ASSIGNMENT - 1

GROUP 5

Premnath N **CS14B022**

R Hemanth Reddy **CS14B024**

M Uday Theja **CS14B044**

K Nearest Neighbour Classifier:

Non-parametric method :

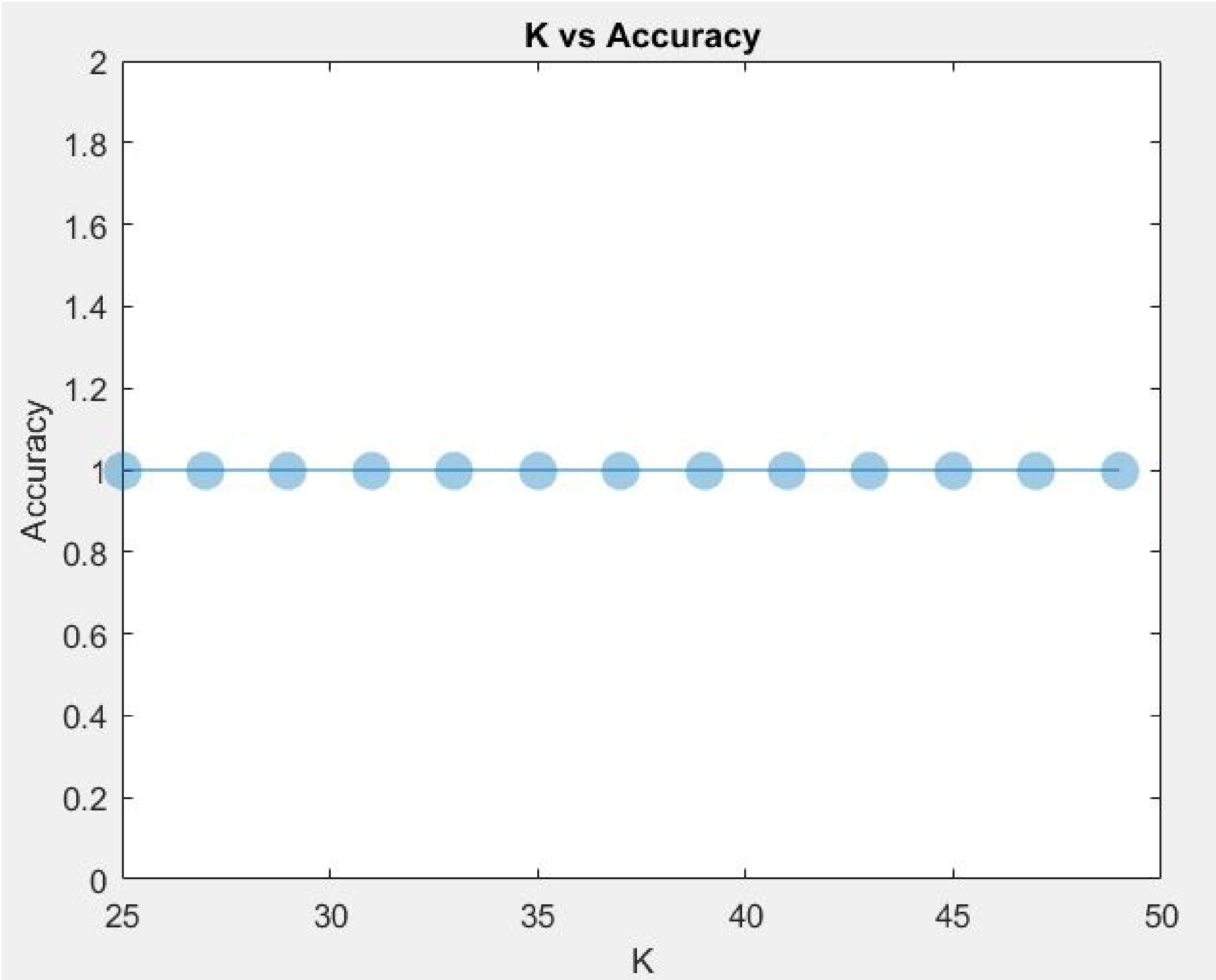
- Weight function W(x)=1/d where
  d=distance of the neighbour to the test point(to be classified)

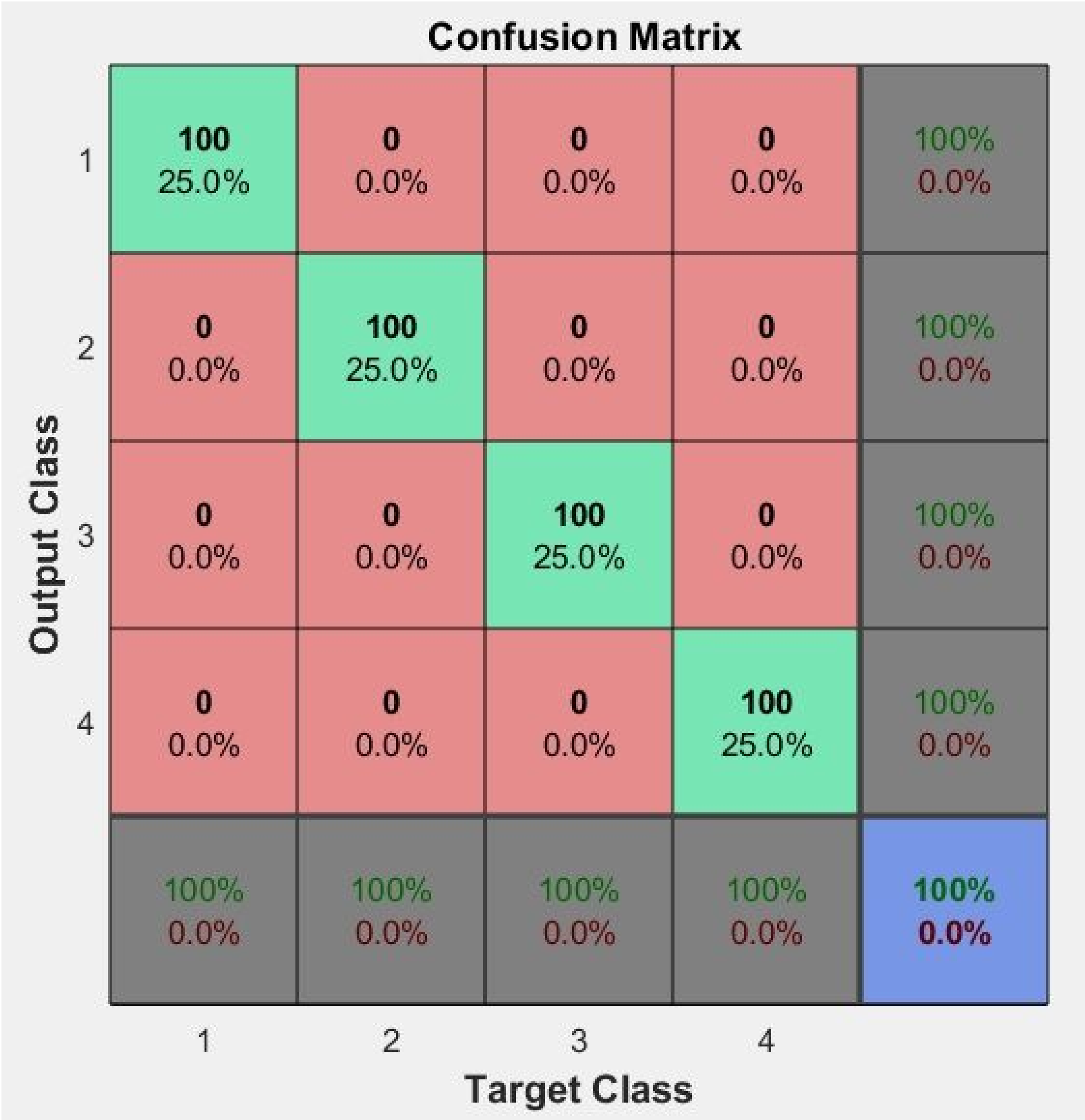
- Hyperparameter Optimization(On validation data) :
  Value of k manually varied from k=25 to 50 (considering training data set to be of 1000 points).Optimum value of k occurred around square root of 100 i.e k=38.

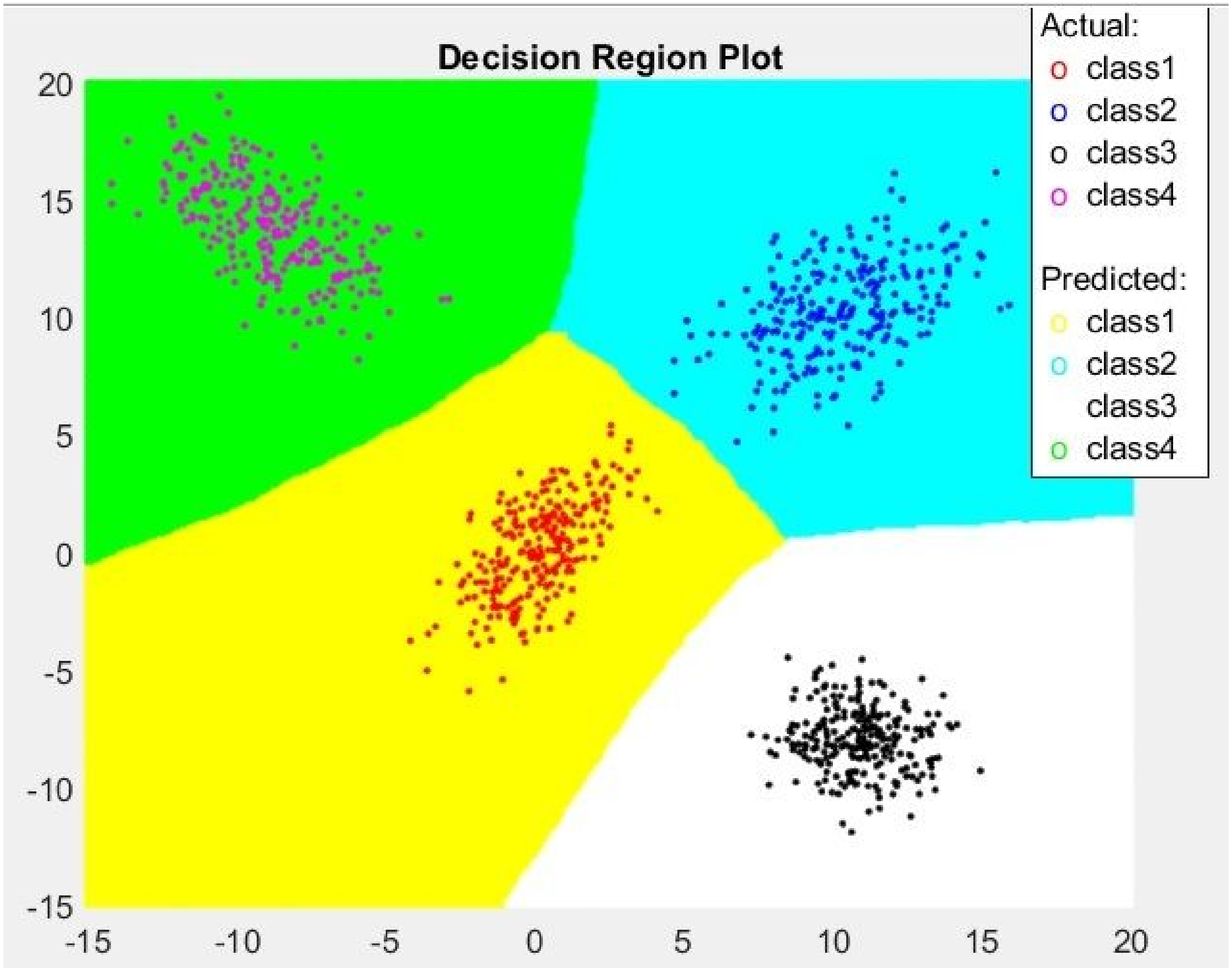
- Performance on data sets:

Linearly Separable data:

- Decision surface plots separates the clusters with 100 percent accuracy.

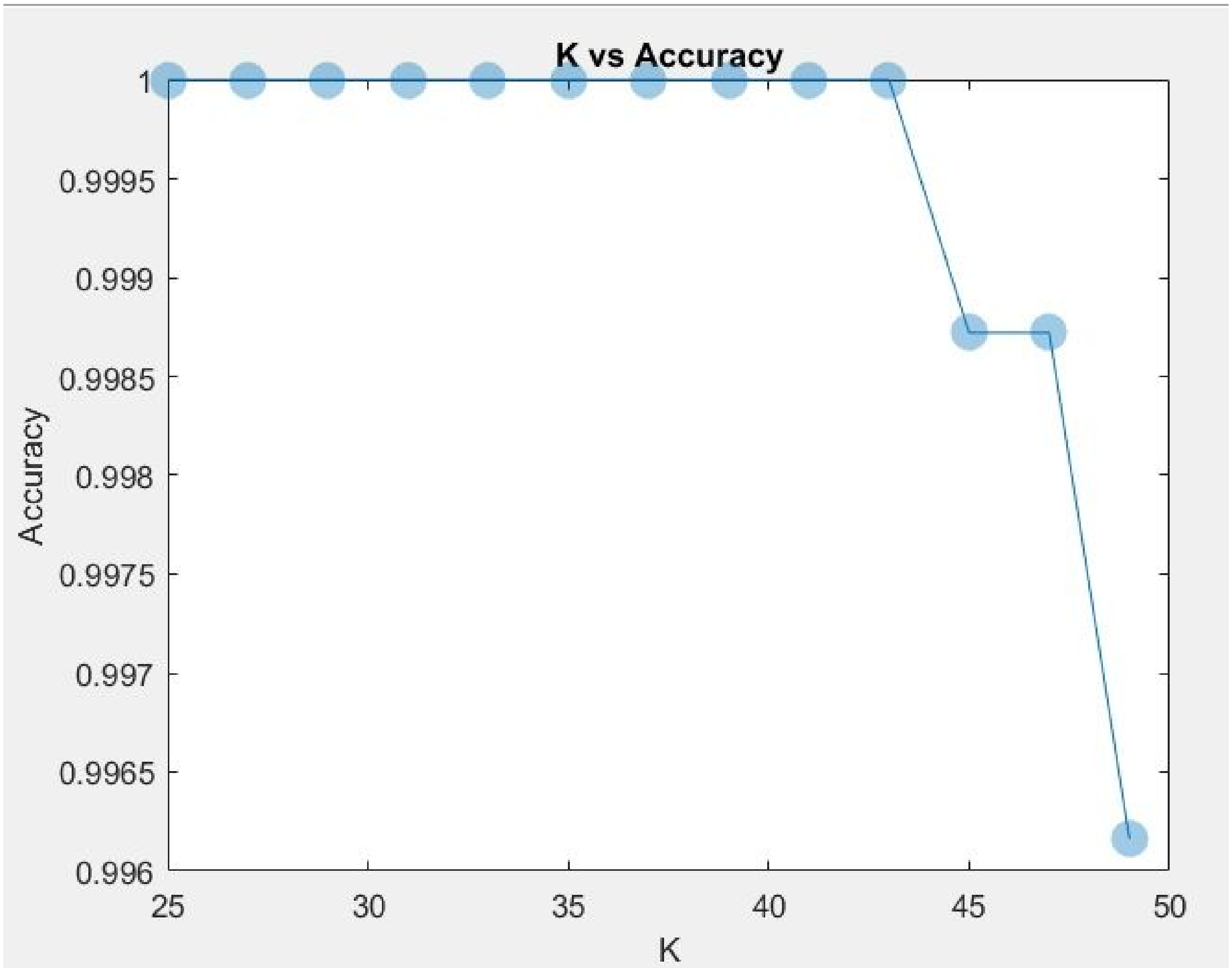

K vs Accuracy
Accuracy
K
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
25
30
35
40
45
50

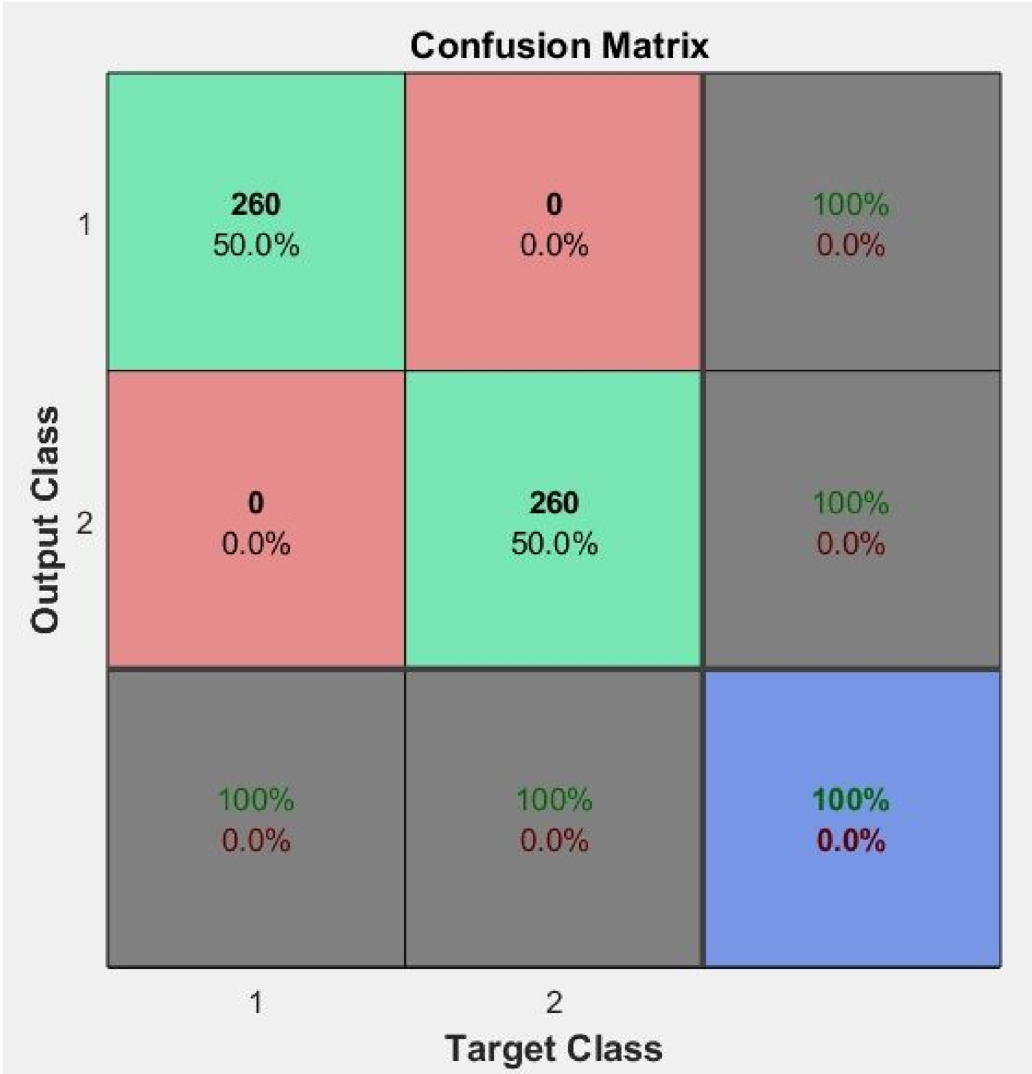

Confusion Matrix
Output Class
Target Class
1
2
3
4
100
25.0%
0
0.0%
100%
0.0%

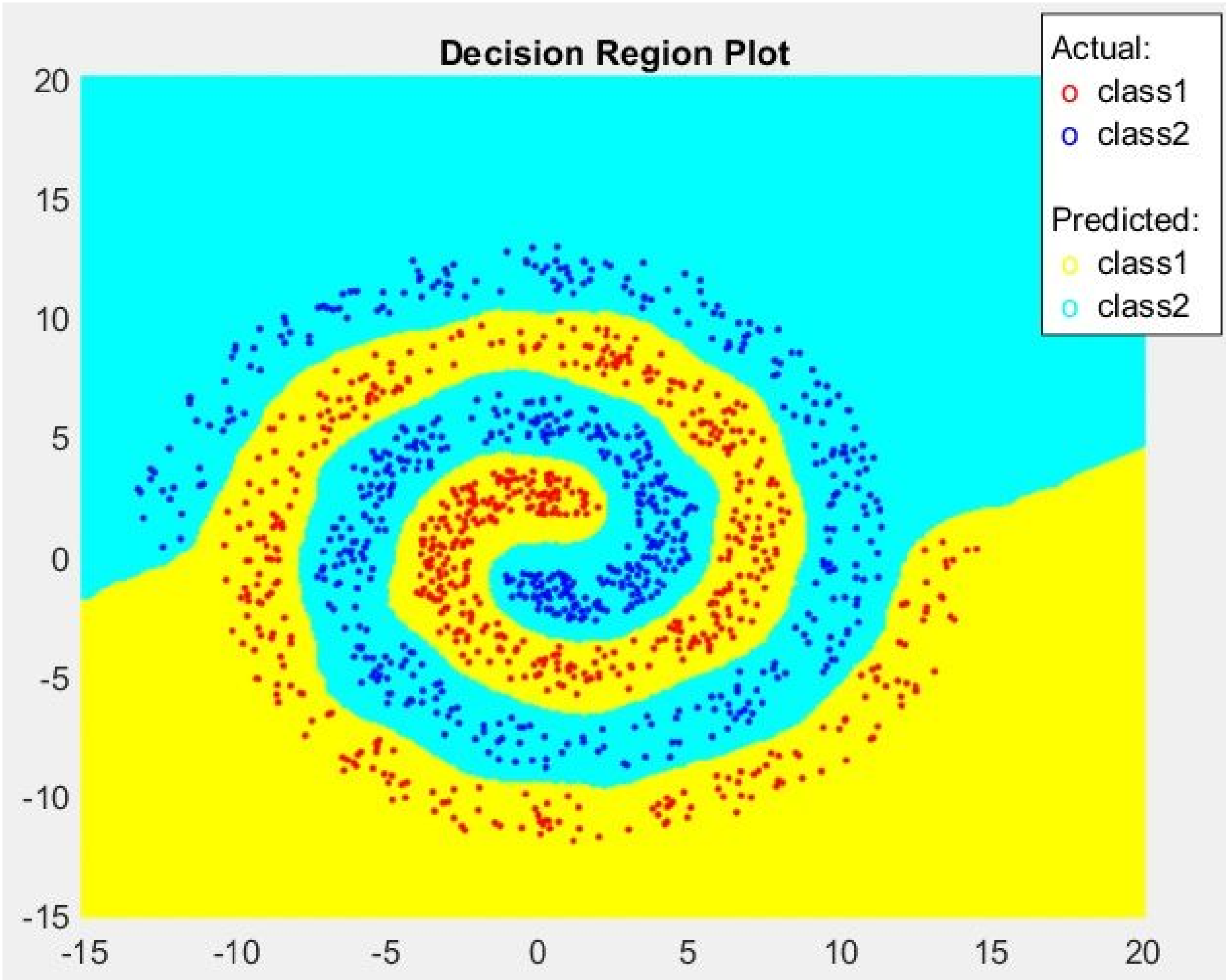


Non-Linearly Separable data:

- Decision surface plots takes the form of training data distribution with 100 percent accuracy.

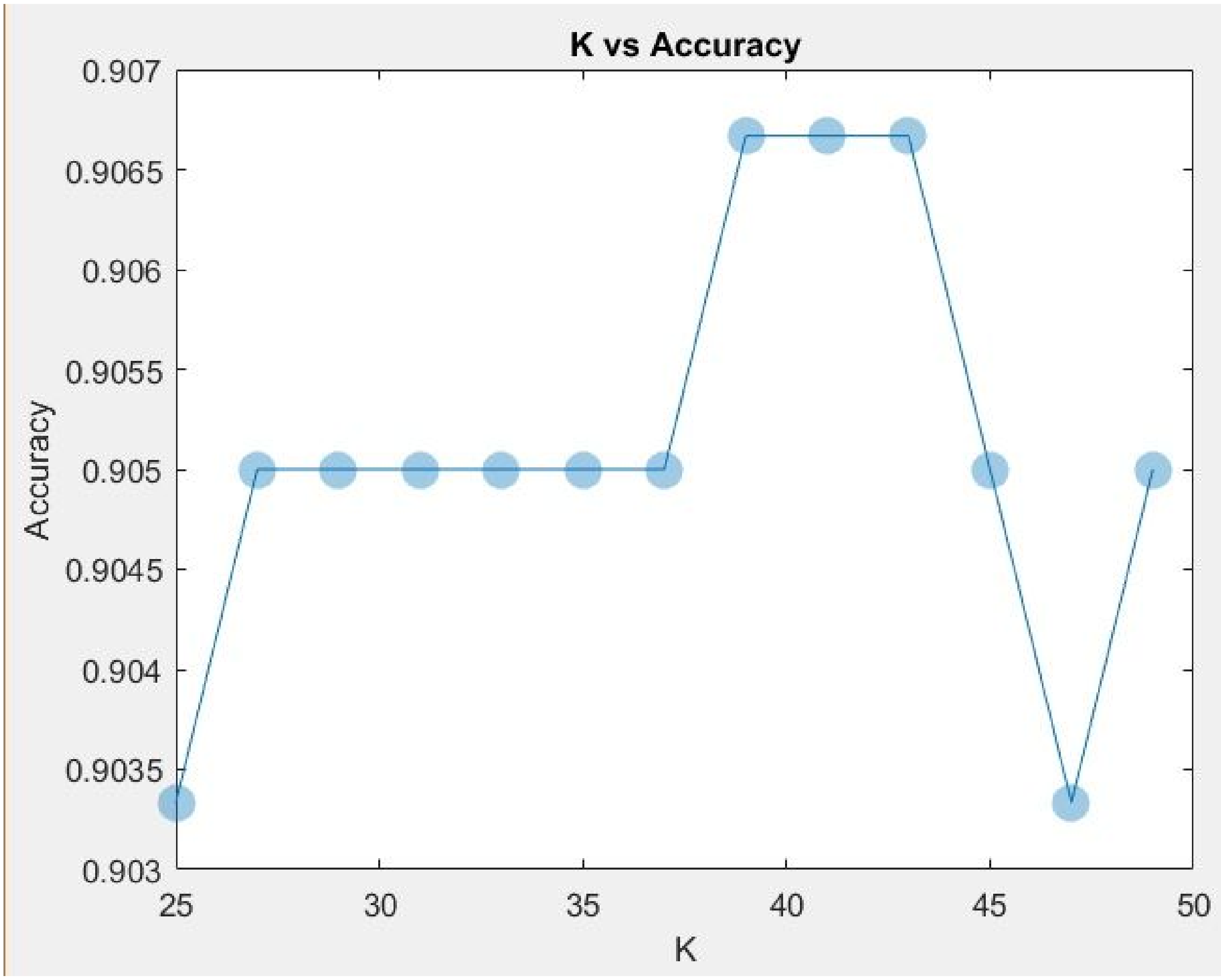

K vs Accuracy
Accuracy
K
1
0.9995
0.999
0.9985
0.998
0.9975
0.997
0.9965
0.996
25
30
35
40
45
50

