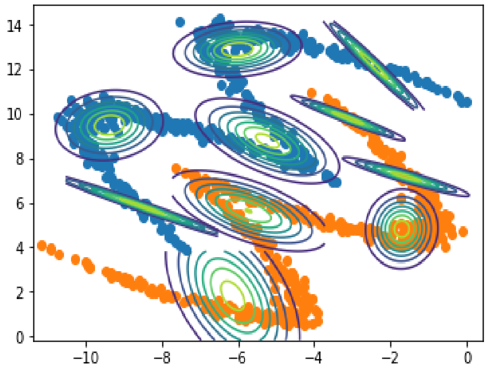
The given data is classified into the Gaussain mixture by the probability of clusters within the mixture .Gaussain mixture model is a generative algorithm.

The algorithm is based on EM maximization technique.The initialization to the algorithm is using the output of kmeans.Then there are series of EM steps going on to maximize likelihood

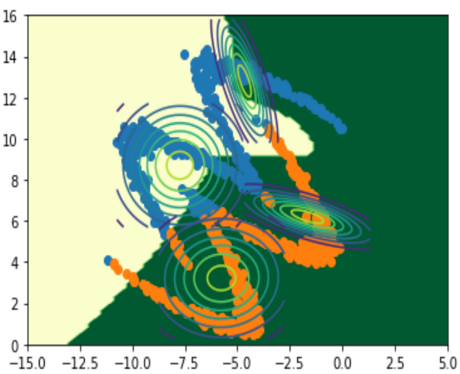
$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$



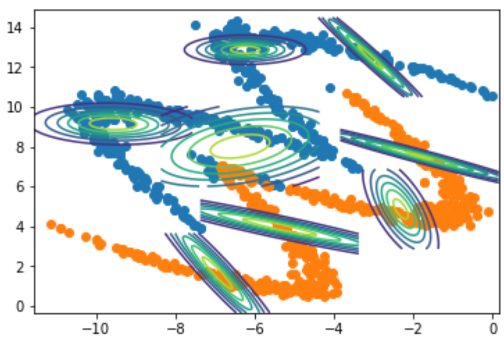
K=2 DIAGONAL



K=5 NON-DIAGONAL

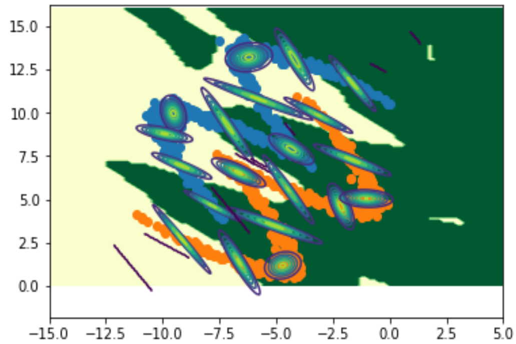


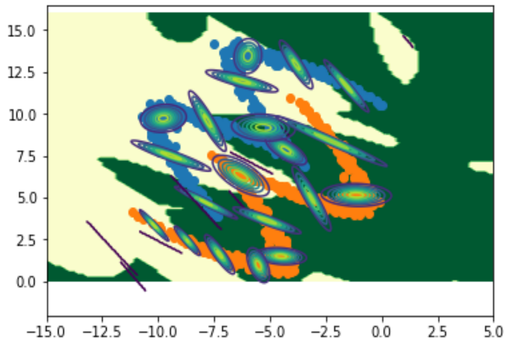
K=2 NON-DIAGONAL



K=4 NON-DIAGONAL

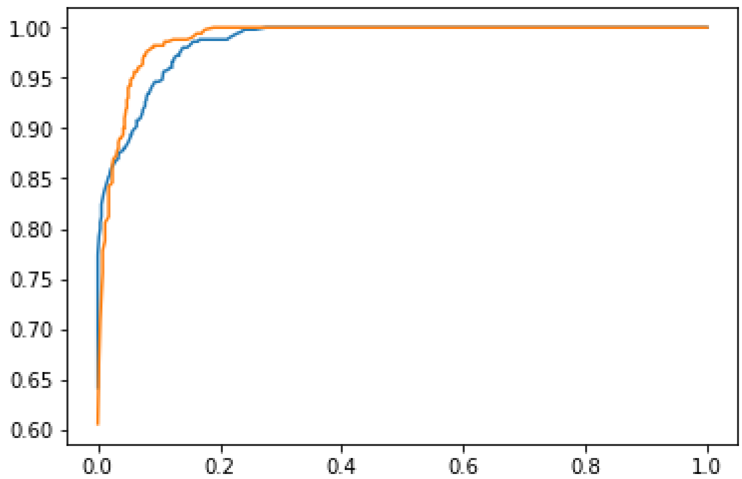
In the E step we caluculate gammas which are corresponding probabilities of data points