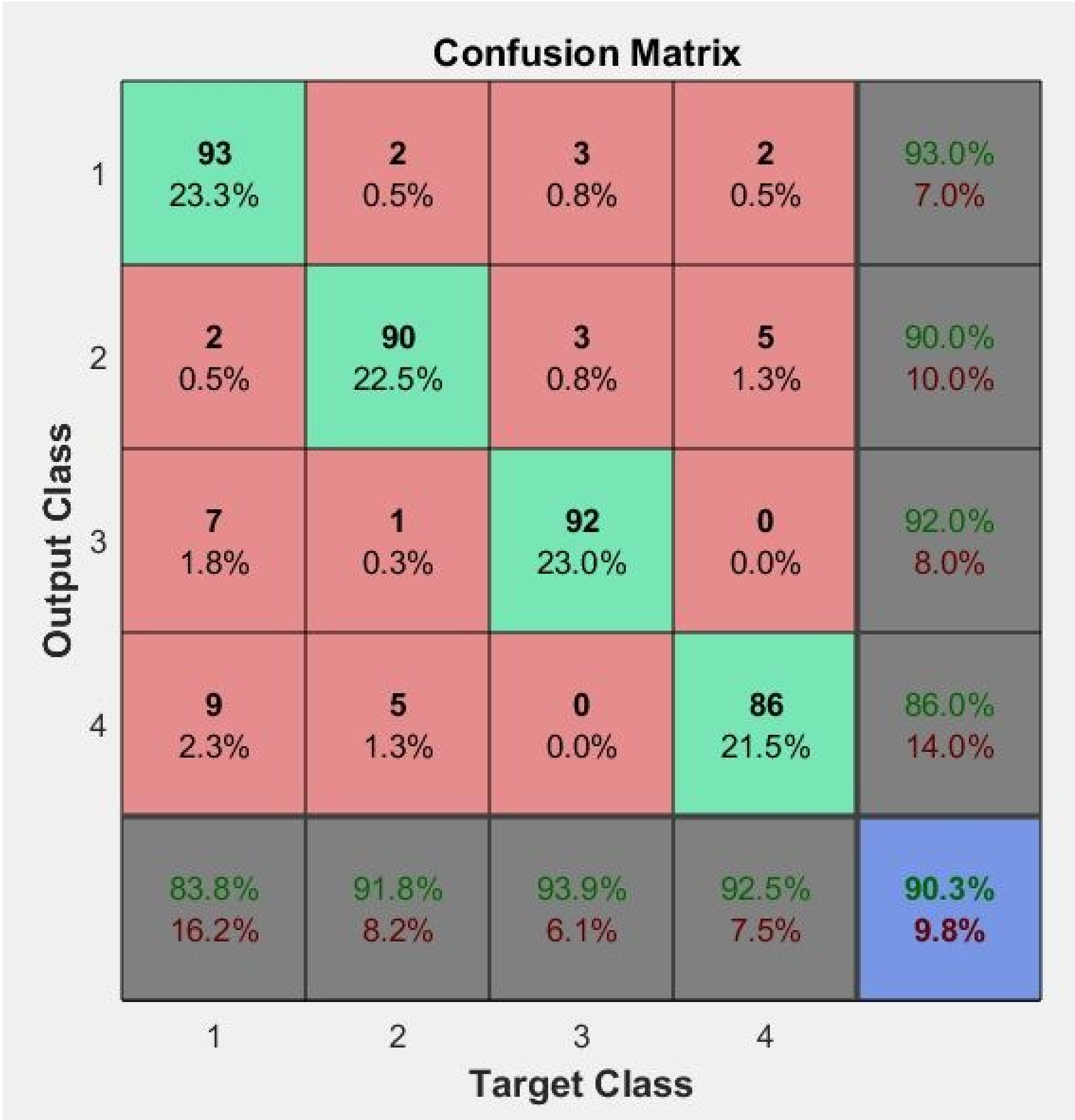

Confusion Matrix
260
50.0%
0
0.0%
100%
0.0%
0
0.0%
260
50.0%
100%
0.0%
100%
0.0%
100%
0.0%
100%
0.0%
Output Class
1
2
1
2
Target Class



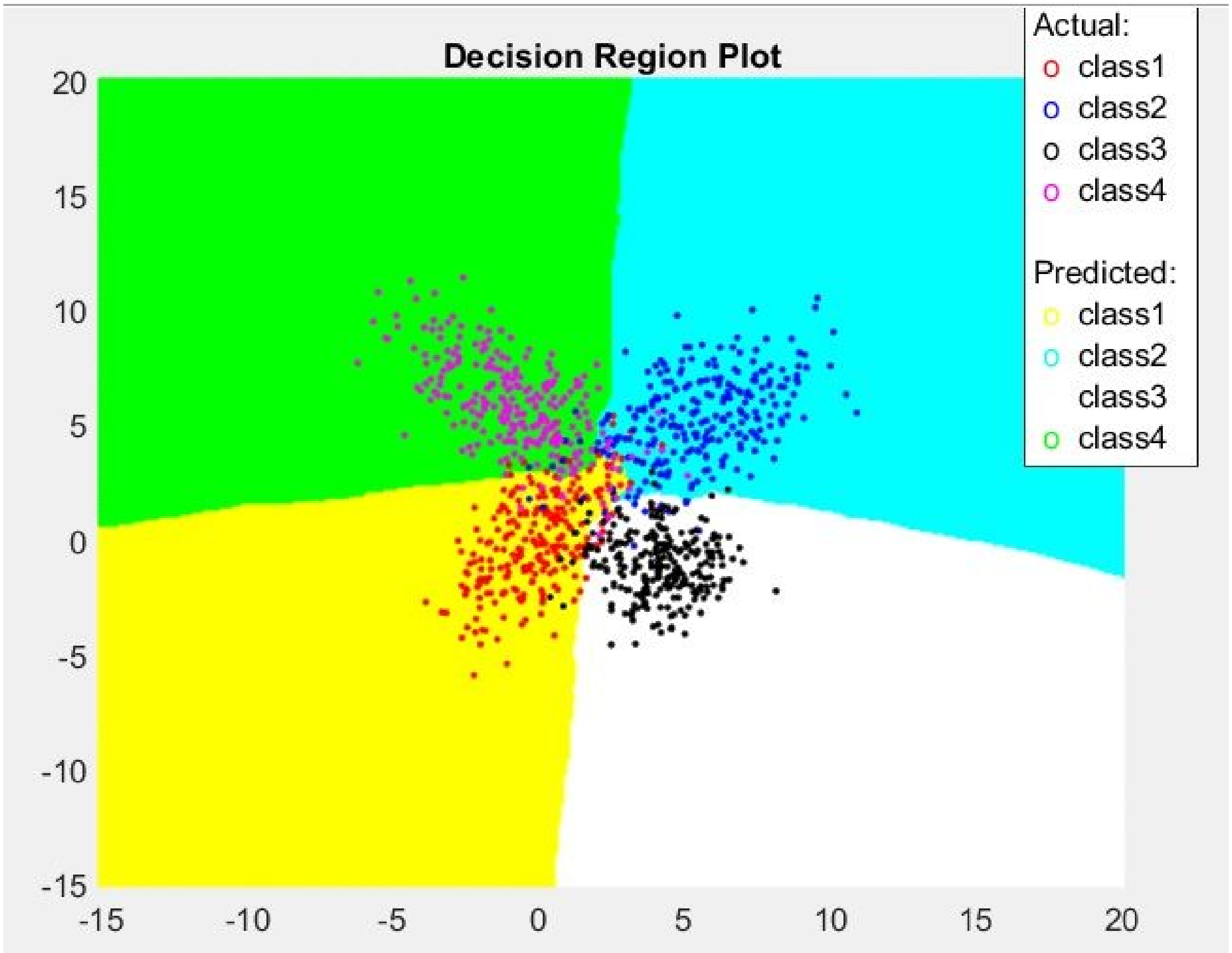
Overlapping data:

- The classification of data is more accurate away from the centre of distribution.


K vs Accuracy
0.907
0.9065
0.906
0.9055
0.905
0.9045
0.904
0.9035
0.903
Accuracy
25
30
35
40
45
50
K



Confusion Matrix
Output Class
Target Class
1
2
3
4
93
23.3%
2
0.5%
3
0.8%
2
0.5%
93.0%
7.0%
2
0.5%
90
22.5%
3
0.8%
5
1.3%
90.0%
10.0%
7
1.8%
1
0.3%
92
23.0%
0
0.0%
92.0%
8.0%
9
2.3%
5
1.3%
0
0.0%
86
21.5%
86.0%
14.0%
83.8%
16.2%
91.8%
8.2%
93.9%
6.1%
92.5%
7.5%
90.3%
9.8%



Bayes and Naive-Bayes classifier

Data Fitting :

Underlying Probability Distribution assumed to be Gaussian or mixture of Gaussians.

For a Gaussian Distribution :

- Maximum Likelihood Estimation method used for estimation of parameters.

- Covariance matrix for all the classes is the same
  MLE estimation done with parameters μ1,μ2,...μN as means for different classes and single covariance matrix C.

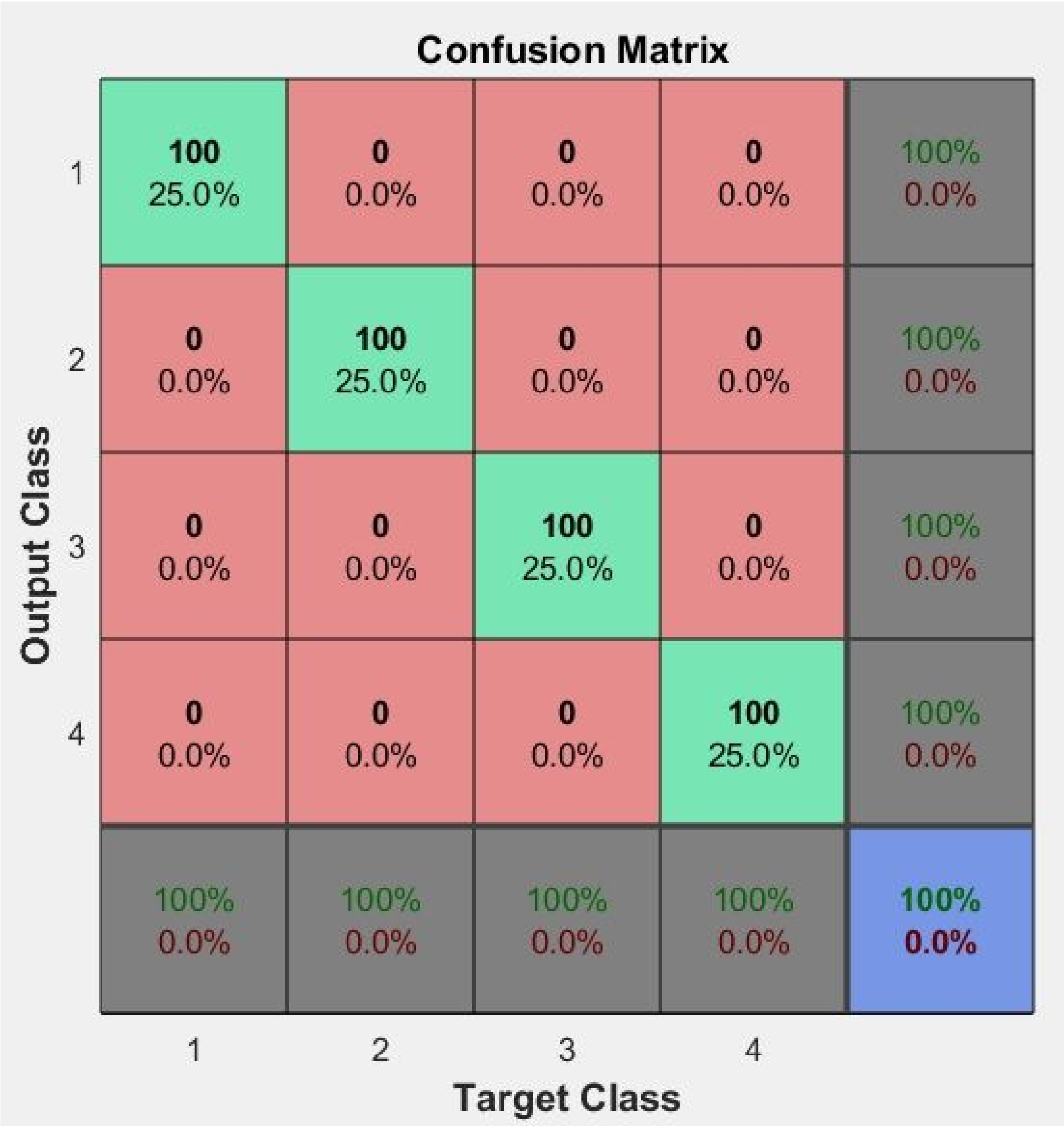
- Covariance matrix for all the classes is different
  MLE estimation done with parameters $\mu_1, \mu_2, \ldots \mu_N$ as means for different classes and $C_1, C_2, \ldots C_N$ as covariance matrix for different classes

- To suit Naive Bayes classifier,the built covariance matrix is made diagonal.

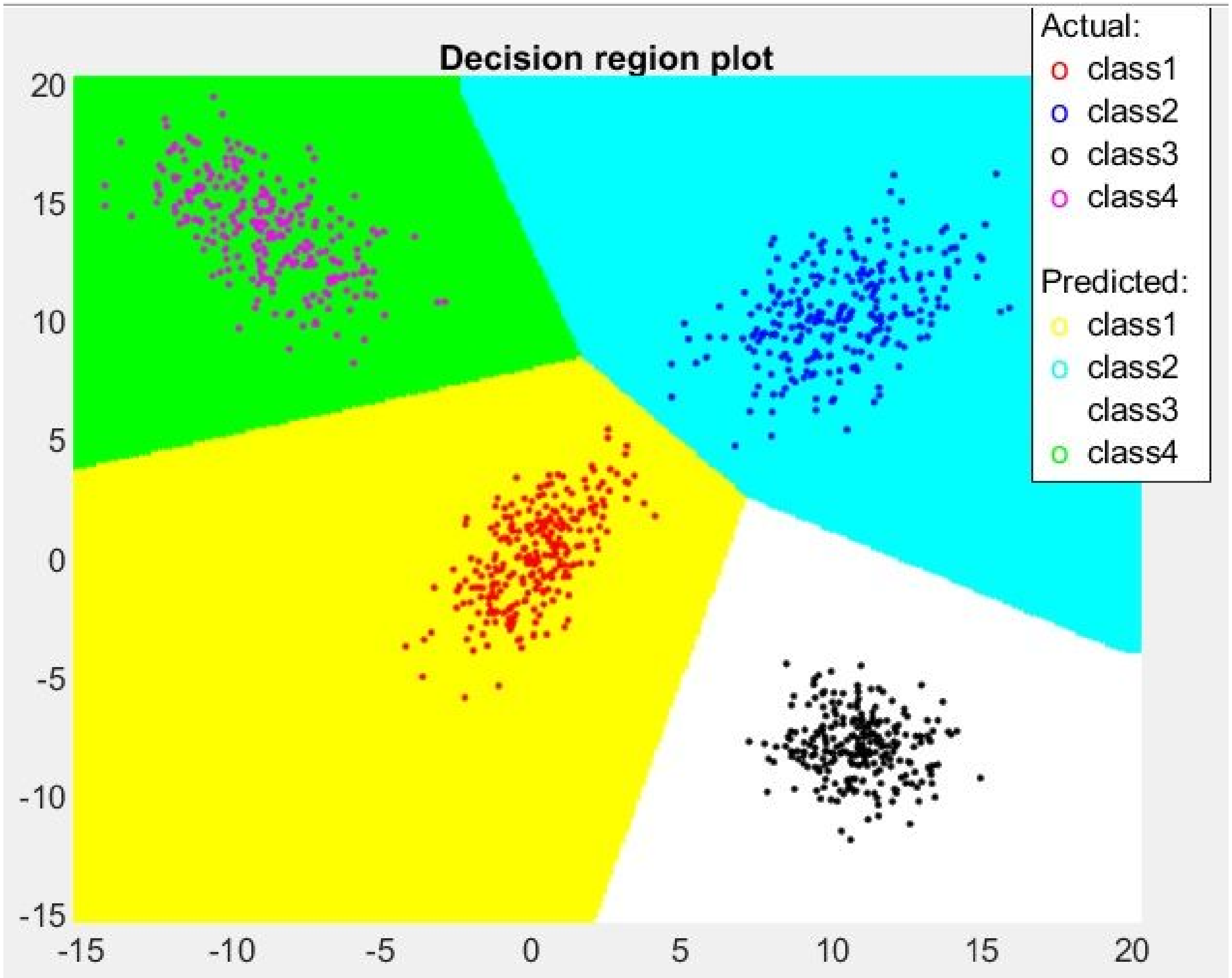
Bayes Classifier:
Same Covariance Matrix:

Linearly Separable Data:

- The decision surface takes the form of hyper planes separating class points with 100 percent accuracy.

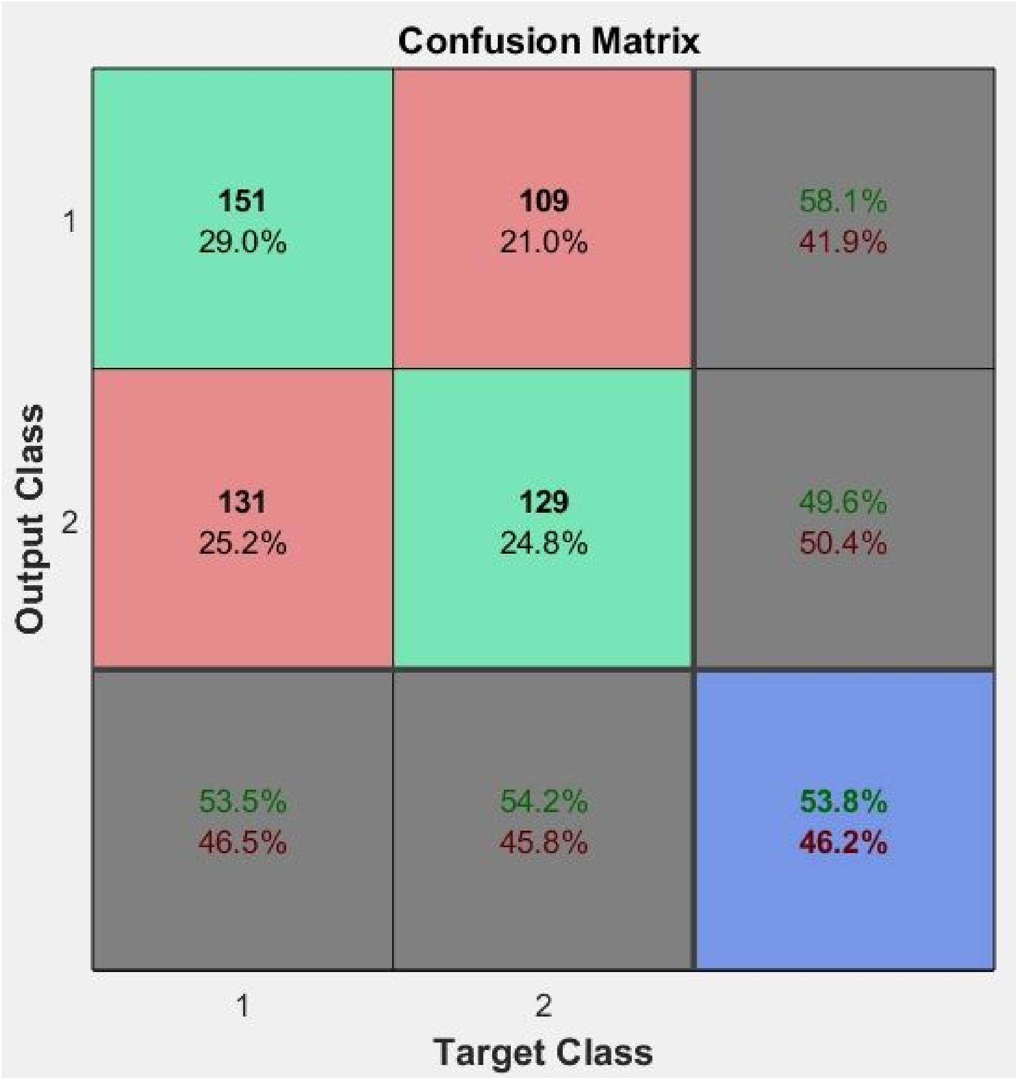


Confusion Matrix
Output Class
Target Class
1
2
3
4
100
25.0%
0
0.0%
100%
0.0%

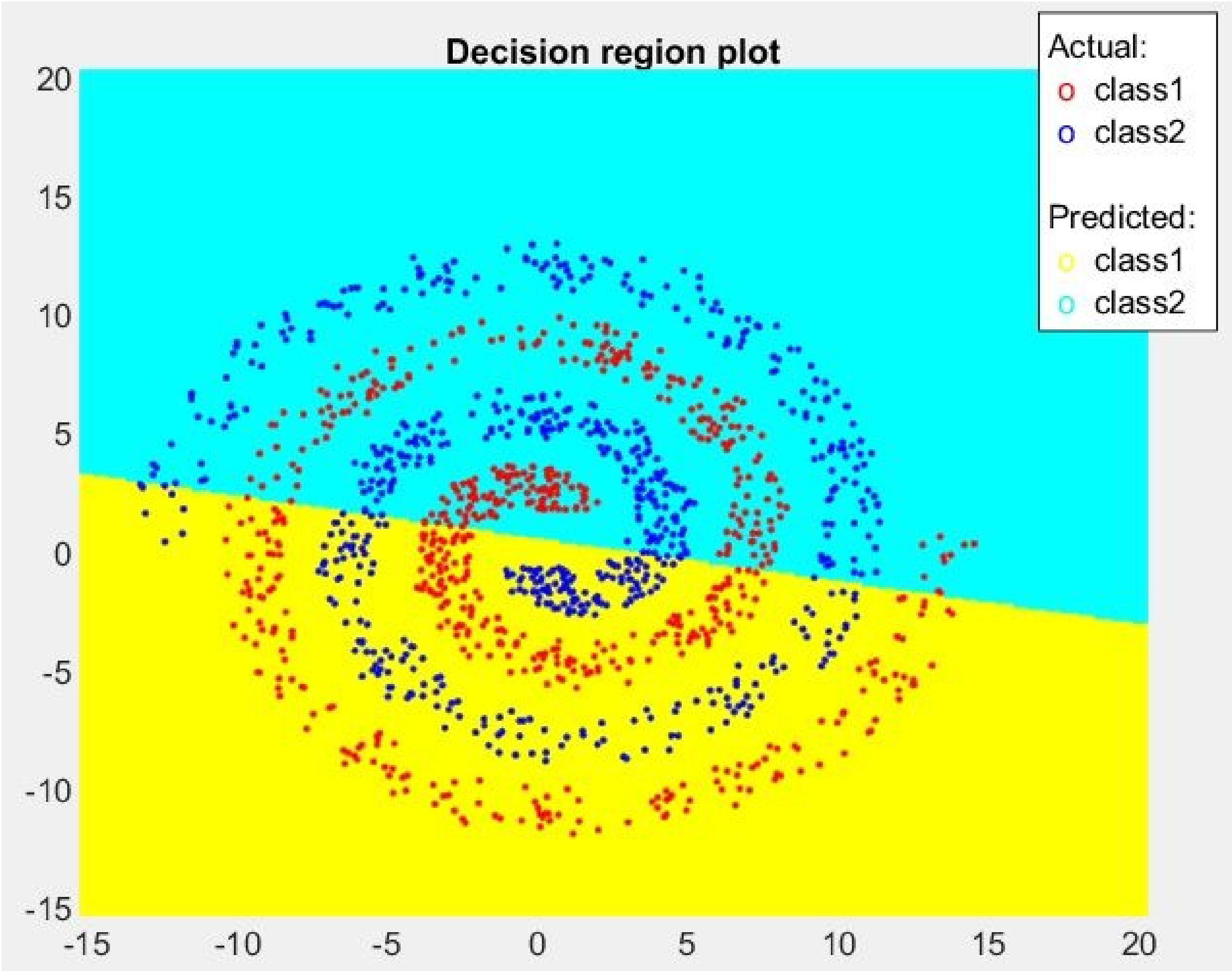


Non Linearly Separable Data:

- The accuracy for test data is hyper plane(as seen from the results) cutting the spiral into approximately two halves.

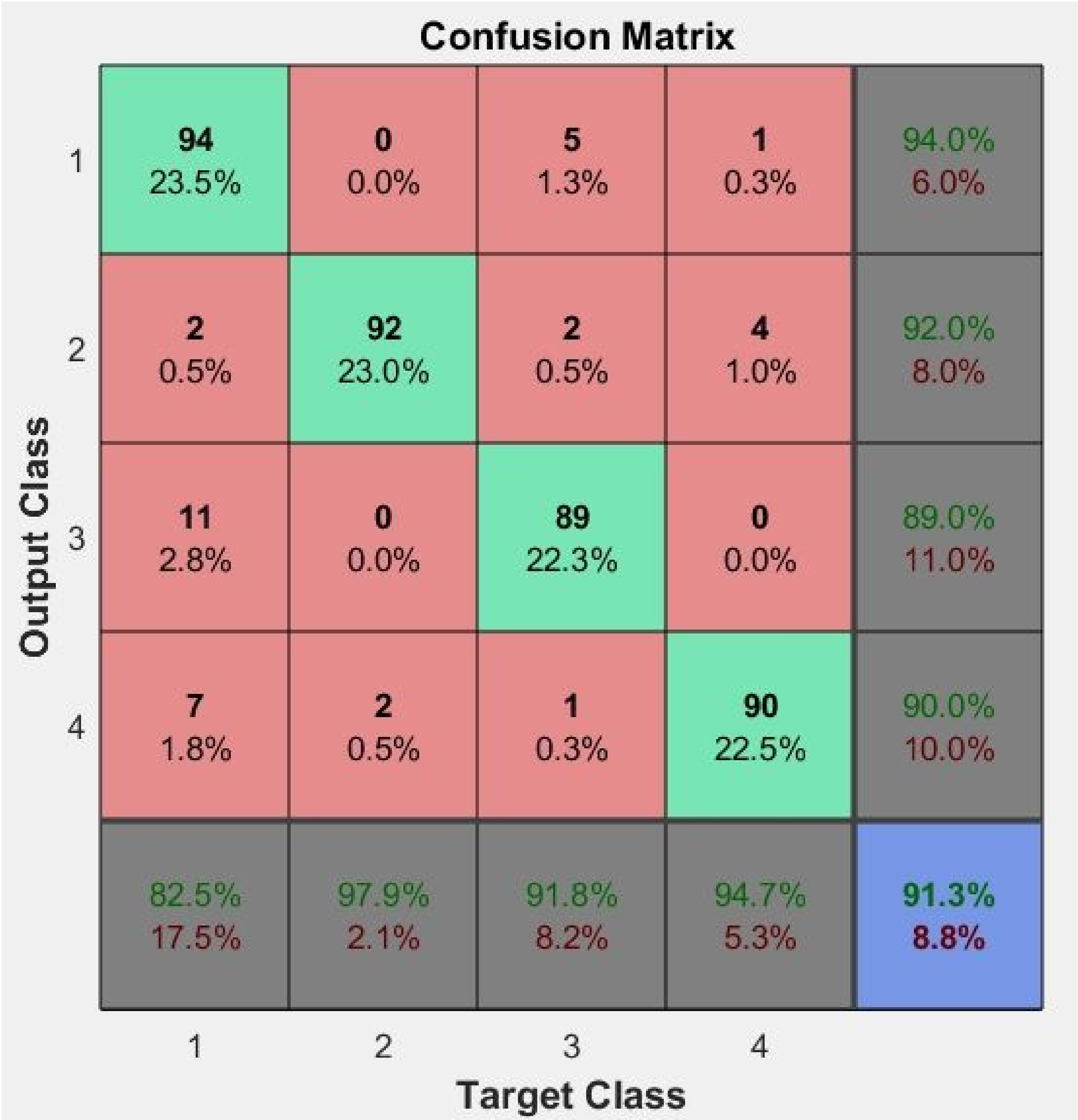


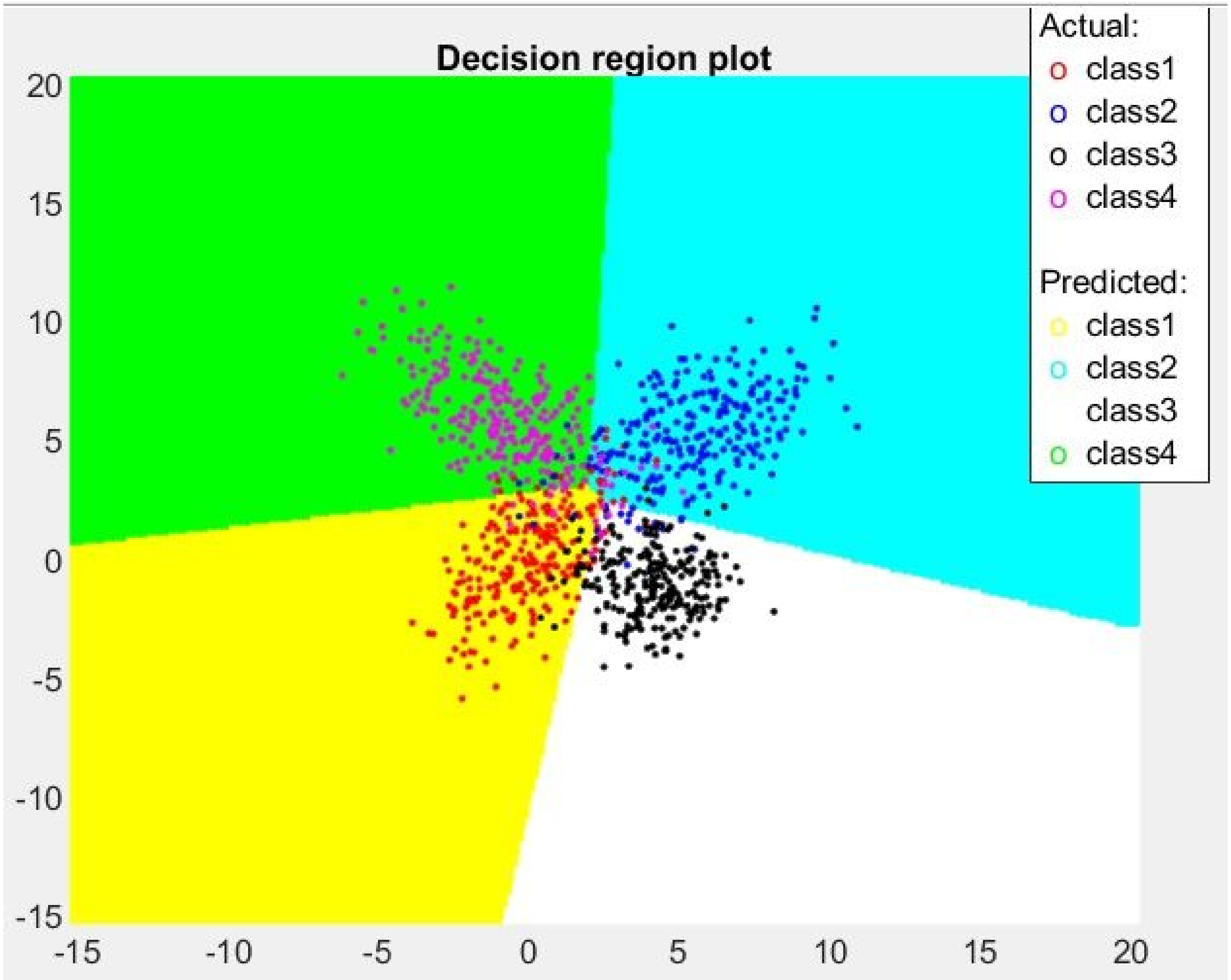
Confusion Matrix
151
29.0%
109
21.0%
58.1%
41.9%
131
25.2%
129
24.8%
49.6%
50.4%
53.5%
46.5%
54.2%
45.8%
53.8%
46.2%
1
2
Output Class
1
2
Target Class



Overlapping Data:

- The accuracy of the results is better since the gaussian distributions fit with centre as means of the data points of that particular class.

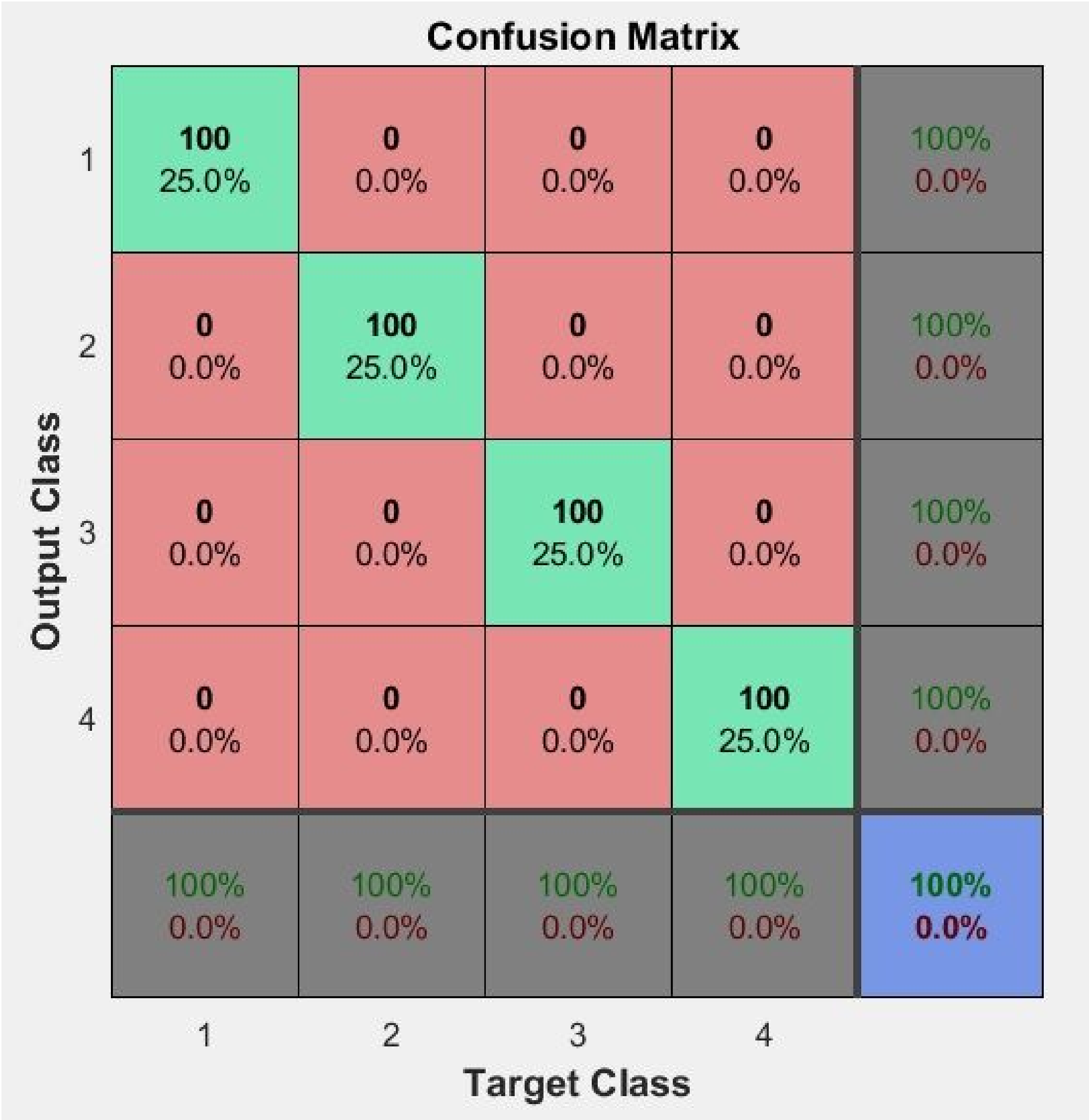

Confusion Matrix
94
23.5%
0
0.0%
5
1.3%
1
0.3%
94.0%
6.0%
2
0.5%
92
23.0%
2
0.5%
4
1.0%
92.0%
8.0%
11
2.8%
0
0.0%
89
22.3%
0
0.0%
89.0%
11.0%
7
1.8%
2
0.5%
1
0.3%
90
22.5%
90.0%
10.0%
82.5%
17.5%
97.9%
2.1%
91.8%
8.2%
94.7%
5.3%
91.3%
8.8%
Output Class
1
2
3
4
Target Class
1
2
3
4

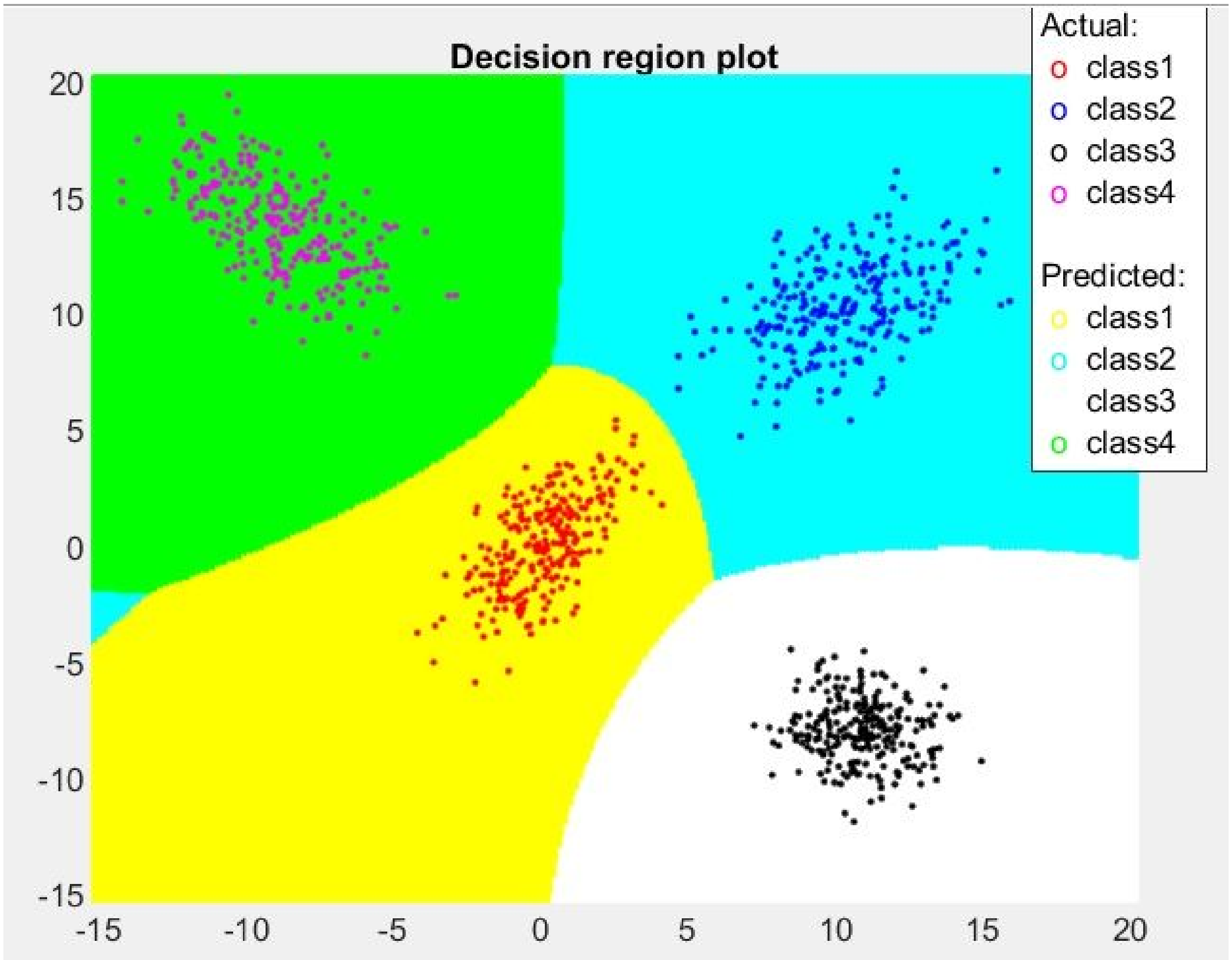


Different Covariance Matrix :

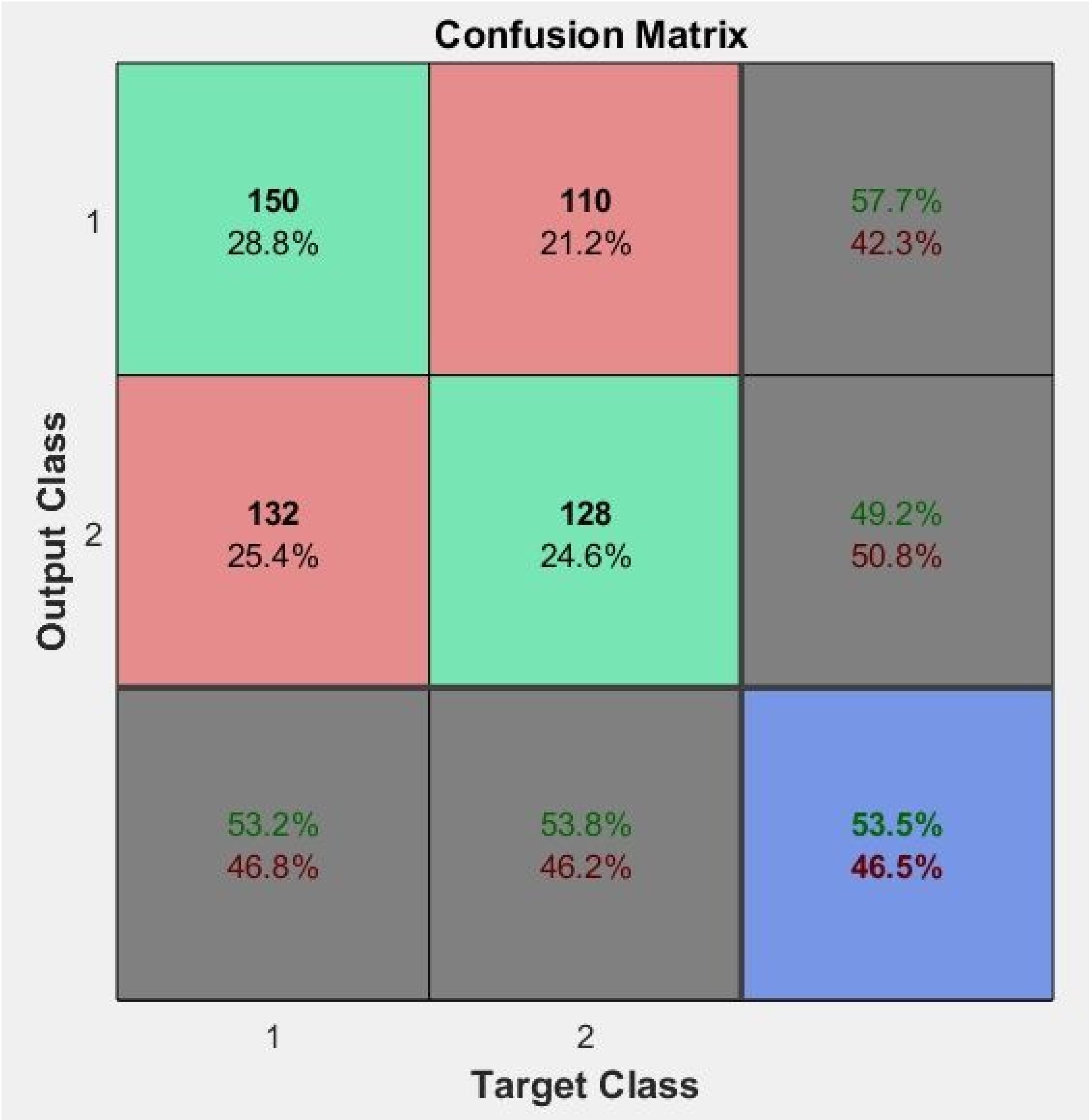
- The decision surfaces are hyper quadric in nature.

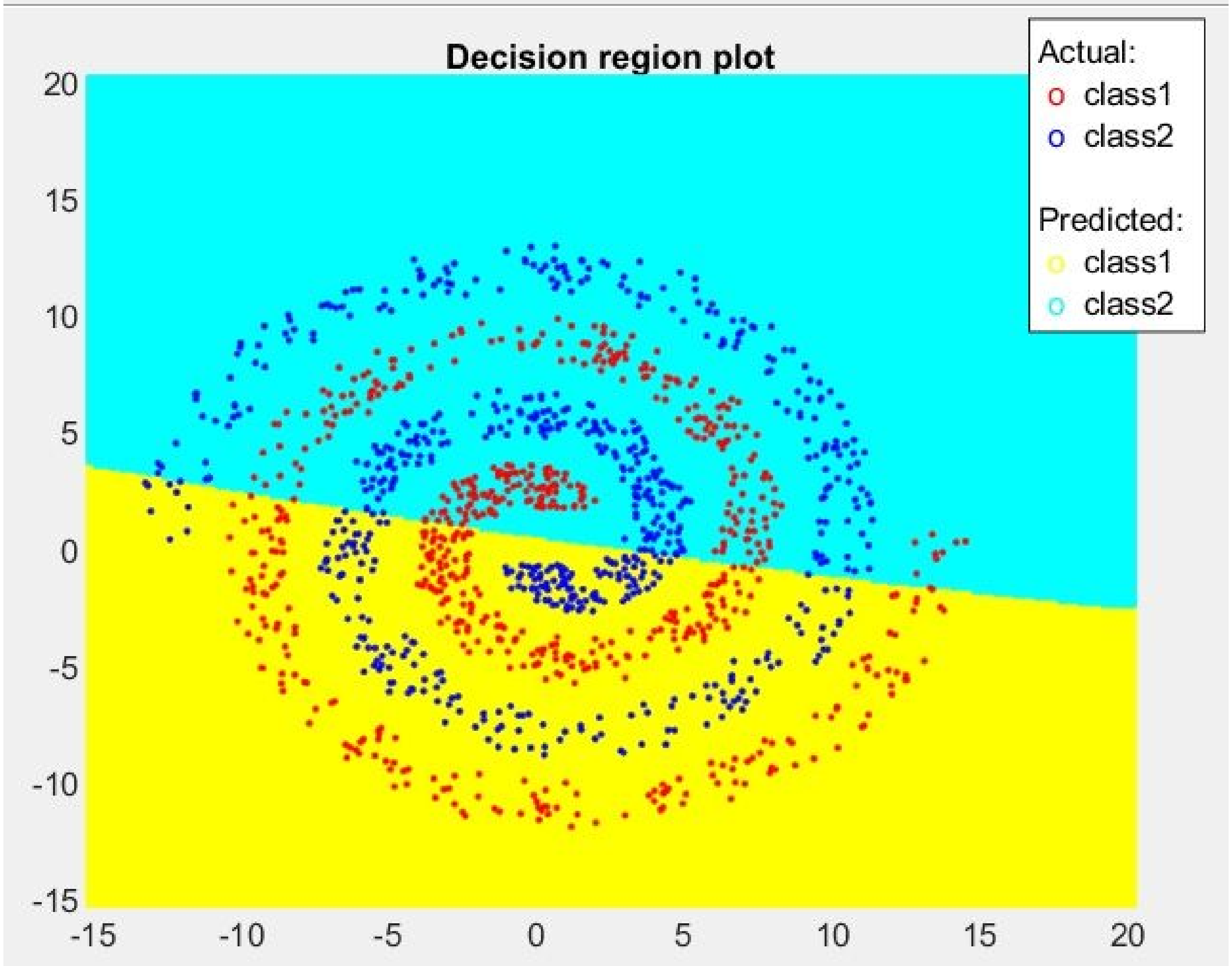
Linearly Separable Data:


Confusion Matrix
Output Class
Target Class
1
2
3
4
100
25.0%
0
0.0%
100%

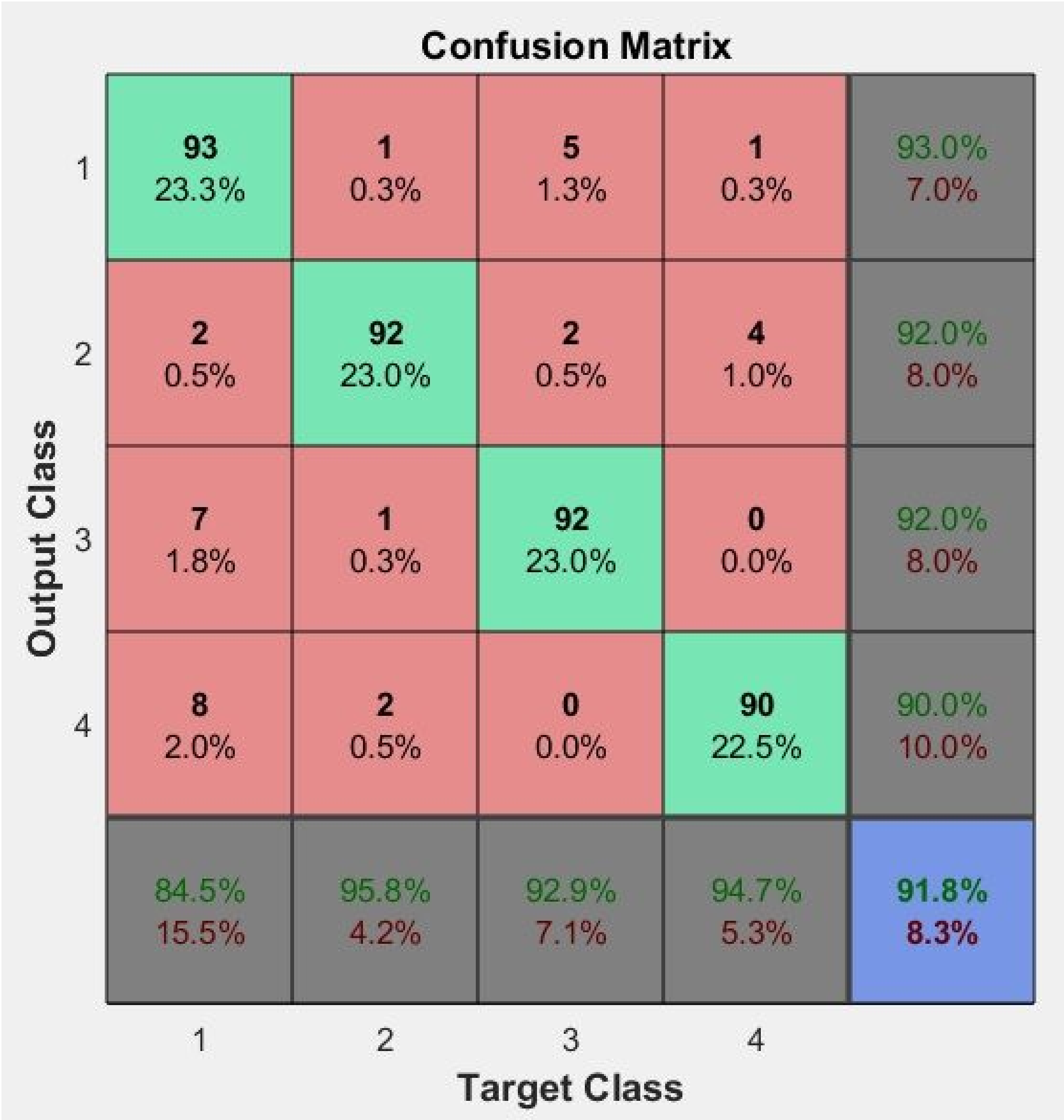


Non Linearly Separable Data:

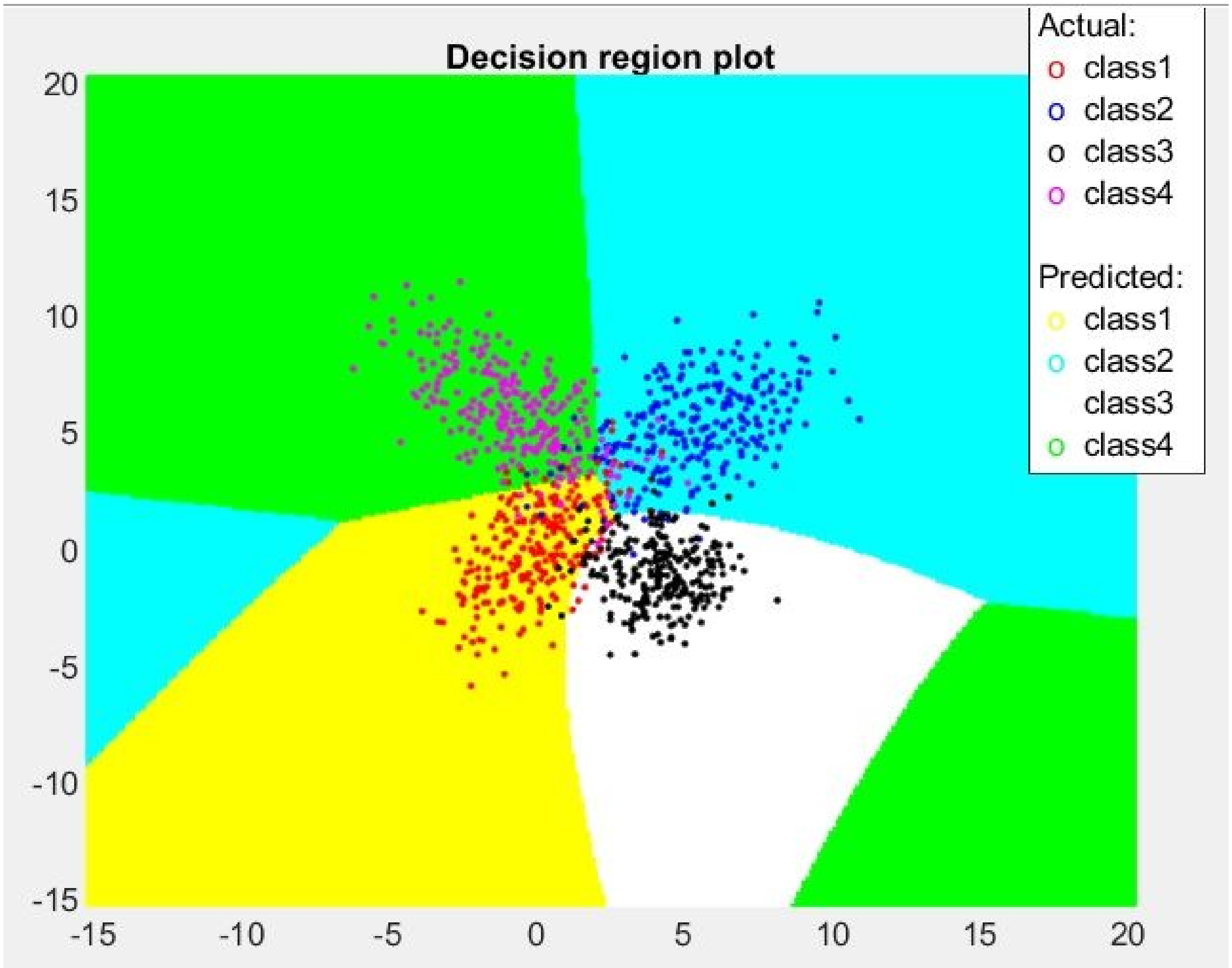

Confusion Matrix
150
28.8%
110
21.2%
57.7%
42.3%
132
25.4%
128
24.6%
49.2%
50.8%
53.2%
46.8%
53.8%
46.2%
53.5%
46.5%
1
2
1
2
Output Class
Target Class