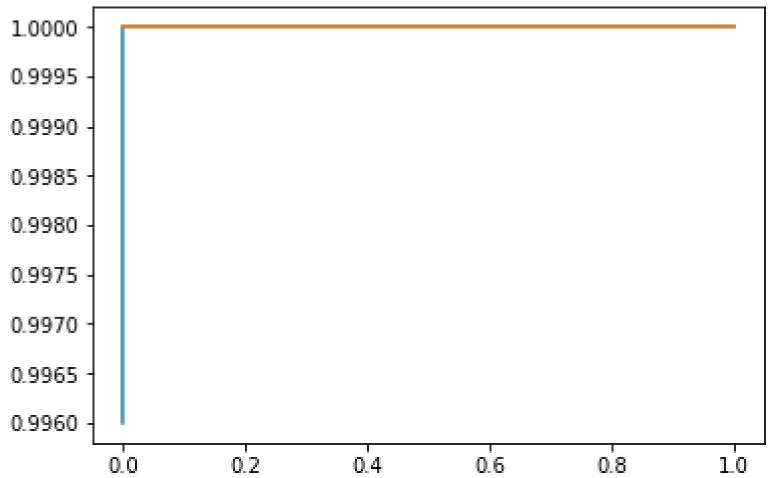


K=10 non diagonal case **k=10 diagonal case**

K values	Non diagonal matrix	Diagonal matrix
2	943	920
5	994	999
8	1000	1000
10	1000	1000



ROC PLOT FOR K=2 DIAGONAL VS NON-DIAGONAL



ROC CURVE FOR K=10 DIAG VS NON-DIAGONAL

Observations

->As you can see above that diagonal matrix has not covered more development points as compared to non diagonal matrix .

Did Gmm with k values 2 and 10 for the plots.as the k values.

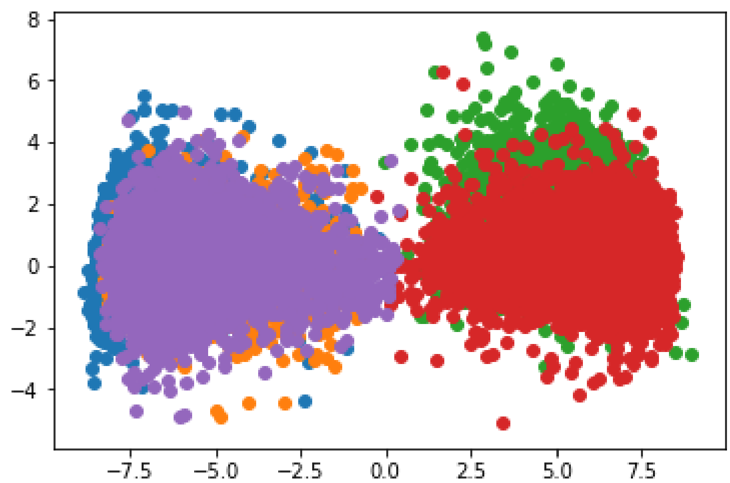
As the k values increase the accuracy increases .The covariance matrix of data is changed in my initialize gmm .

Function to diagonal whenever necessary.

The accuracy for the diagonal matrices is low as compared to non diagonal.As we don't consider the relation between off diagonal features is corelation is not taken into account for gmm .

2.IMAGE DATA

VISUALIZATION



OBSERVATIONS IN VISUALIZATION

The given datasets forest ,mountain,street,insidecity,highway The given datasets are dimensionally reduced from 23 -dimensons to 2 -dimensions and shown in above graph

For k=2 the correctly classified points in forest and mountain is 3261 where the percentage is too less than 50%