Overlapping Data:



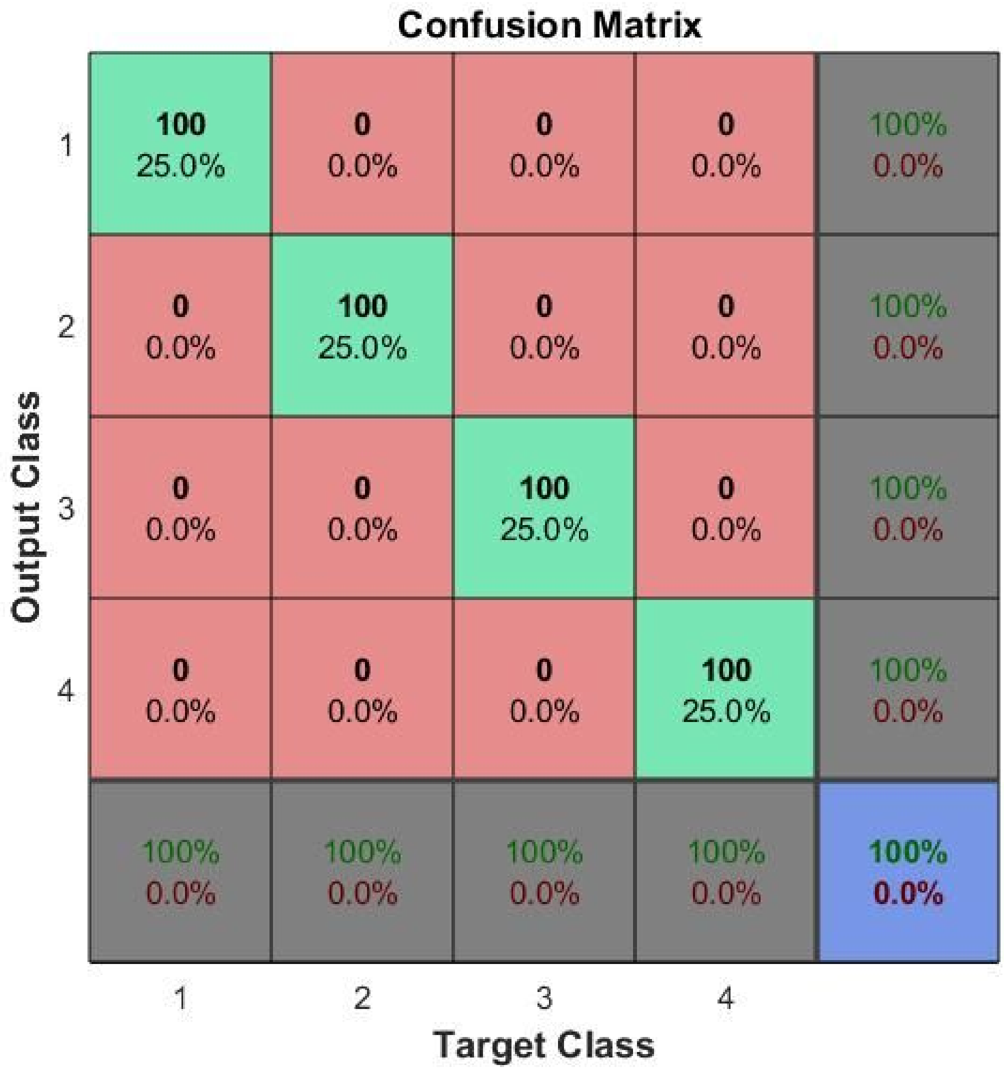
Confusion Matrix
Output Class
Target Class
1
2
3
4
93
23.3%
1
0.3%
5
1.3%
1
0.3%
93.0%
7.0%
2
0.5%
92
23.0%
2
0.5%
4
1.0%
92.0%
8.0%
7
1.8%
1
0.3%
92
23.0%
0
0.0%
92.0%
8.0%
8
2.0%
2
0.5%
0
0.0%
90
22.5%
90.0%
10.0%
84.5%
15.5%
95.8%
4.2%
92.9%
7.1%
94.7%
5.3%
91.8%
8.3%

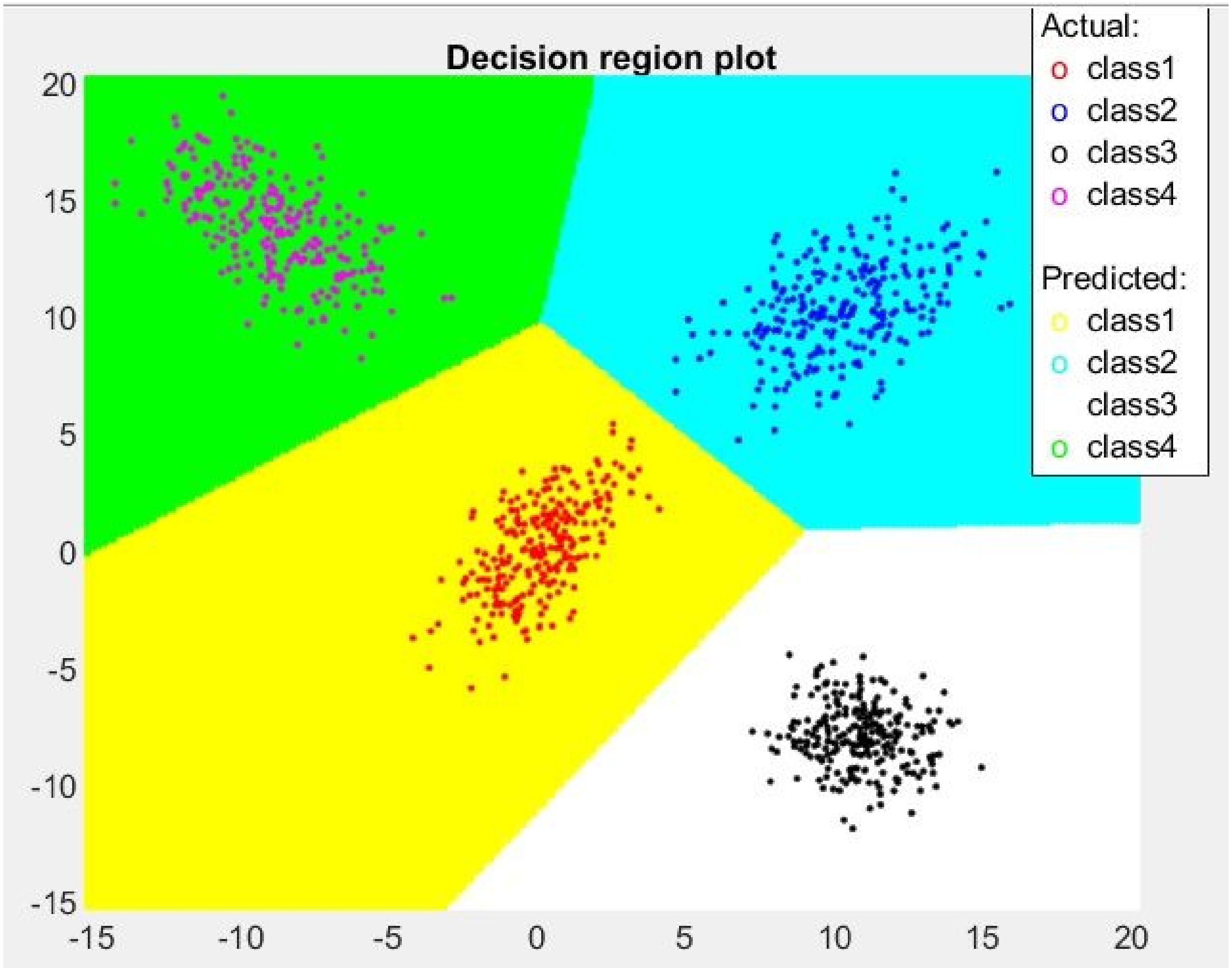


<u>Naive Bayes Classifier</u>

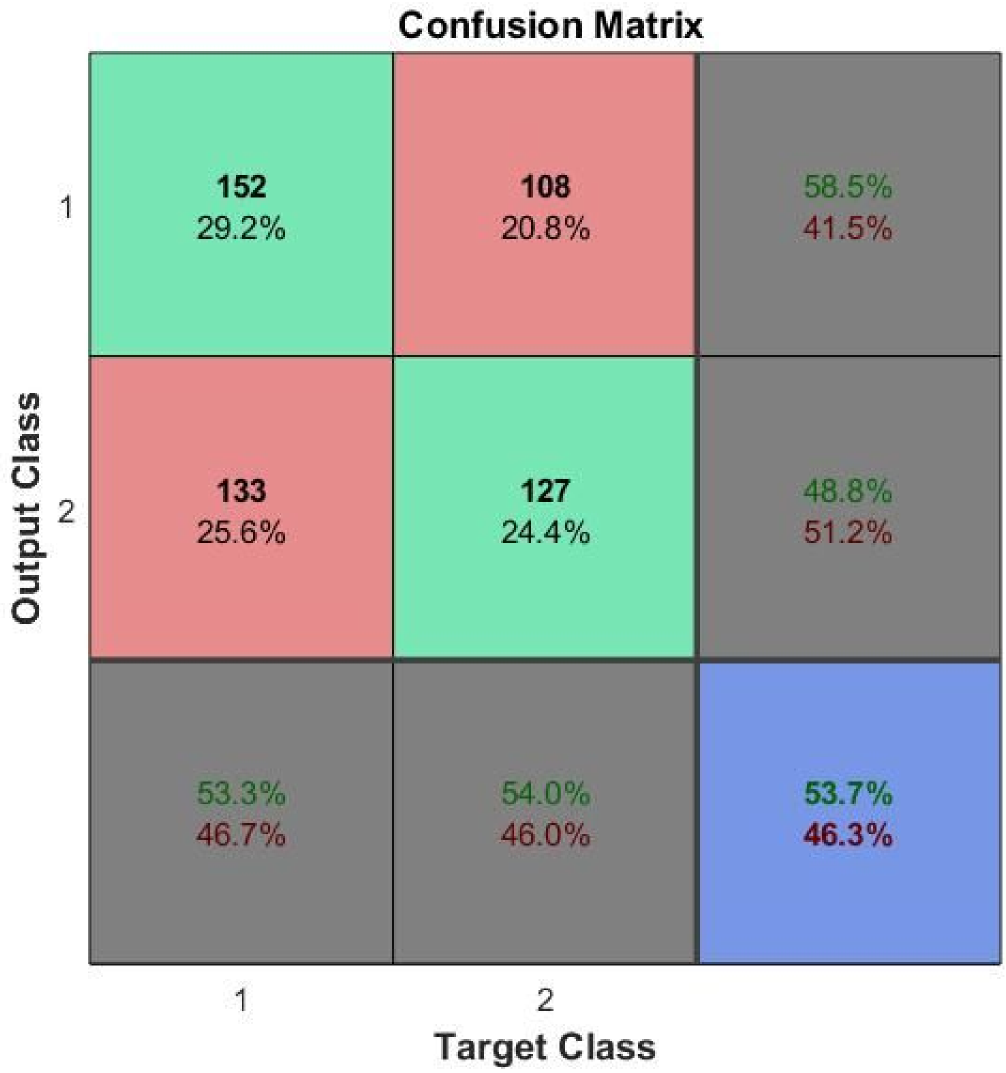
The covariance matrix is diagonal.
The decision surfaces are of hyper planes in nature.
<u>C = sigma^2 I</u>

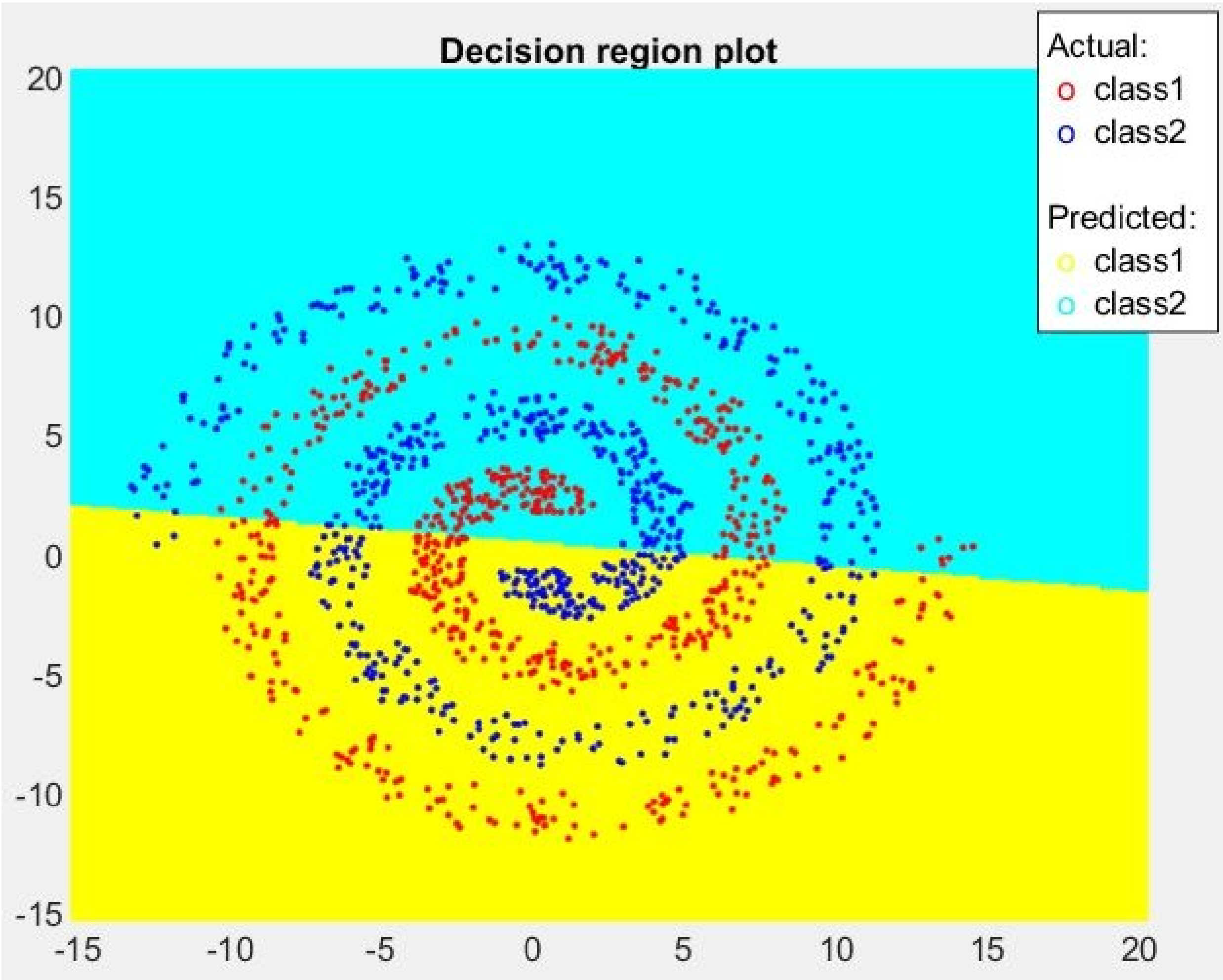
Linearly Separable Data:


Confusion Matrix
Output Class
Target Class
1
2
3
4
100
25.0%
0
0.0%
100%

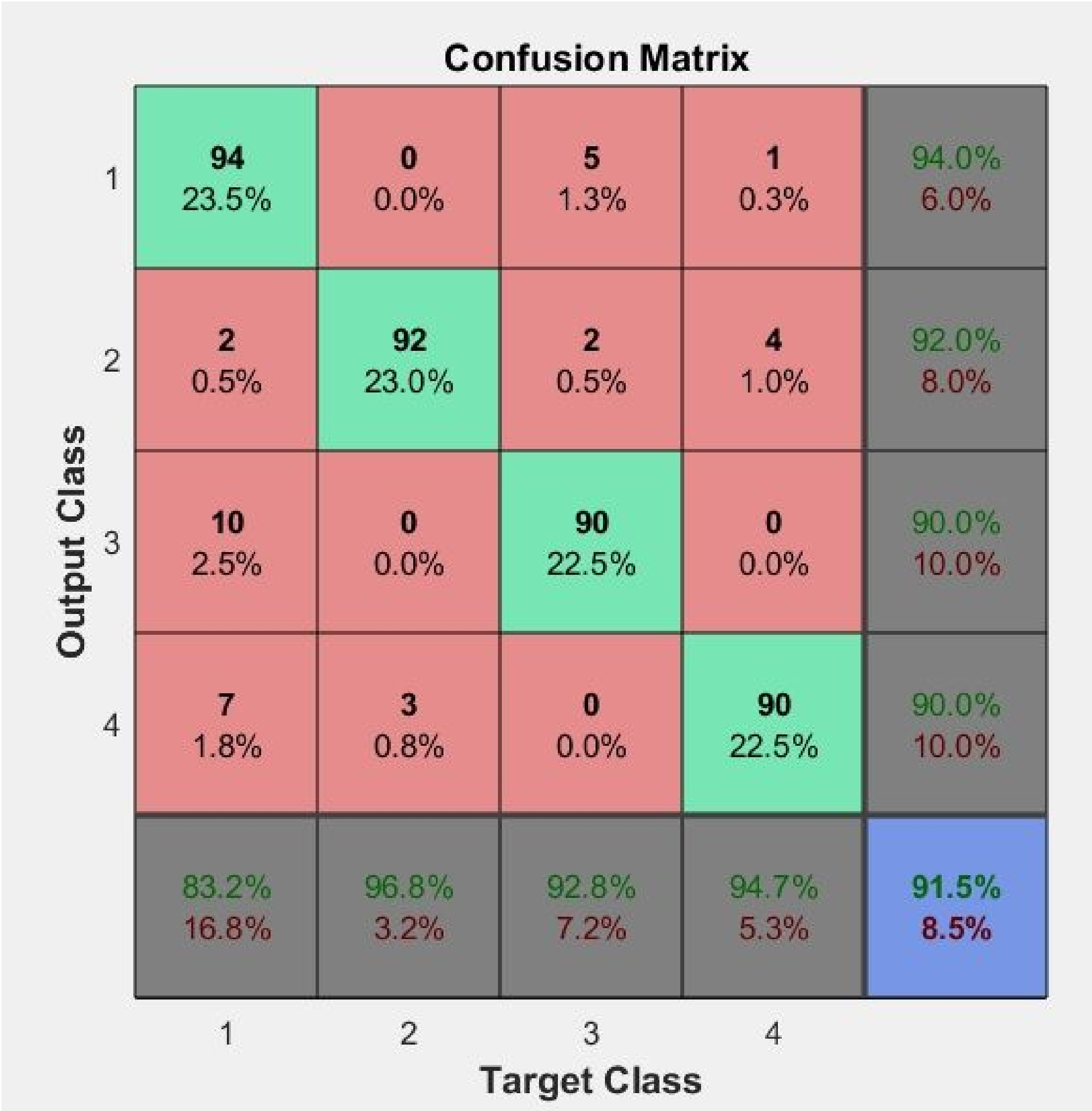


Non Linearly Separable Data:

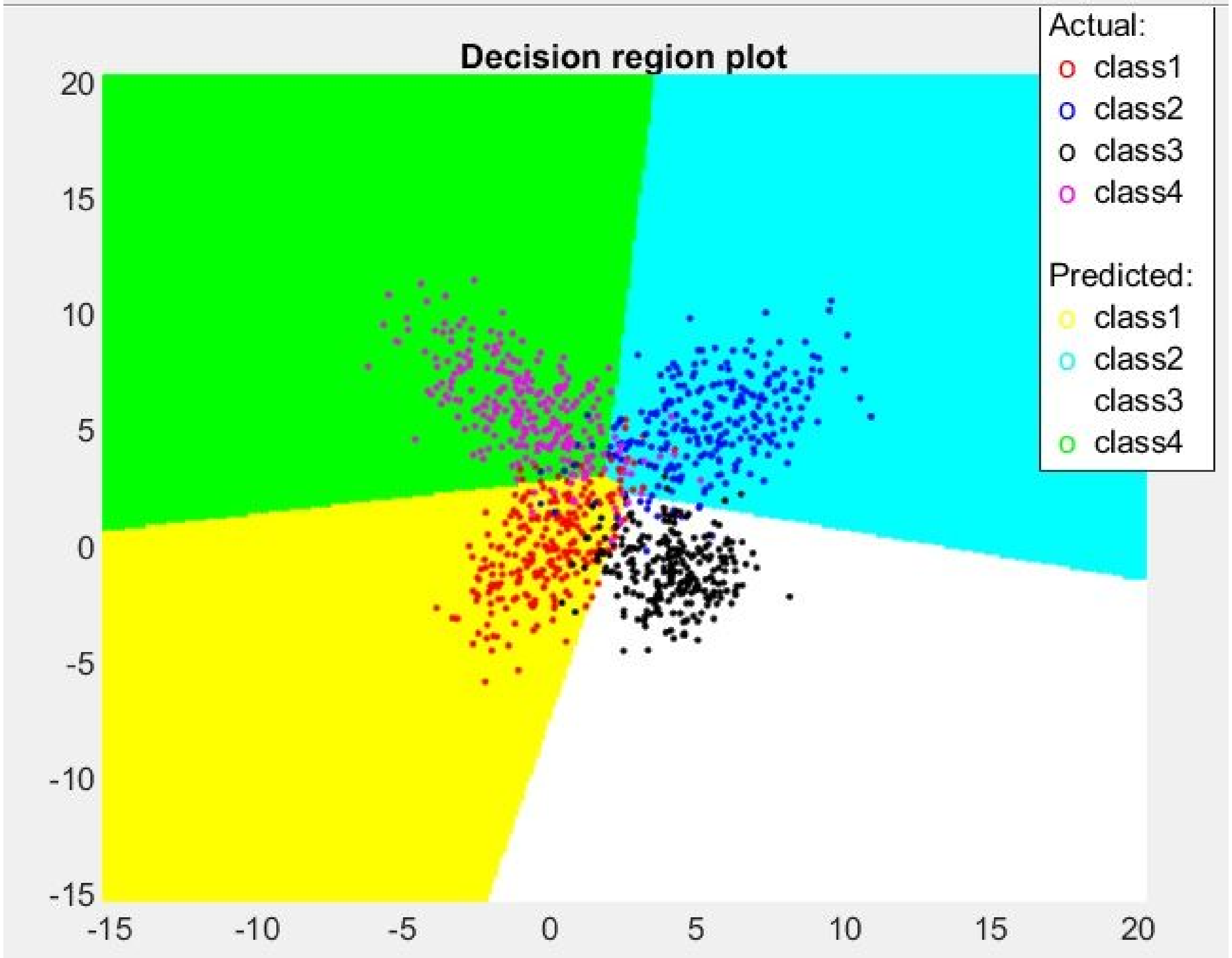

Confusion Matrix
152
29.2%
108
20.8%
58.5%
41.5%
133
25.6%
127
24.4%
48.8%
51.2%
53.3%
46.7%
54.0%
46.0%
53.7%
46.3%
1
2
Output Class
1
2
Target Class



Overlapping Data:



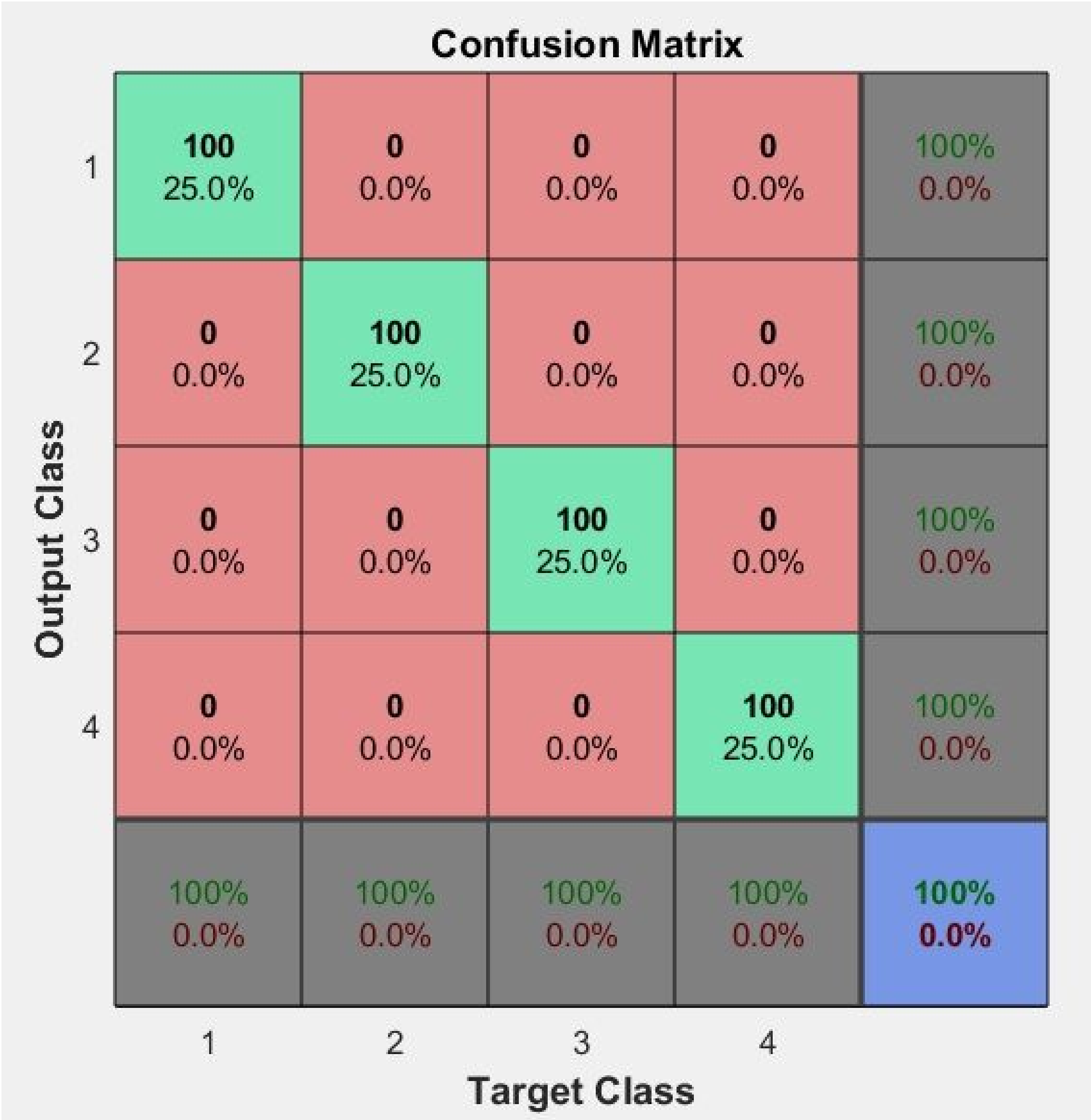
Confusion Matrix
94
23.5%
0
0.0%
5
1.3%
1
0.3%
94.0%
6.0%
2
0.5%
92
23.0%
2
0.5%
4
1.0%
92.0%
8.0%
10
2.5%
0
0.0%
90
22.5%
0
0.0%
90.0%
10.0%
7
1.8%
3
0.8%
0
0.0%
90
22.5%
90.0%
10.0%
83.2%
16.8%
96.8%
3.2%
92.8%
7.2%
94.7%
5.3%
91.5%
8.5%
Output Class
1
2
3
4
Target Class
1
2
3
4

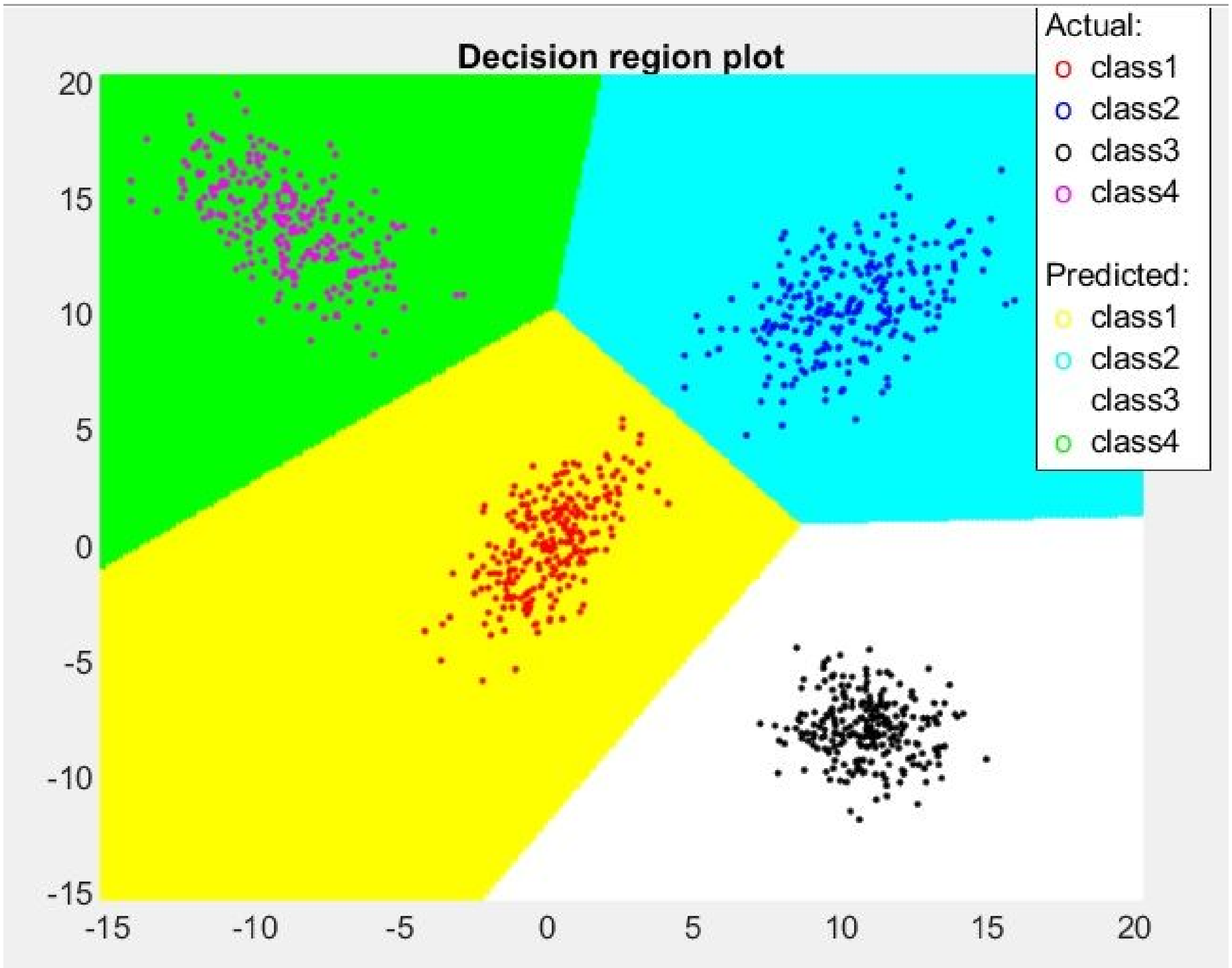


<u>**Same Covariance Matrix:**</u>

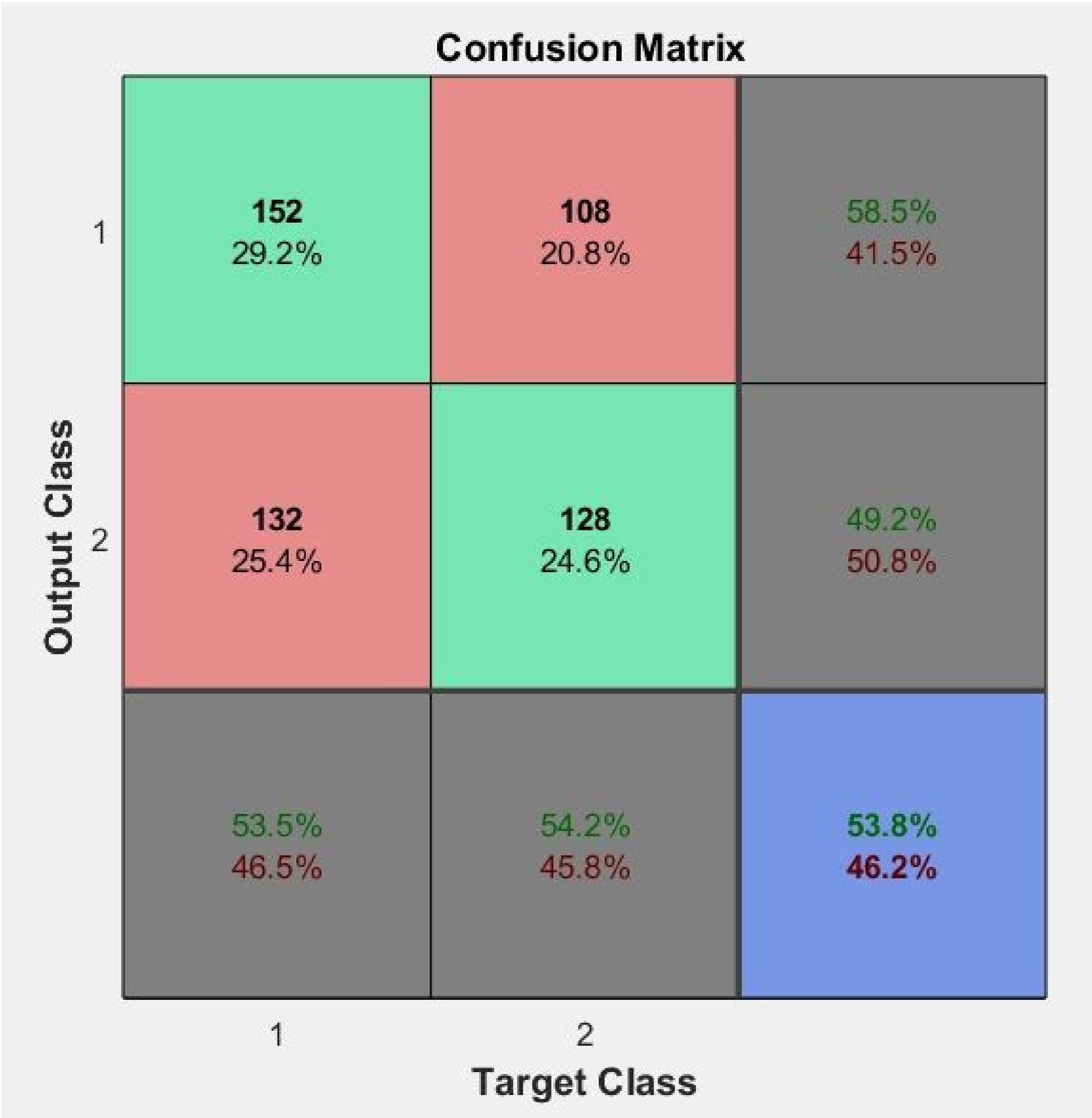
- The decision surfaces are of hyper planes in nature

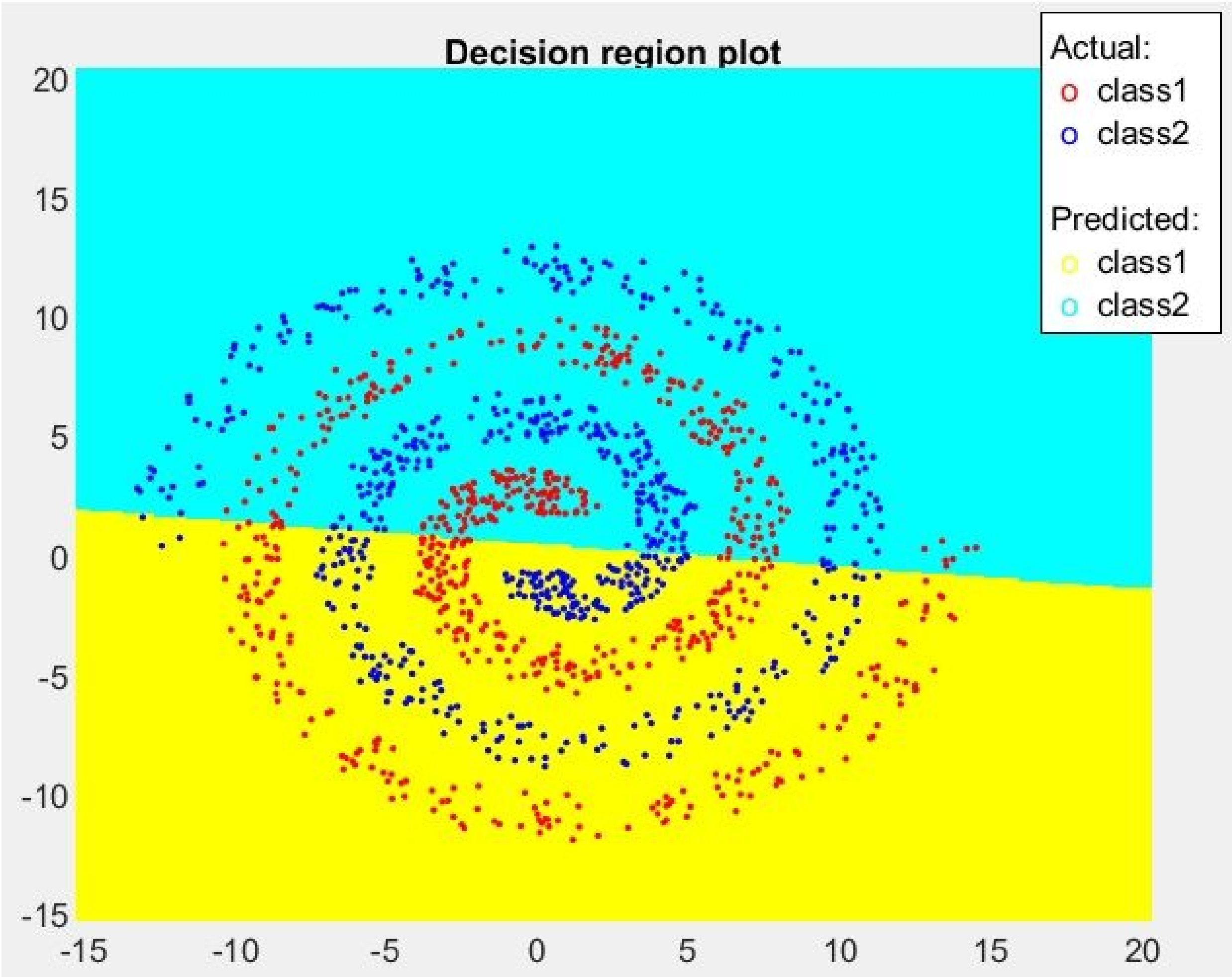
Linearly Separable Data:


Confusion Matrix
100
25.0%
0
0.0%
0
0.0%
0
0.0%
100%
0.0%
0
0.0%
100
25.0%
0
0.0%
0
0.0%
100%
0.0%
0
0.0%
0
0.0%
100
25.0%
0
0.0%
100%
0.0%
0
0.0%
0
0.0%
0
0.0%
100
25.0%
100%
0.0%
100%
0.0%
100%
0.0%
100%
0.0%
100%
0.0%
100%
0.0%
Output Class
1
2
3
4
Target Class
1
2
3
4

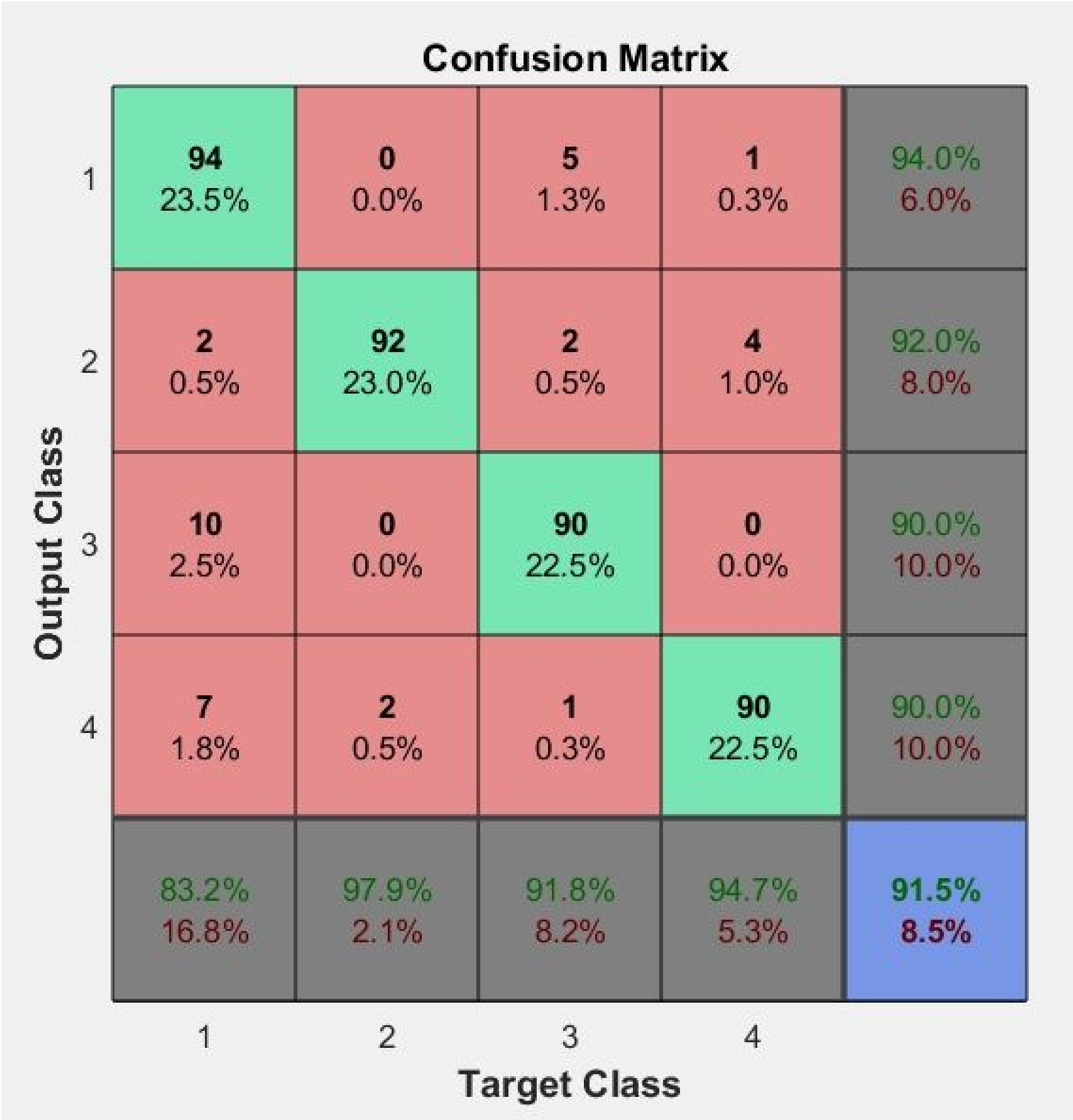


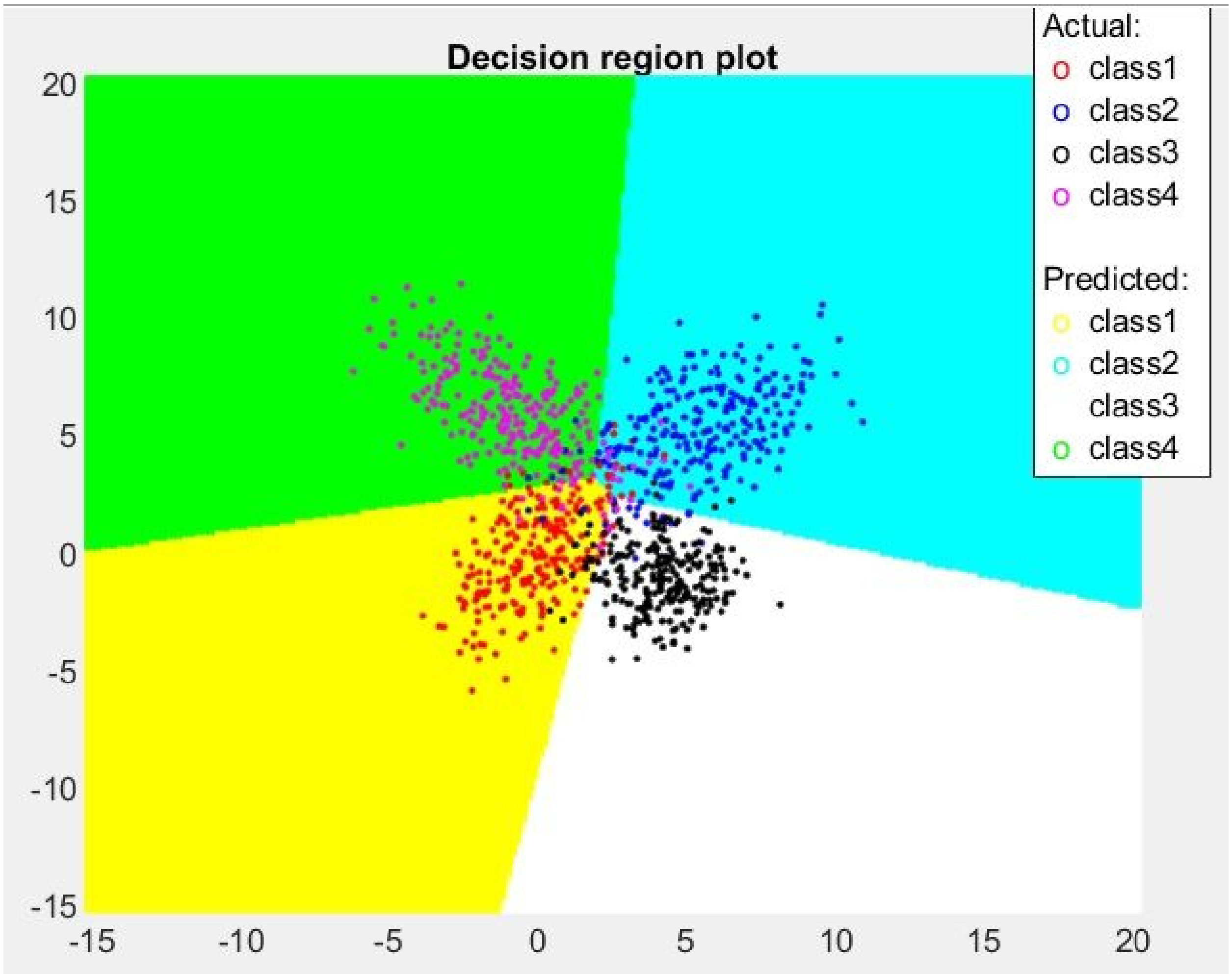
Non Linearly Separable Data:


Confusion Matrix
152
29.2%
108
20.8%
58.5%
41.5%
132
25.4%
128
24.6%
49.2%
50.8%
53.5%
46.5%
54.2%
45.8%
53.8%
46.2%
1
2
Output Class
1
2
Target Class



Overlapping Data:

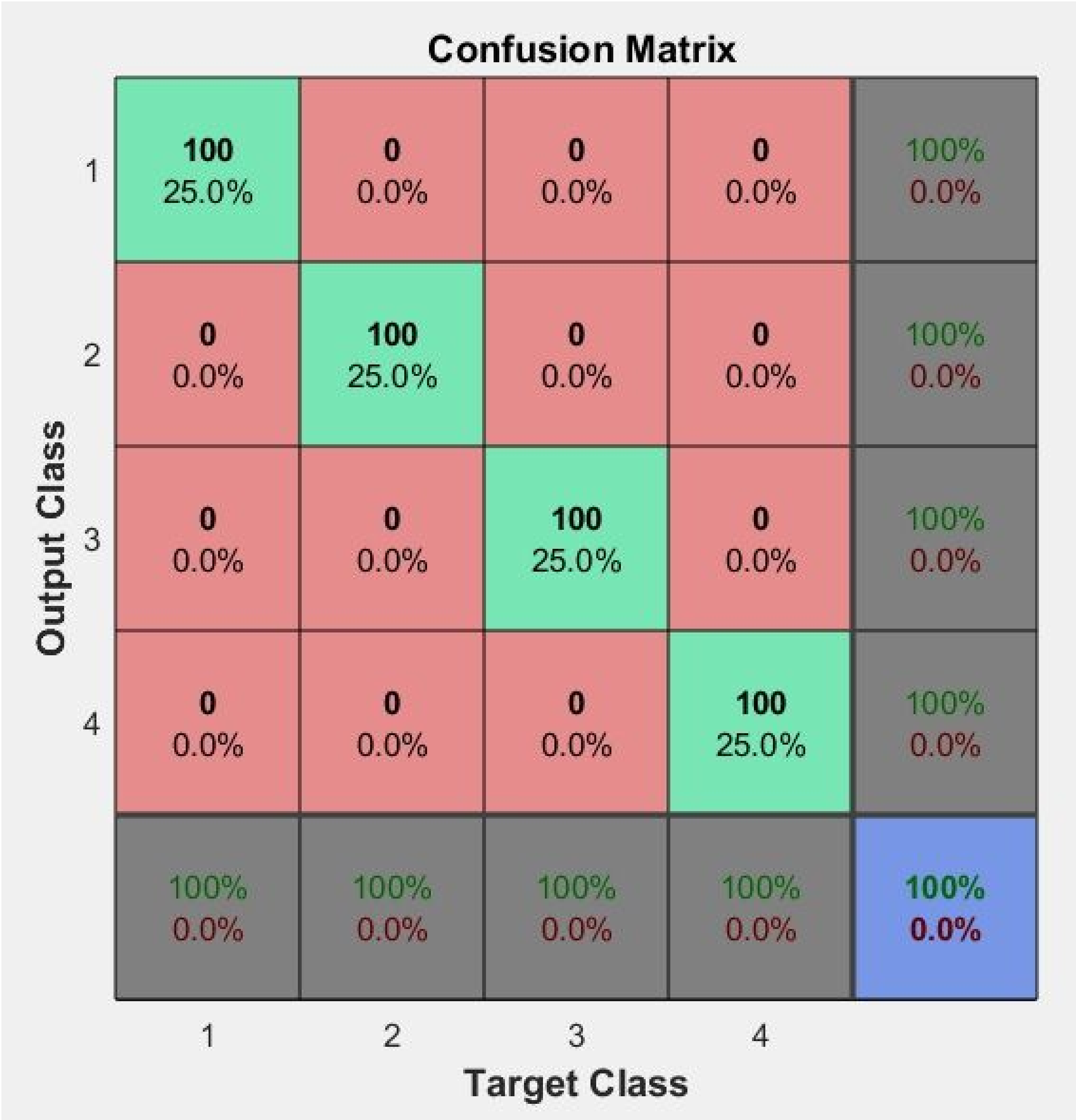

Confusion Matrix
94
23.5%
0
0.0%
5
1.3%
1
0.3%
94.0%
6.0%
2
0.5%
92
23.0%
2
0.5%
4
1.0%
92.0%
8.0%
10
2.5%
0
0.0%
90
22.5%
0
0.0%
90.0%
10.0%
7
1.8%
2
0.5%
1
0.3%
90
22.5%
90.0%
10.0%
83.2%
16.8%
97.9%
2.1%
91.8%
8.2%
94.7%
5.3%
91.5%
8.5%
Output Class
1
2
3
4
Target Class
1
2
3
4

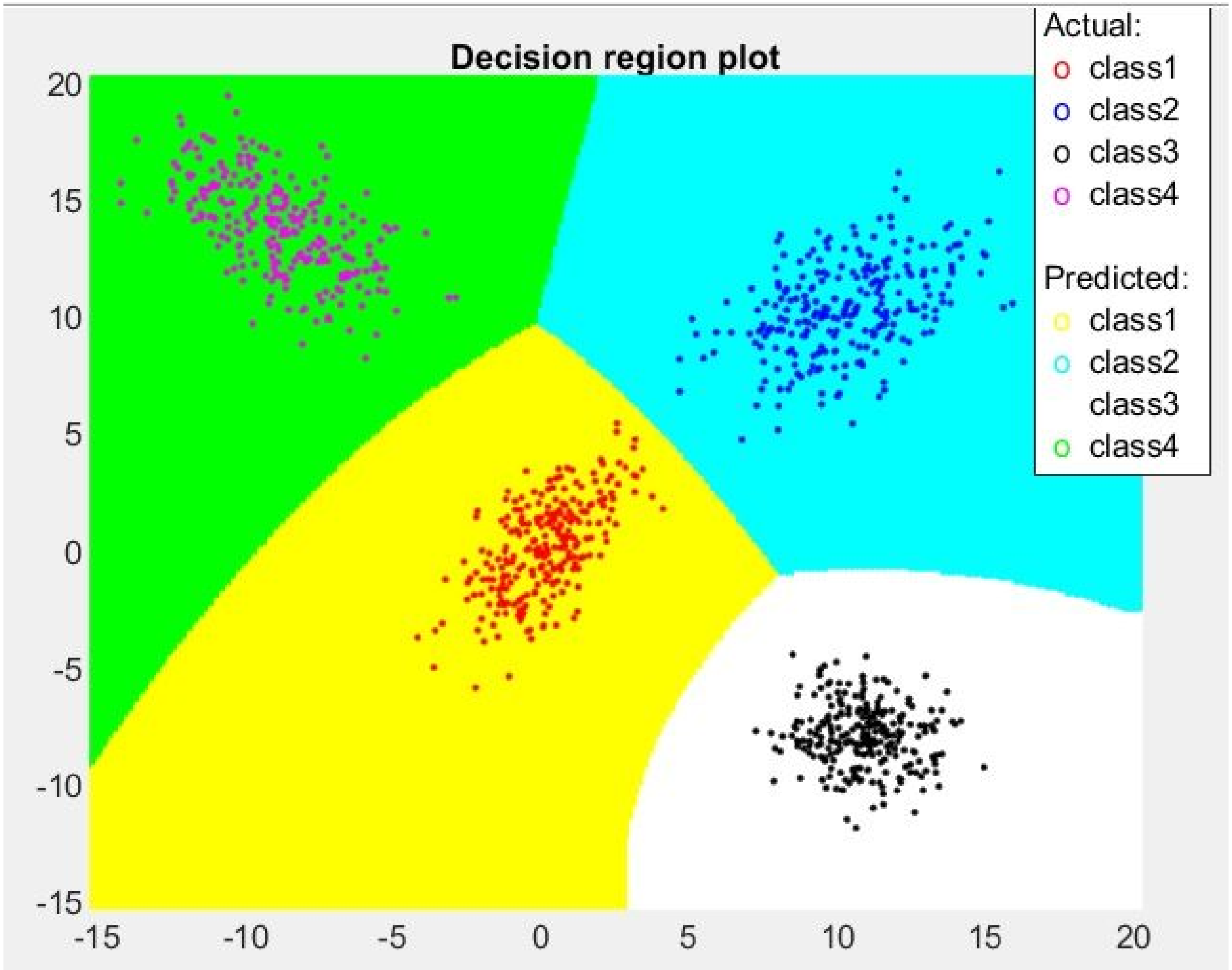


Different Covariance Matrix :

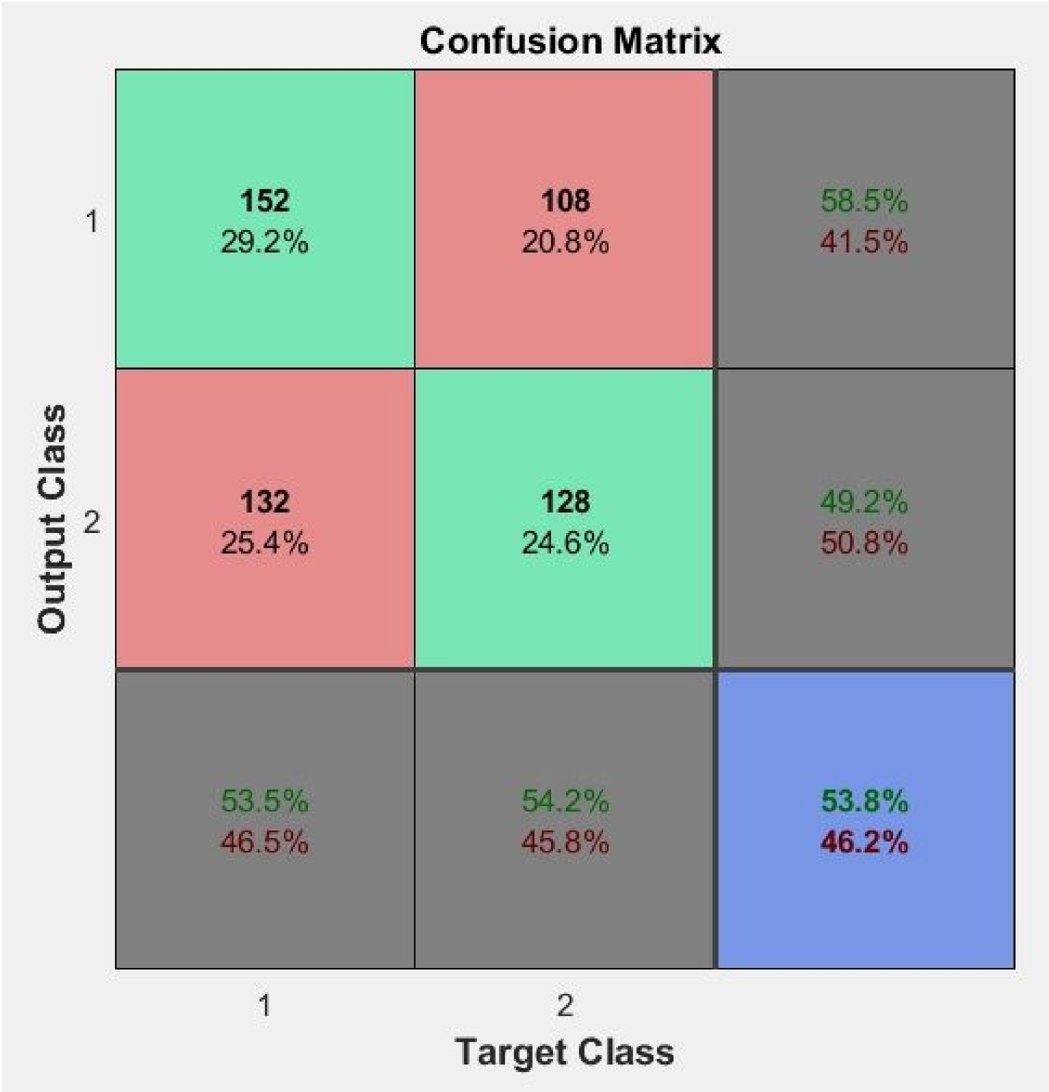
- The decision surfaces are hyper quadric in nature.

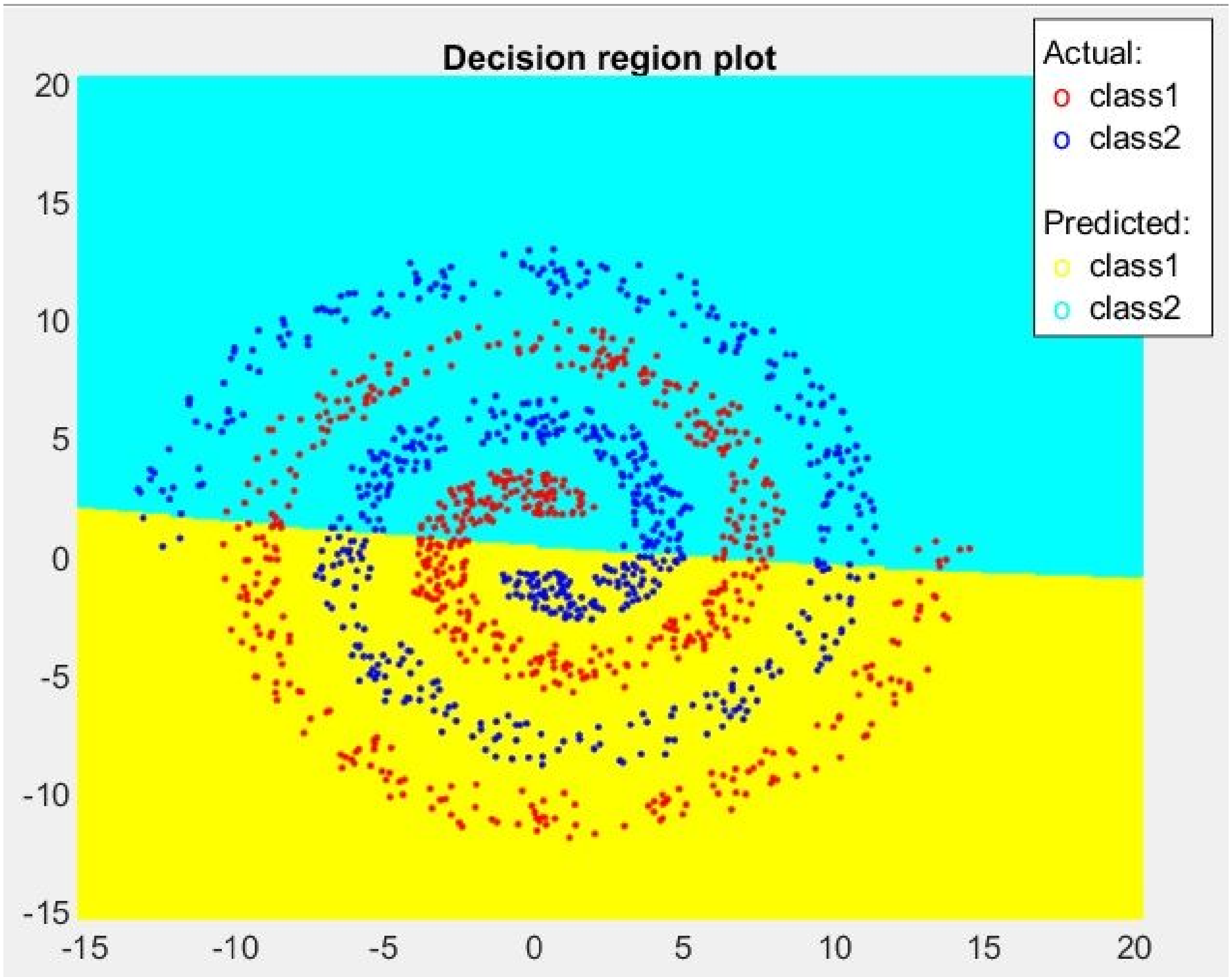
Linearly Separable Data:


Confusion Matrix
Output Class
Target Class
1
2
3
4
100
25.0%
0
0.0%
100%

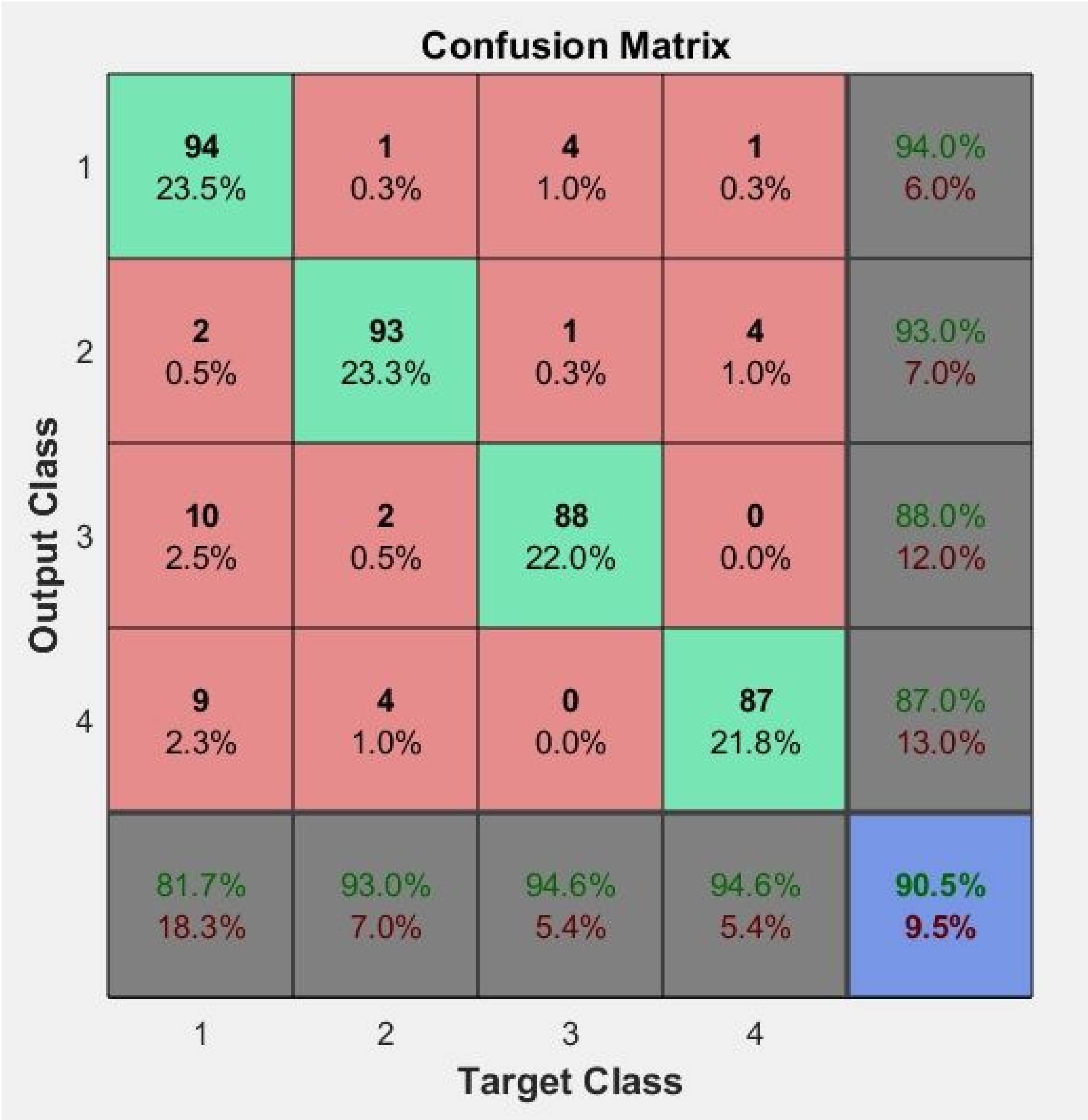


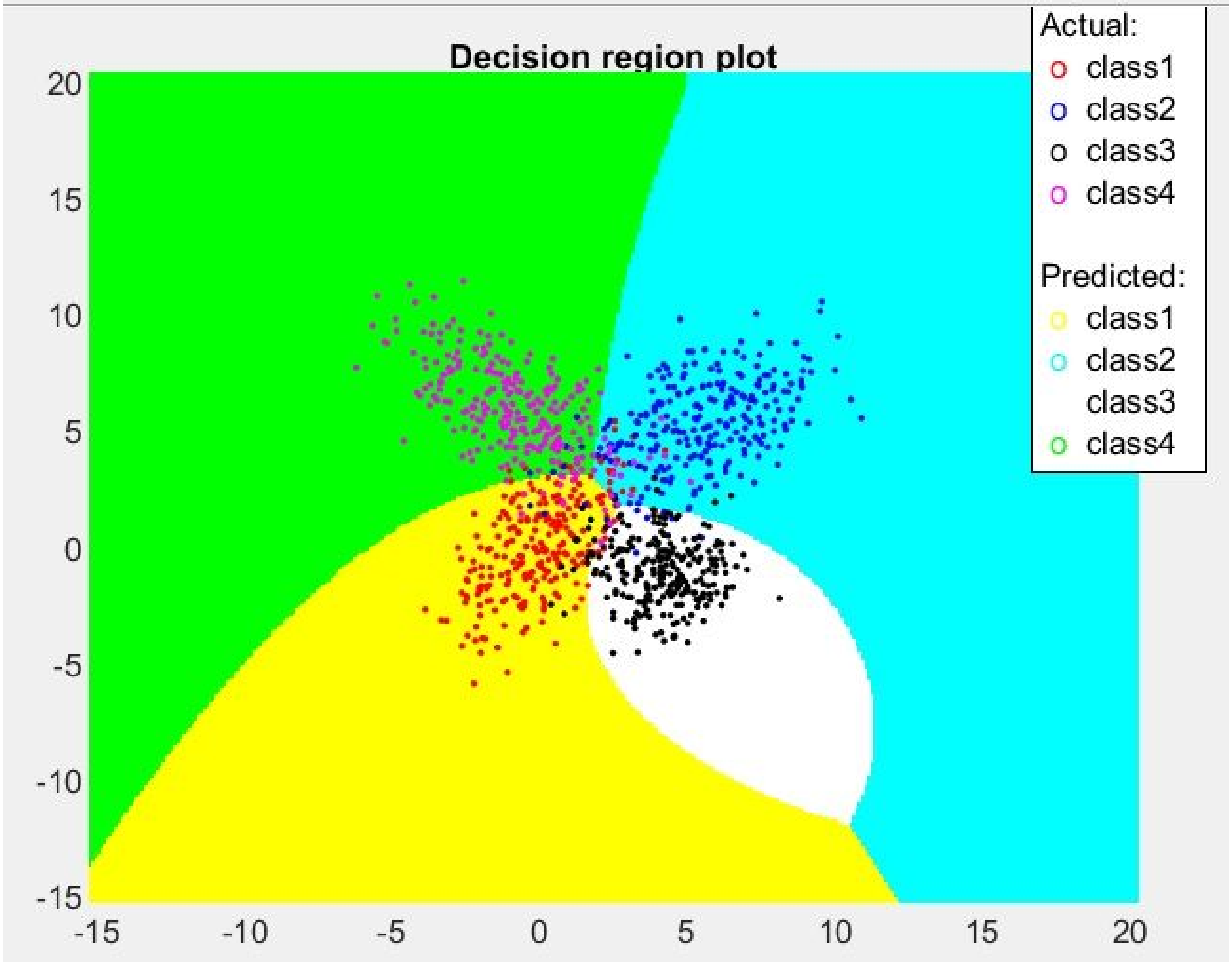
Non Linearly Separable Data:


Confusion Matrix
152
29.2%
108
20.8%
58.5%
41.5%
132
25.4%
128
24.6%
49.2%
50.8%
53.5%
46.5%
54.2%
45.8%
53.8%
46.2%
1
2
Output Class
1
2
Target Class



Overlapping Data:


Confusion Matrix
94
23.5%
1
0.3%
4
1.0%
1
0.3%
94.0%
6.0%
2
0.5%
93
23.3%
1
0.3%
4
1.0%
93.0%
7.0%
10
2.5%
2
0.5%
88
22.0%
0
0.0%
88.0%
12.0%
9
2.3%
4
1.0%
0
0.0%
87
21.8%
87.0%
13.0%
81.7%
18.3%
93.0%
7.0%
94.6%
5.4%
94.6%
5.4%
90.5%
9.5%
Output Class
1
2
3
4
Target Class
1
2
3
4



For mixture of Gaussians (GMMs) :

- Expectation Maximization method used for estimation of parameters.

- Hyper-parameter Optimization:

  The number of clusters representing the GMM model is manually varied and model is run on the validation data.

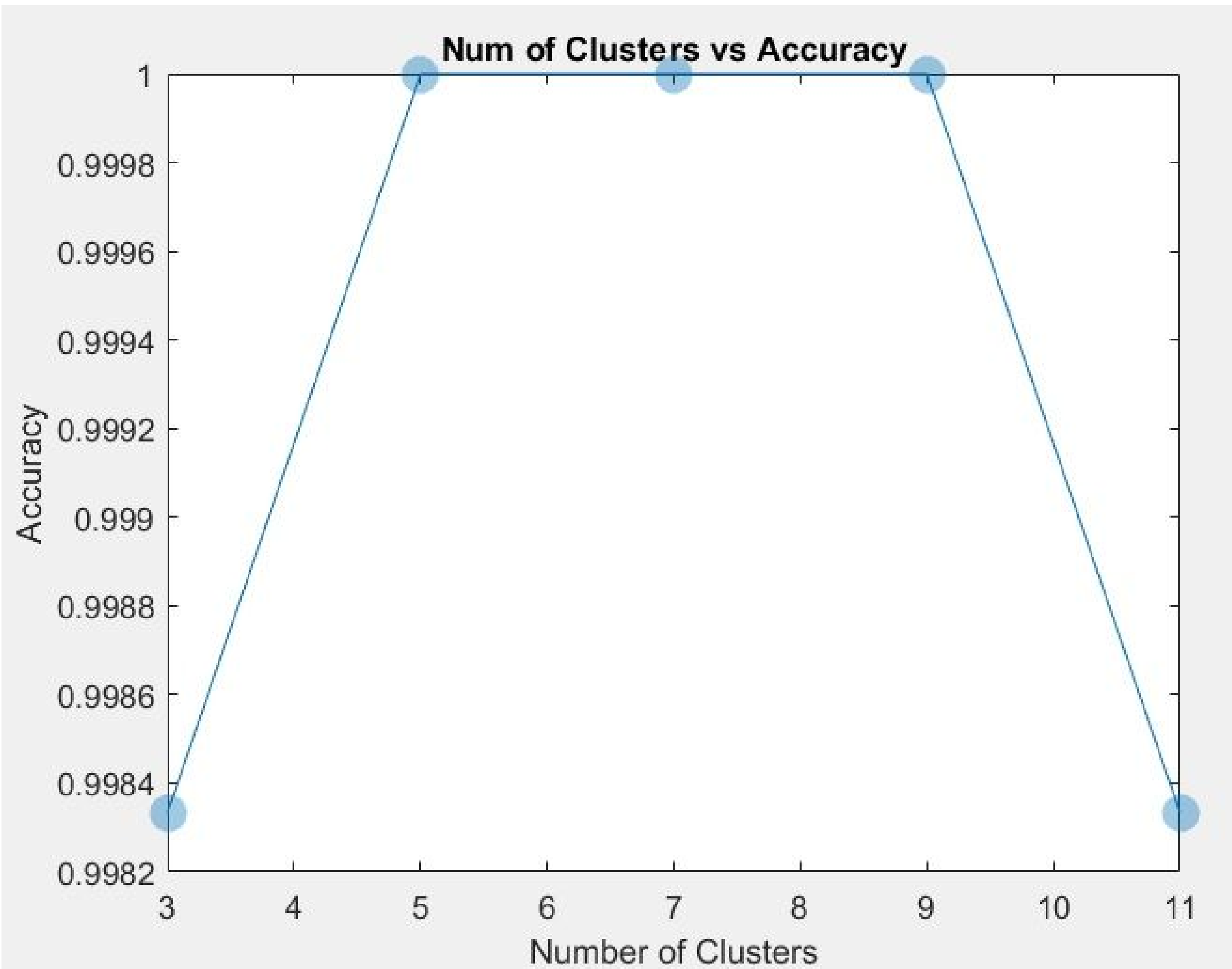
- Discriminant Function

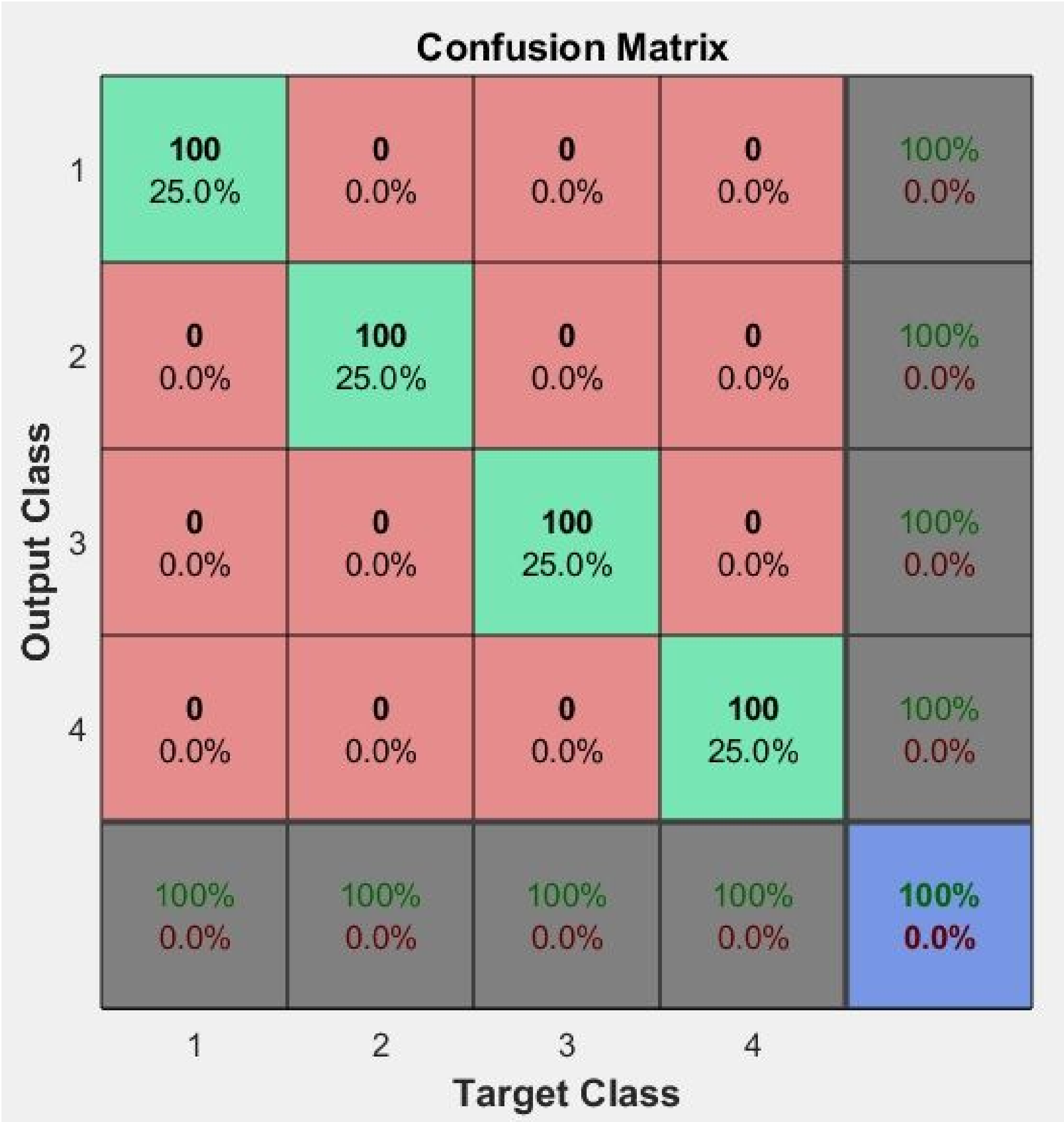
  The calculated likelihood function is the multiplied by the prior probability(Ni/N) to obtain posterior probability.
- Classifier : argmax(Posterior Probability) and thereby assigned to specific class.

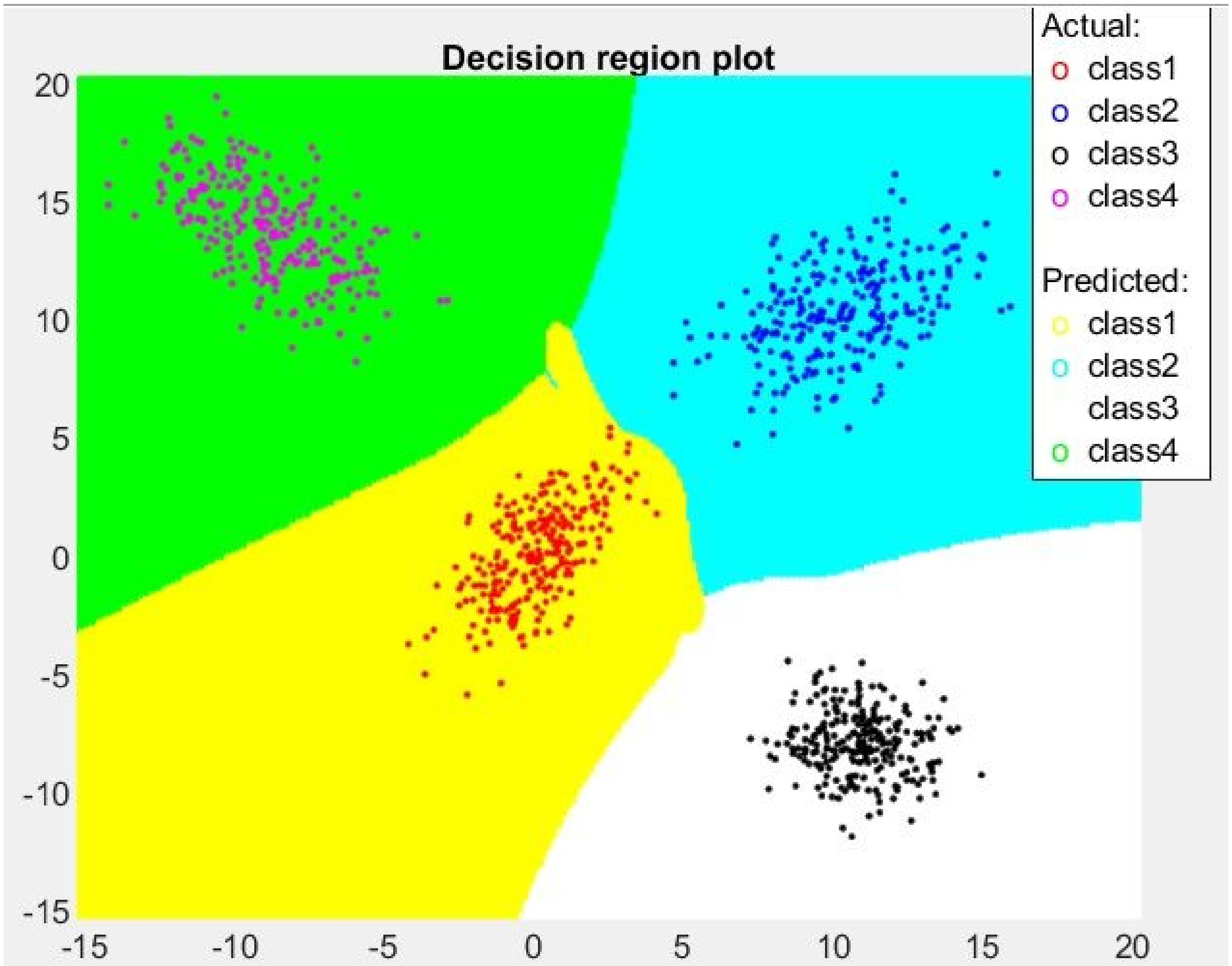
Bayes Classifier (GMM)

Different Covariance Matrix:

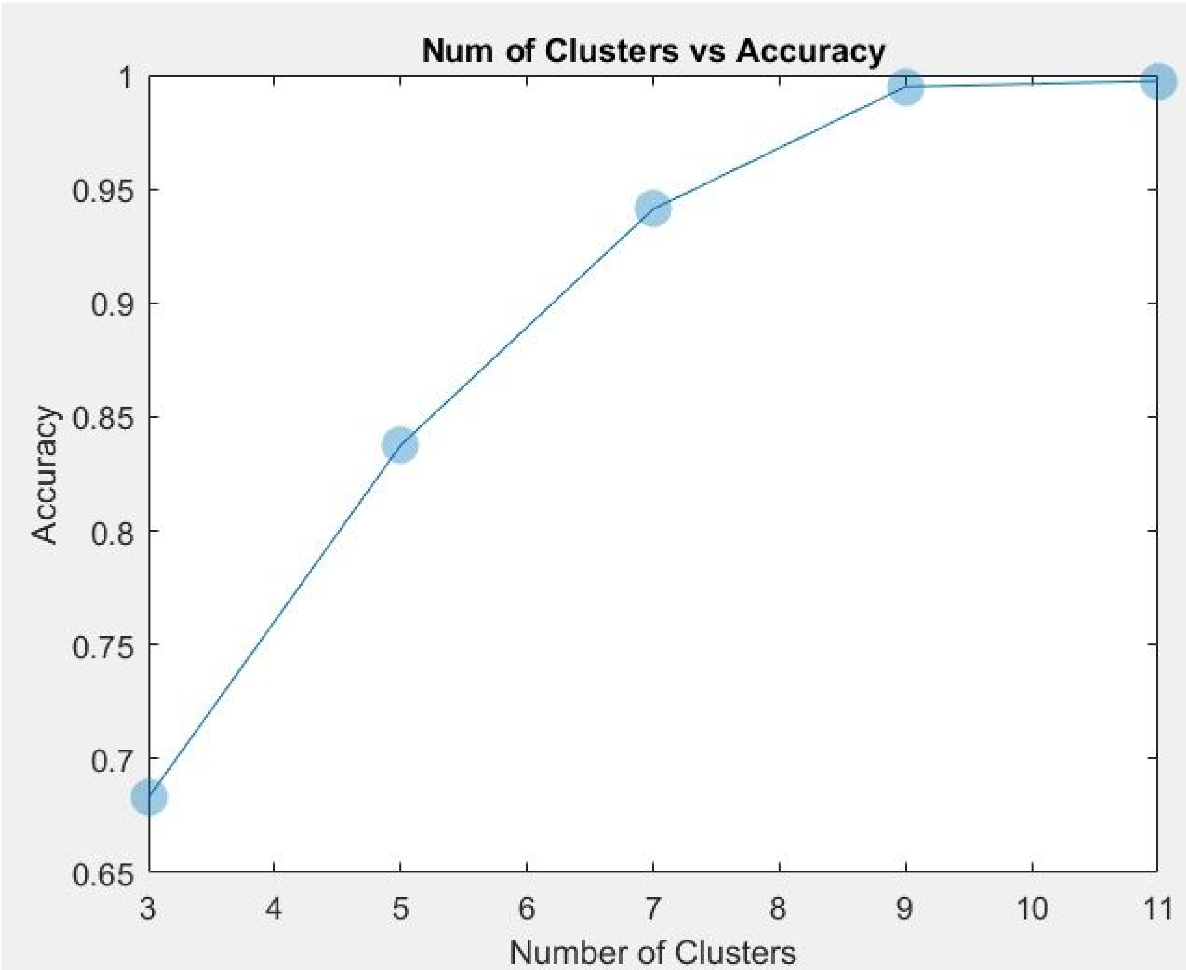
Linearly Separable Data:

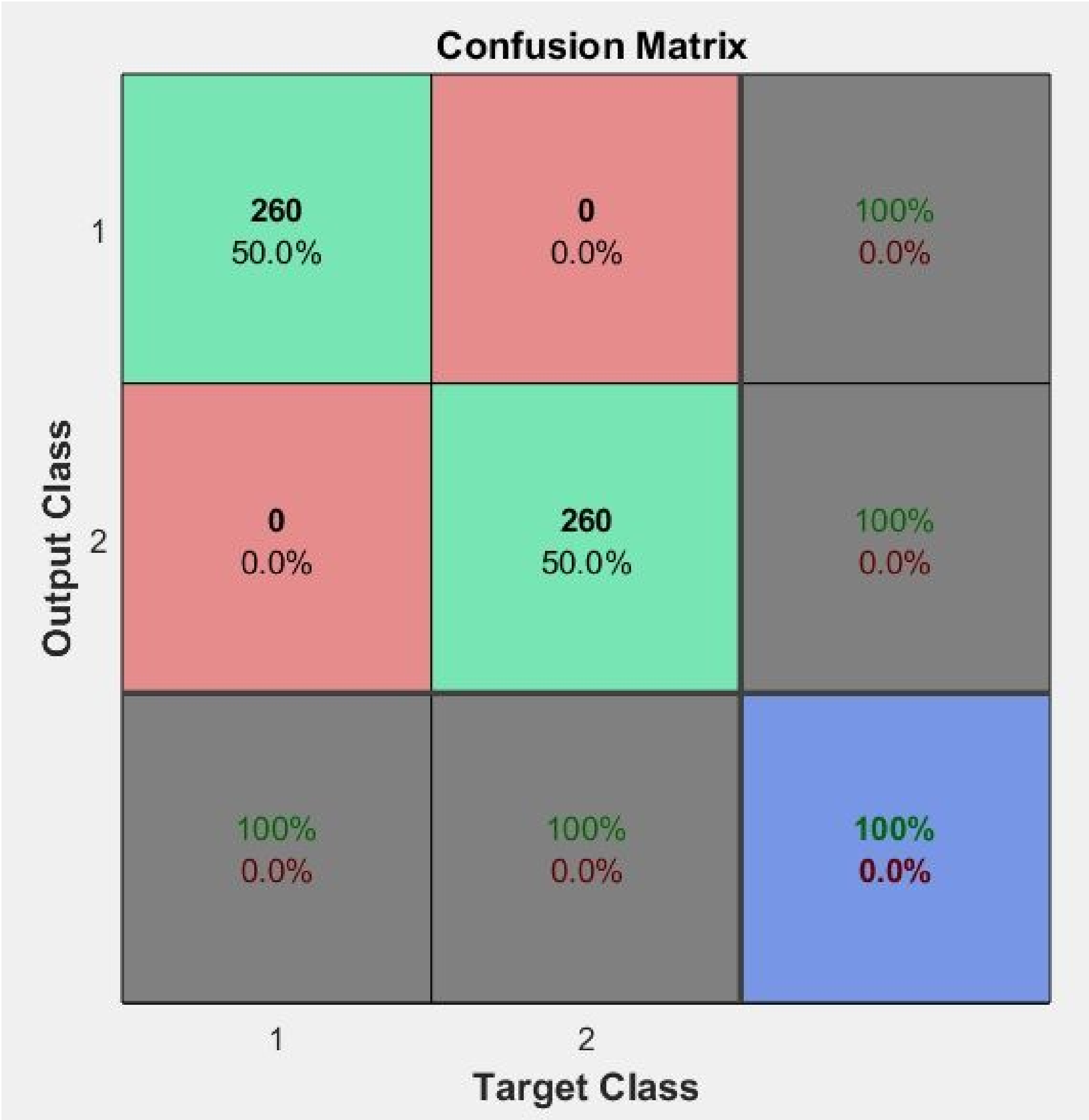


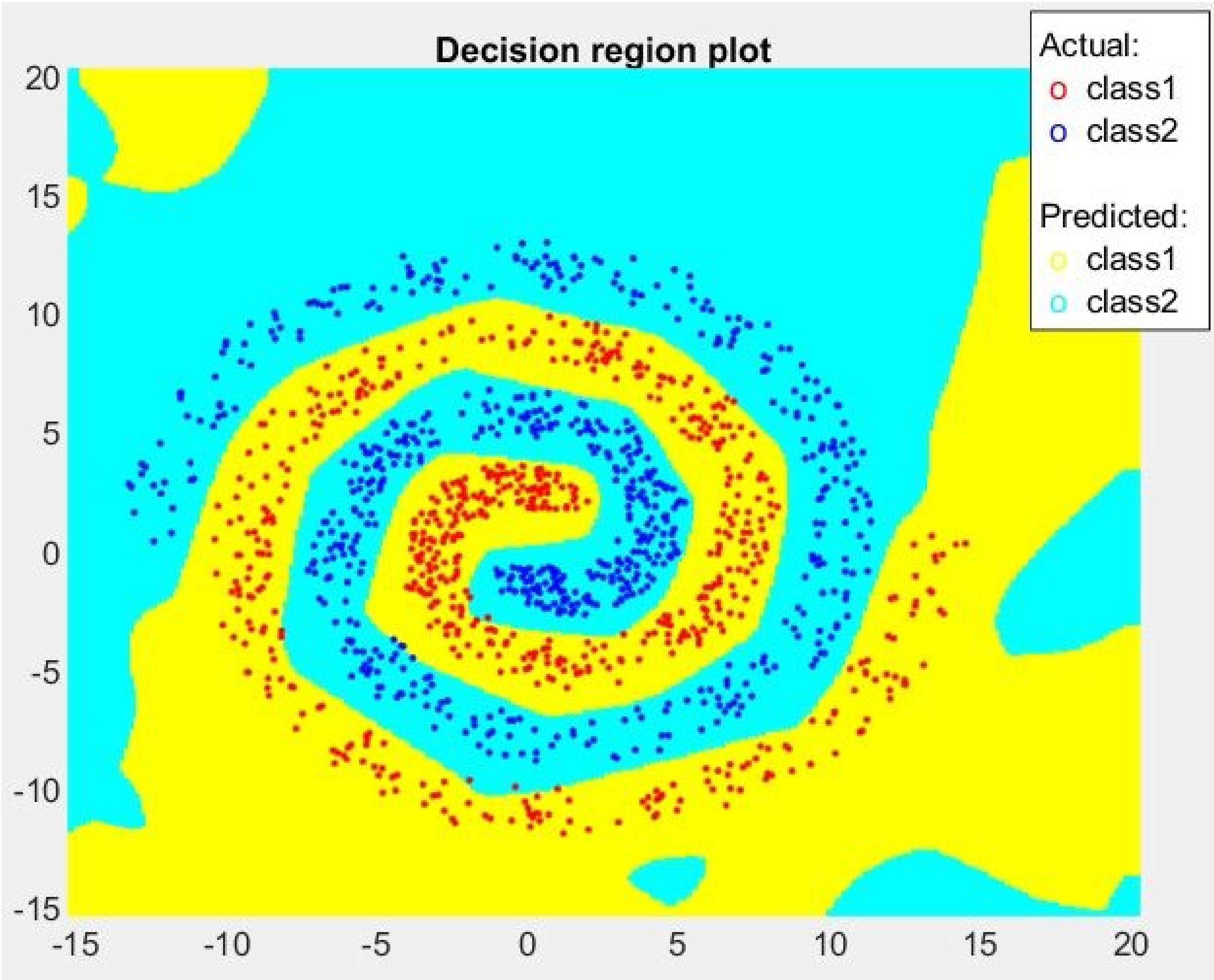

Confusion Matrix
Output Class
Target Class
1
2
3
4
100
25.0%
0
0.0%
100%
0.0%



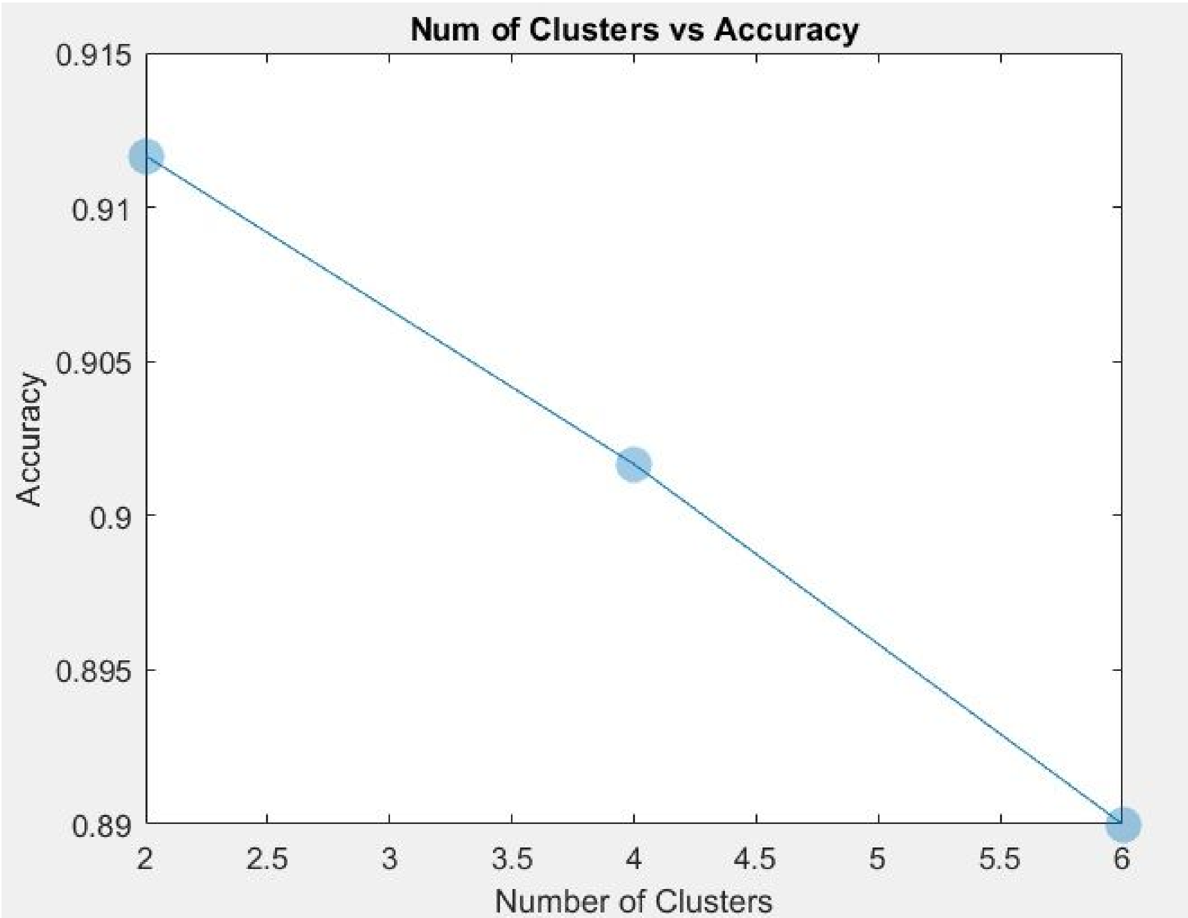
Non Linearly Separable Data:

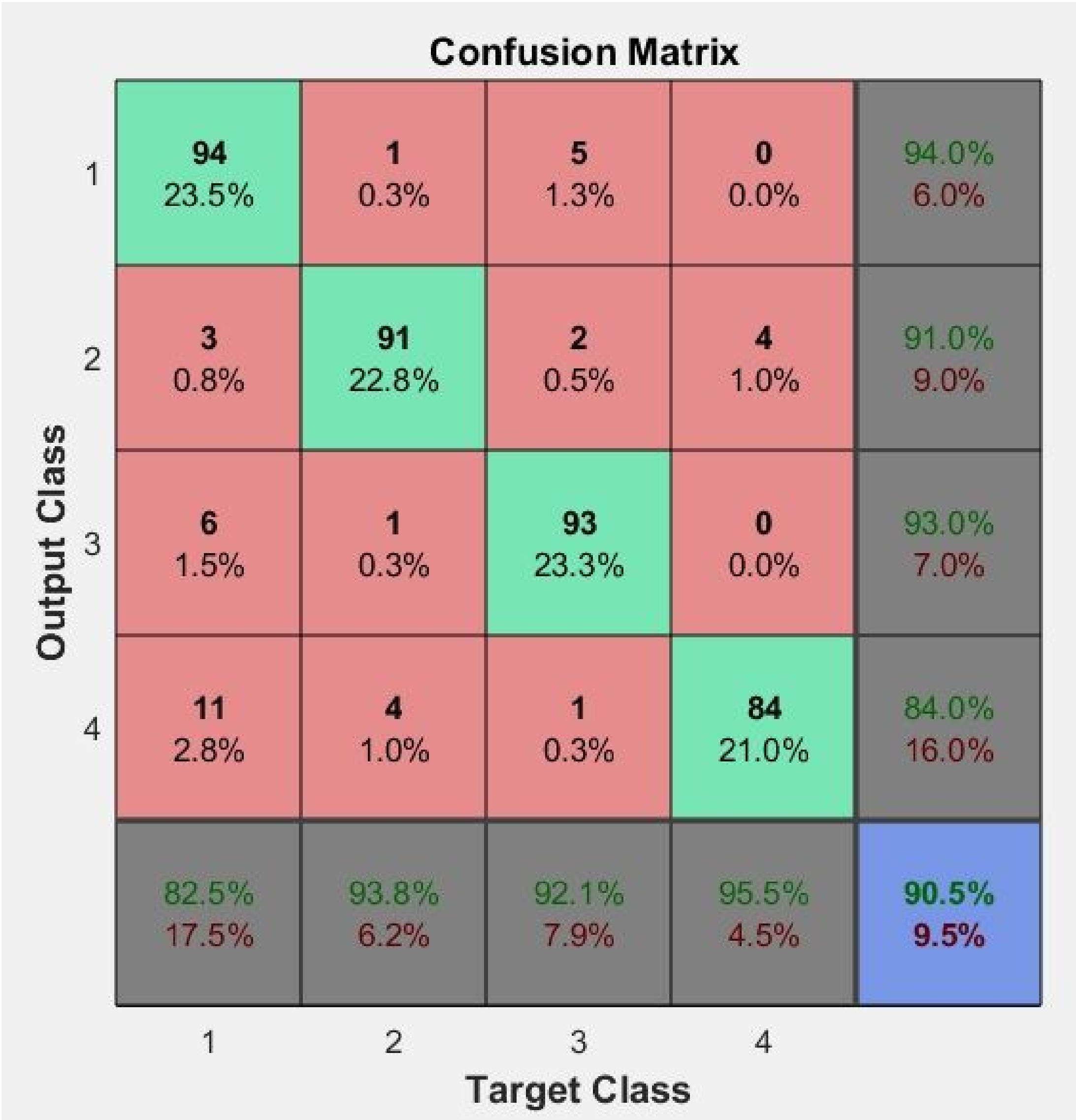

Num of Clusters vs Accuracy
Accuracy
Number of Clusters
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
3
4
5
6
7
8
9
10
11


Confusion Matrix
260
50.0%
0
0.0%
100%
0.0%
0
0.0%
260
50.0%
100%
0.0%
100%
0.0%
100%
0.0%
100%
0.0%
1
2
Output Class
1
2
Target Class



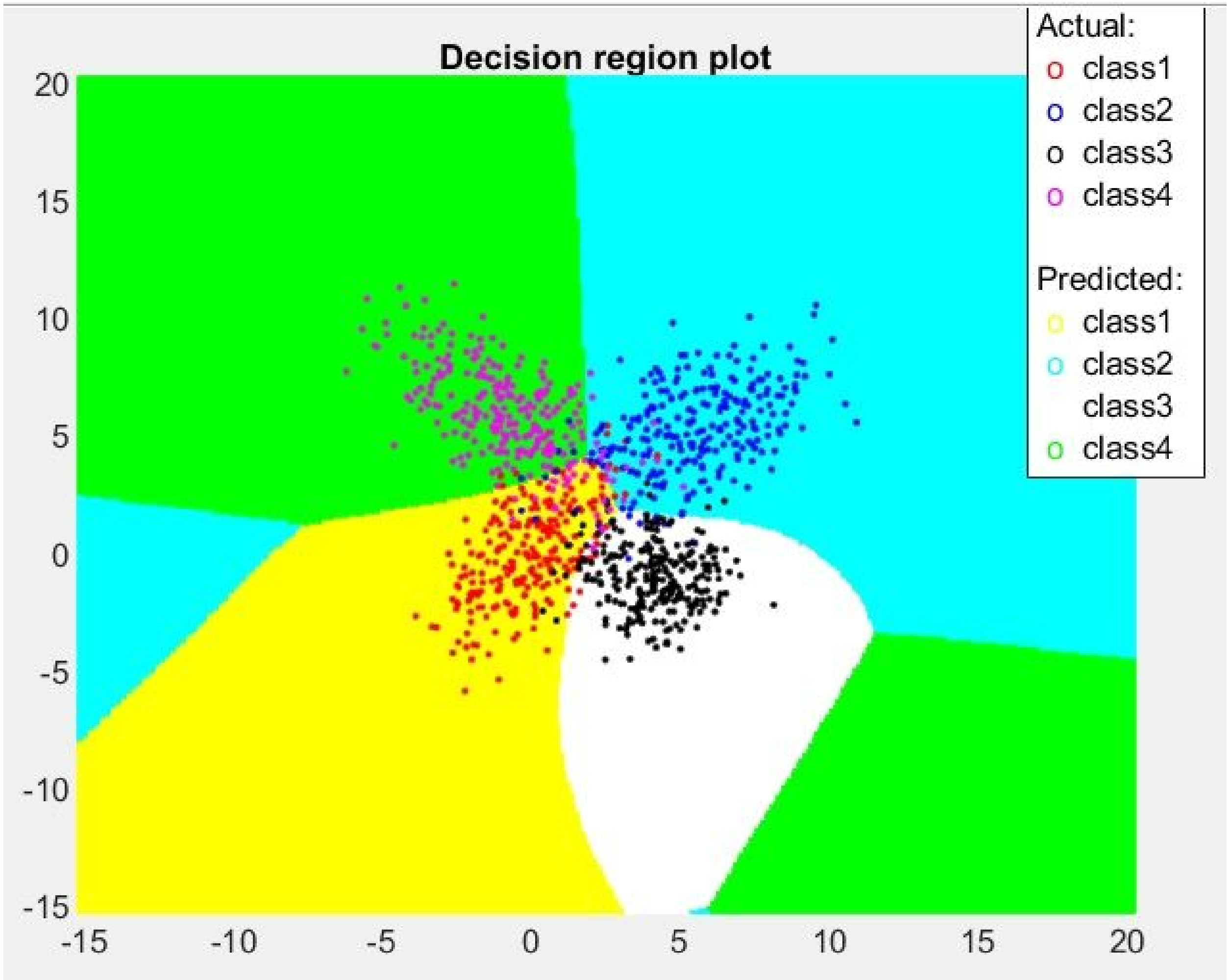
Overlapping Data:


Num of Clusters vs Accuracy
0.915
0.91
0.905
0.9
0.895
0.89
Accuracy
2
2.5
3
3.5
4
4.5
5
5.5
6
Number of Clusters



Confusion Matrix
Output Class
Target Class
1
2
3
4
94
23.5%
1
0.3%
5
1.3%
0
0.0%
94.0%
6.0%
3
0.8%
91
22.8%
2
0.5%
4
1.0%
91.0%
9.0%
6
1.5%
1
0.3%
93
23.3%
0
0.0%
93.0%
7.0%
11
2.8%
4
1.0%
1
0.3%
84
21.0%
84.0%
16.0%
82.5%
17.5%
93.8%
6.2%
92.1%
7.9%
95.5%
4.5%
90.5%
9.5%

Performance on Real world Data-



Sets :

Before performing calculations on the real world data set, we standardise the data.

Standardisation of data:

If we are given n data points with d features each, then for every feature, we calculate the Mean(μ) and Sigma (sqrt(variance)) and transform each data point using the formula:

X = X- μ / Sigma .

- Data Fitting

  Underlying Probability Distribution assumed to be mixture of Gaussians.

- Hyper-parameter Optimization:
  The number of clusters representing the GMM model is manually varied and model is run on the validation data.

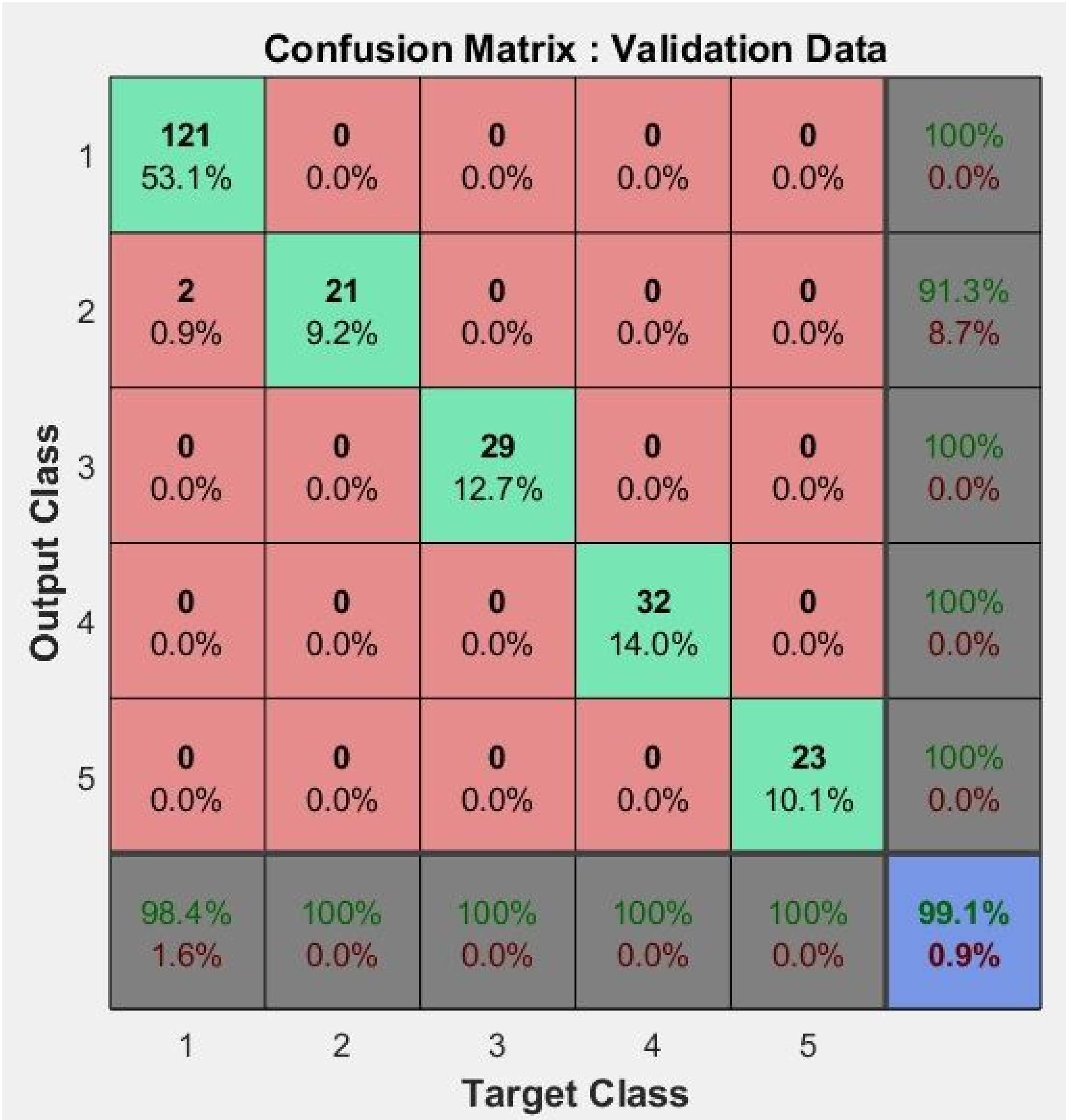
- Classification:
  Compute the probability of the test data being generated by each of the GMMs

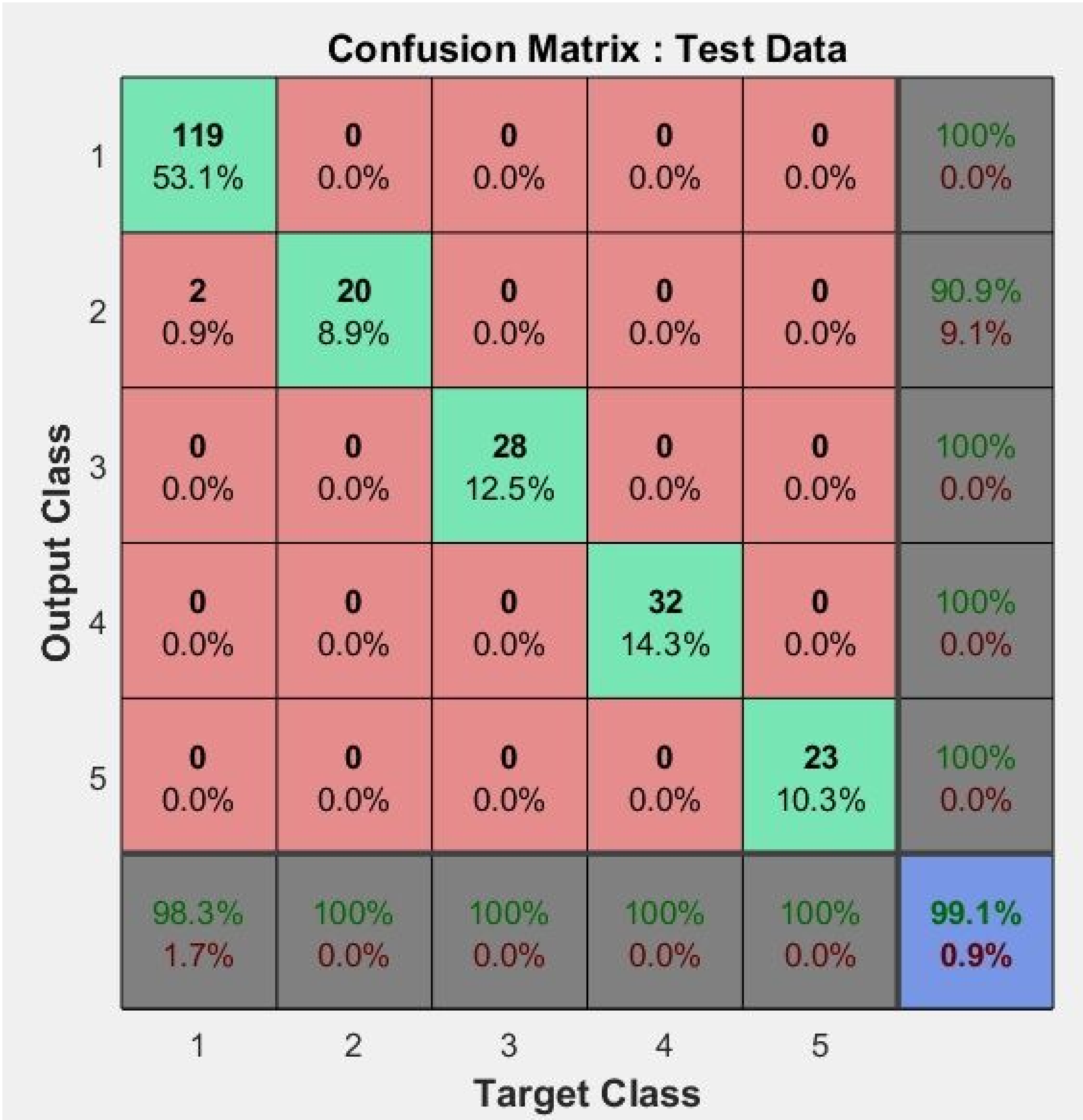
Image classification data-set:

Bayes Classifier:

(Assuming probability distributions to be Gaussian)

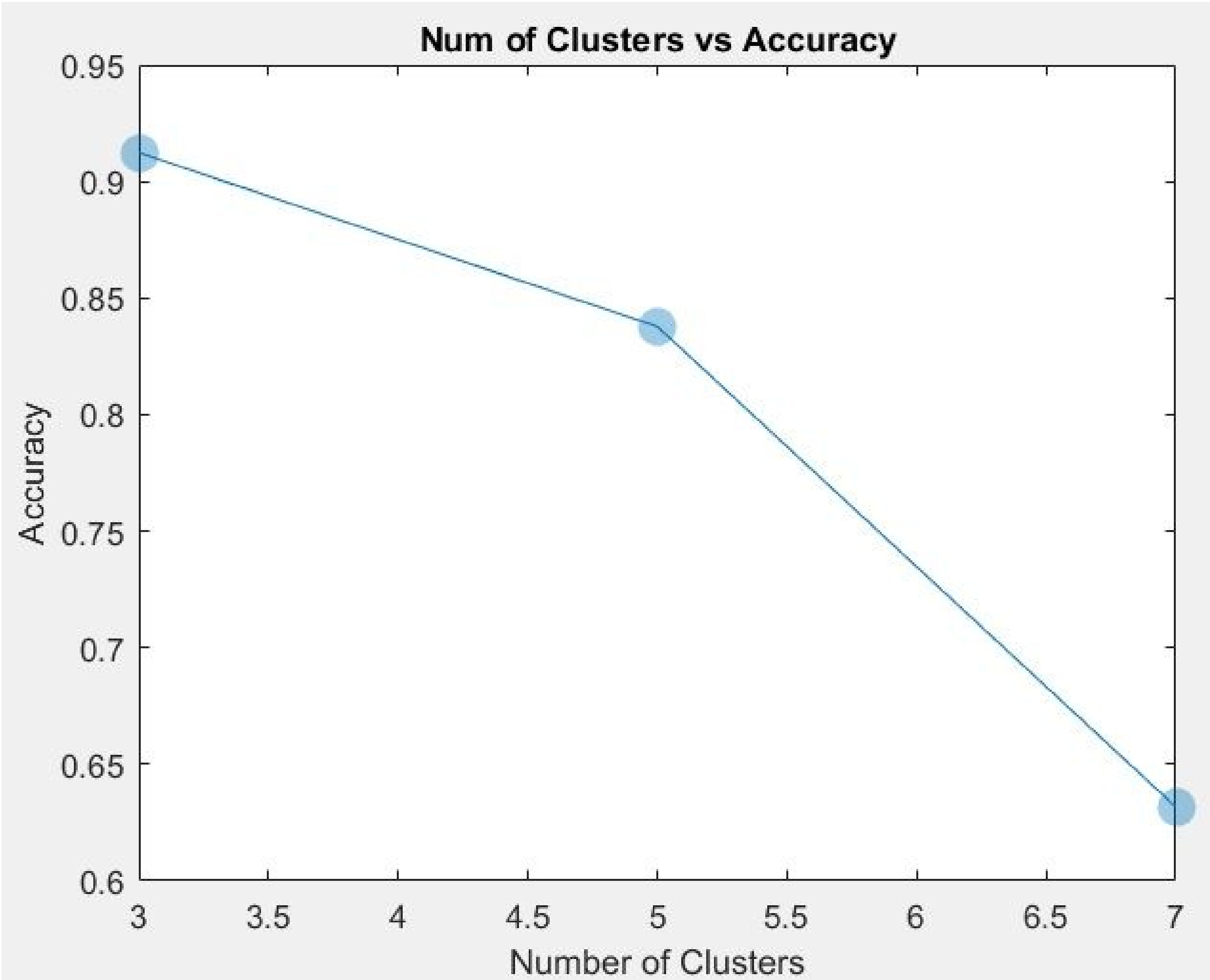
We first use validation data for choosing the ideal random sampling of the given data and then build our classifier based on that particular sampling.

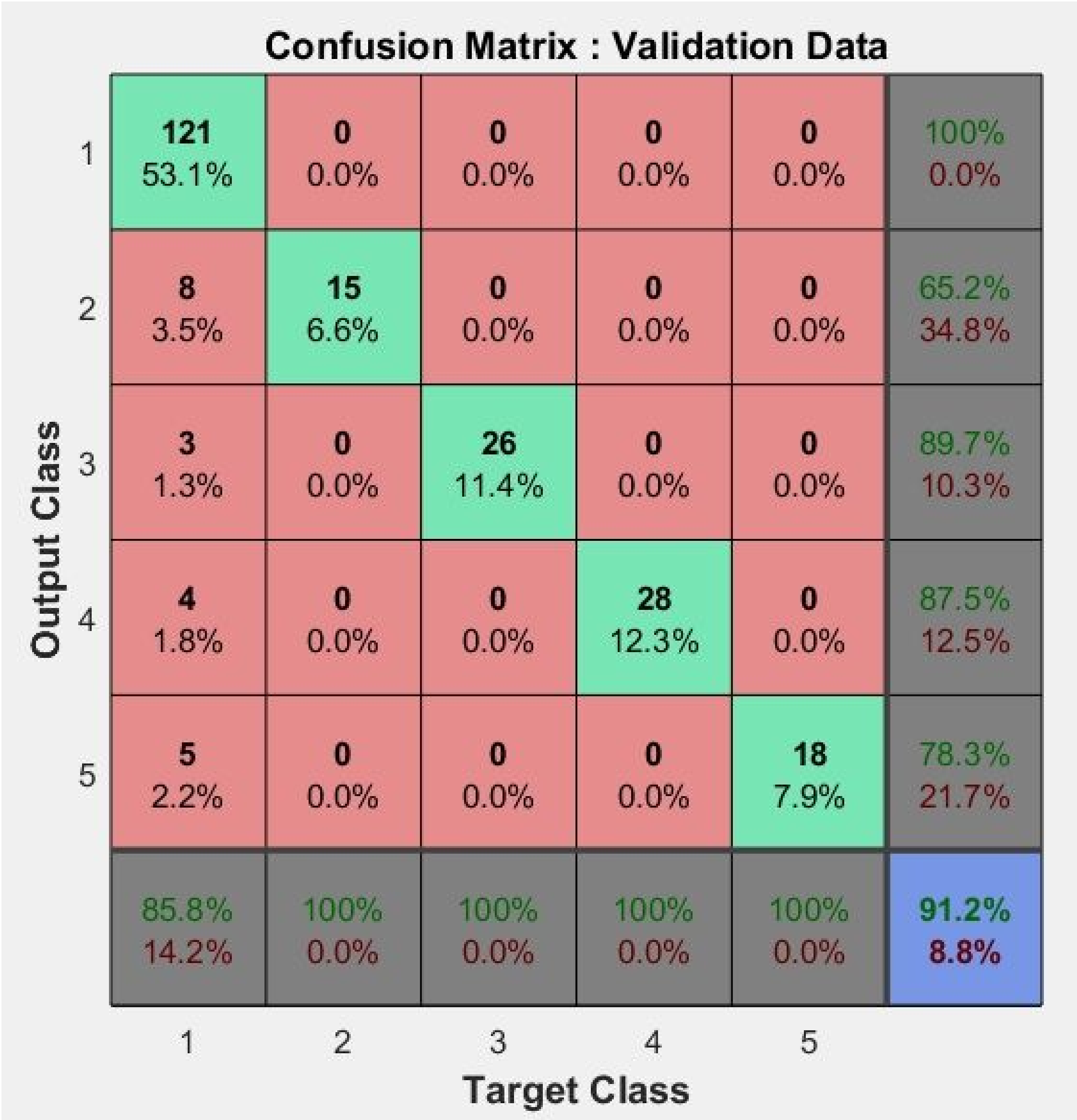

Confusion Matrix : Validation Data
121
53.1%
0
0.0%
100%
0.0%
2
0.9%
21
9.2%
91.3%
8.7%
29
12.7%
32
14.0%
23
10.1%
98.4%
1.6%
99.1%
0.9%
Output Class
Target Class
1
2
3
4
5

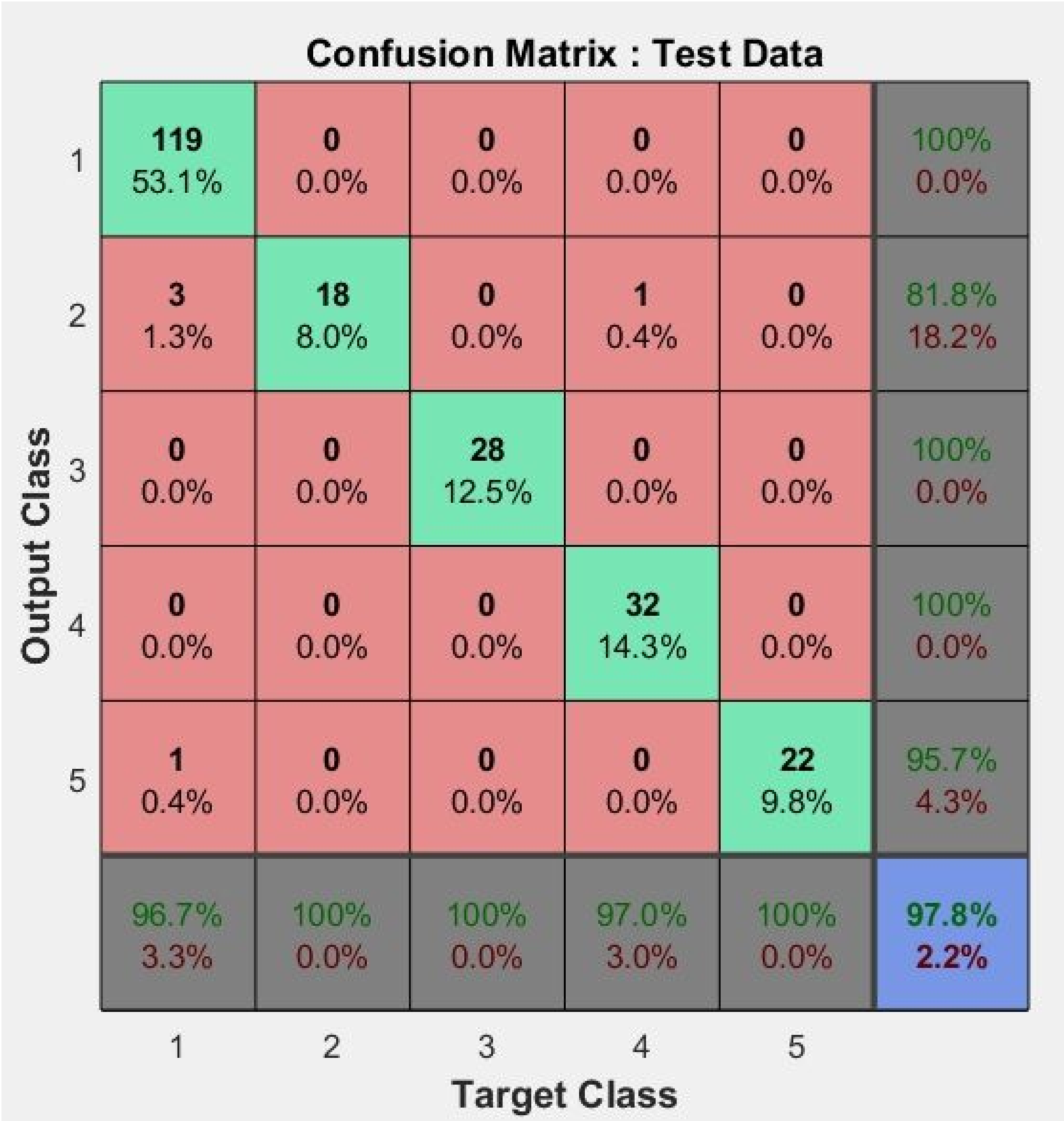


Bayes Classifier :

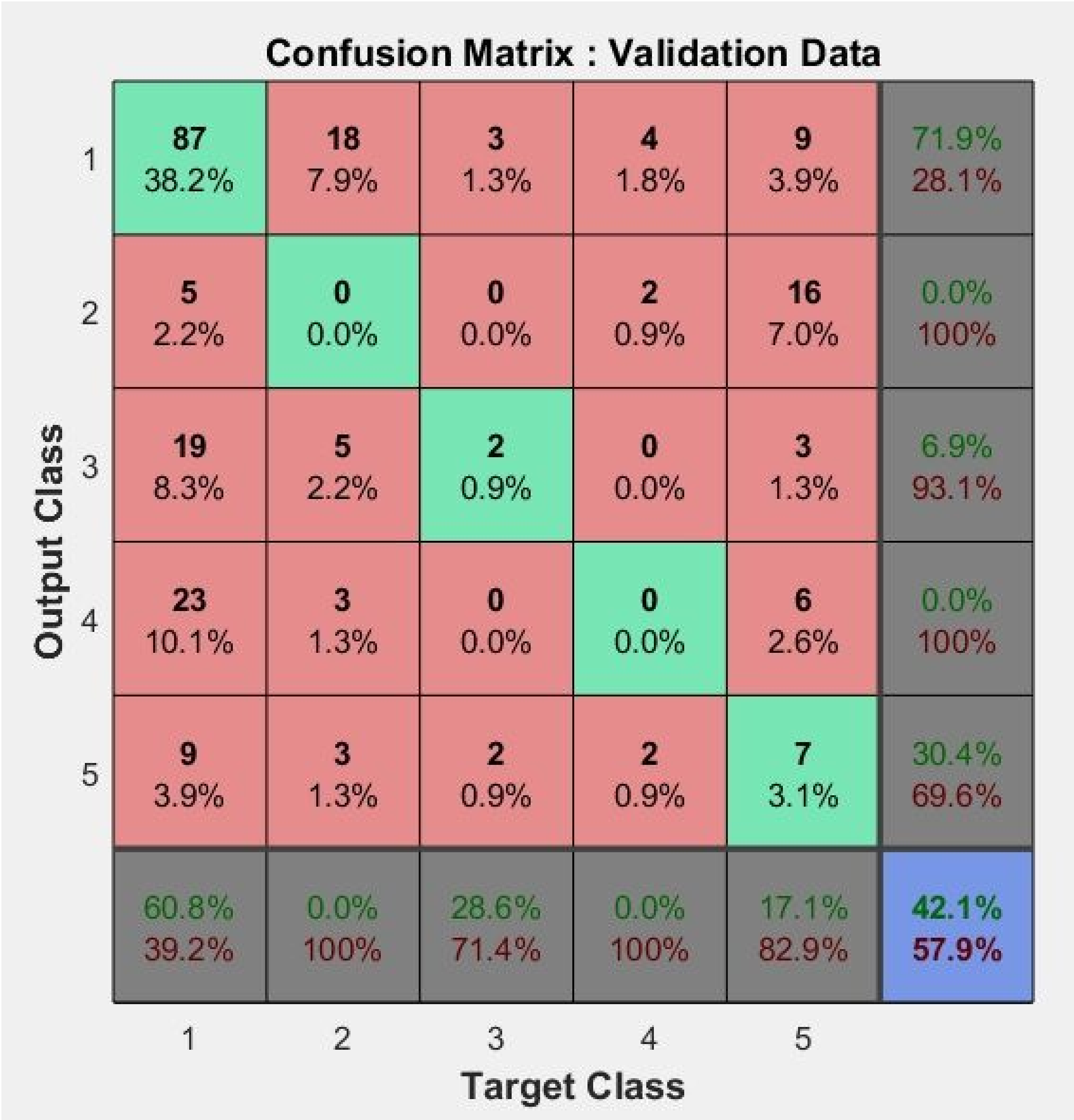
(Assuming probability distribution to be mixture of GMMs.)

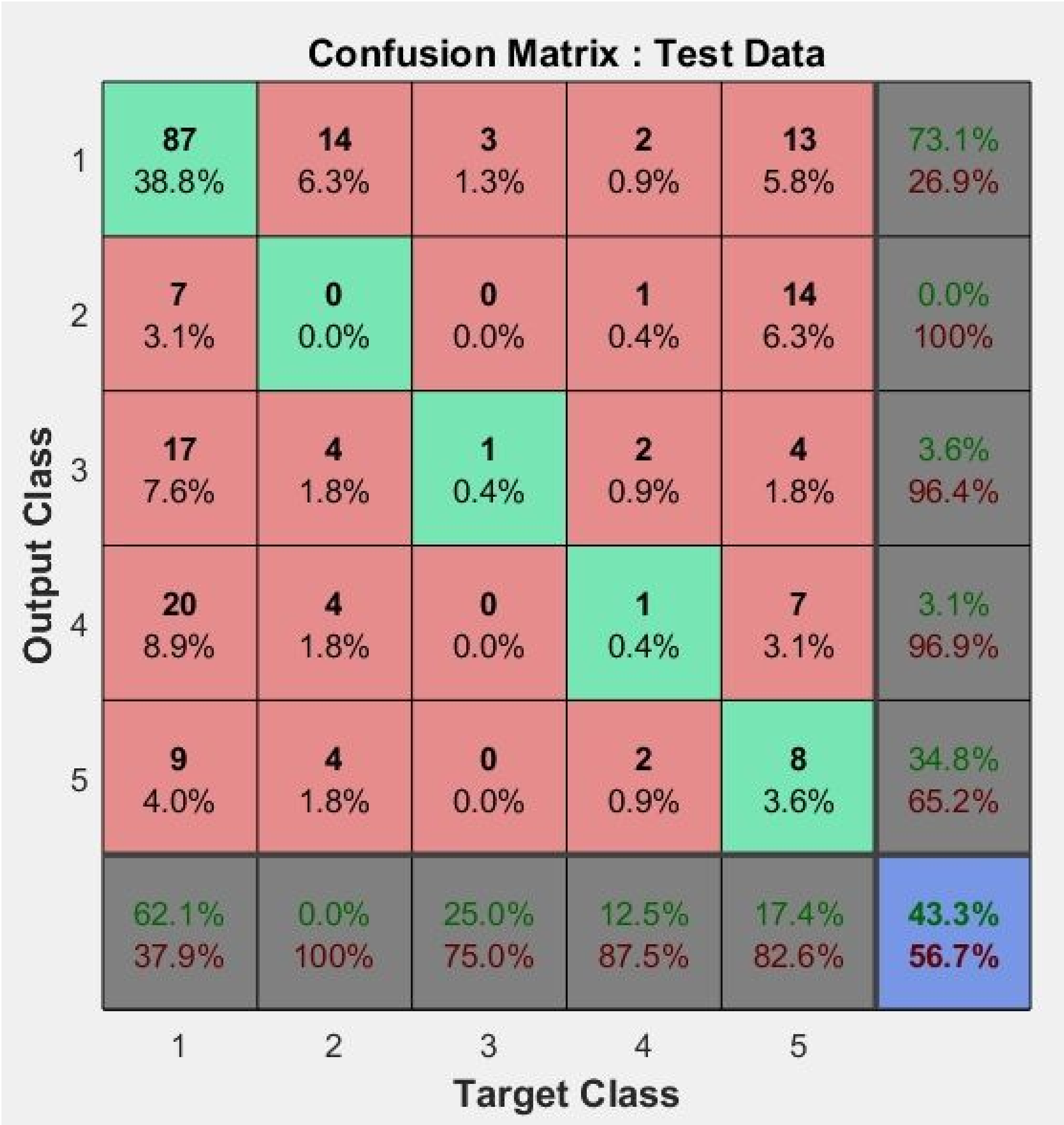

Num of Clusters vs Accuracy
Accuracy
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
3
3.5
4
4.5
5
5.5
6
6.5
7
Number of Clusters


Confusion Matrix : Validation Data
Output Class
Target Class
1
2
3
4
5
121
53.1%
0
0.0%
100%
0.0%
8
3.5%
15
6.6%
65.2%
34.8%
3
1.3%
26
11.4%
89.7%
10.3%
4
1.8%
28
12.3%
87.5%
12.5%
5
2.2%
18
7.9%
78.3%
21.7%
85.8%
14.2%
100%
0.0%
91.2%
8.8%



Naive Bayes Classifier:


Confusion Matrix : Validation Data
87
38.2%
18
7.9%
3
1.3%
4
1.8%
9
3.9%
71.9%
28.1%
5
2.2%
0
0.0%
0
0.0%
2
0.9%
16
7.0%
0.0%
100%
19
8.3%
5
2.2%
2
0.9%
0
0.0%
3
1.3%
6.9%
93.1%
23
10.1%
3
1.3%
0
0.0%
0
0.0%
6
2.6%
0.0%
100%
9
3.9%
3
1.3%
2
0.9%
2
0.9%
7
3.1%
30.4%
69.6%
60.8%
39.2%
0.0%
100%
28.6%
71.4%
0.0%
100%
17.1%
82.9%
42.1%
57.9%
1
2
3
4
5
Output Class
1
2
3
4
5
Target Class



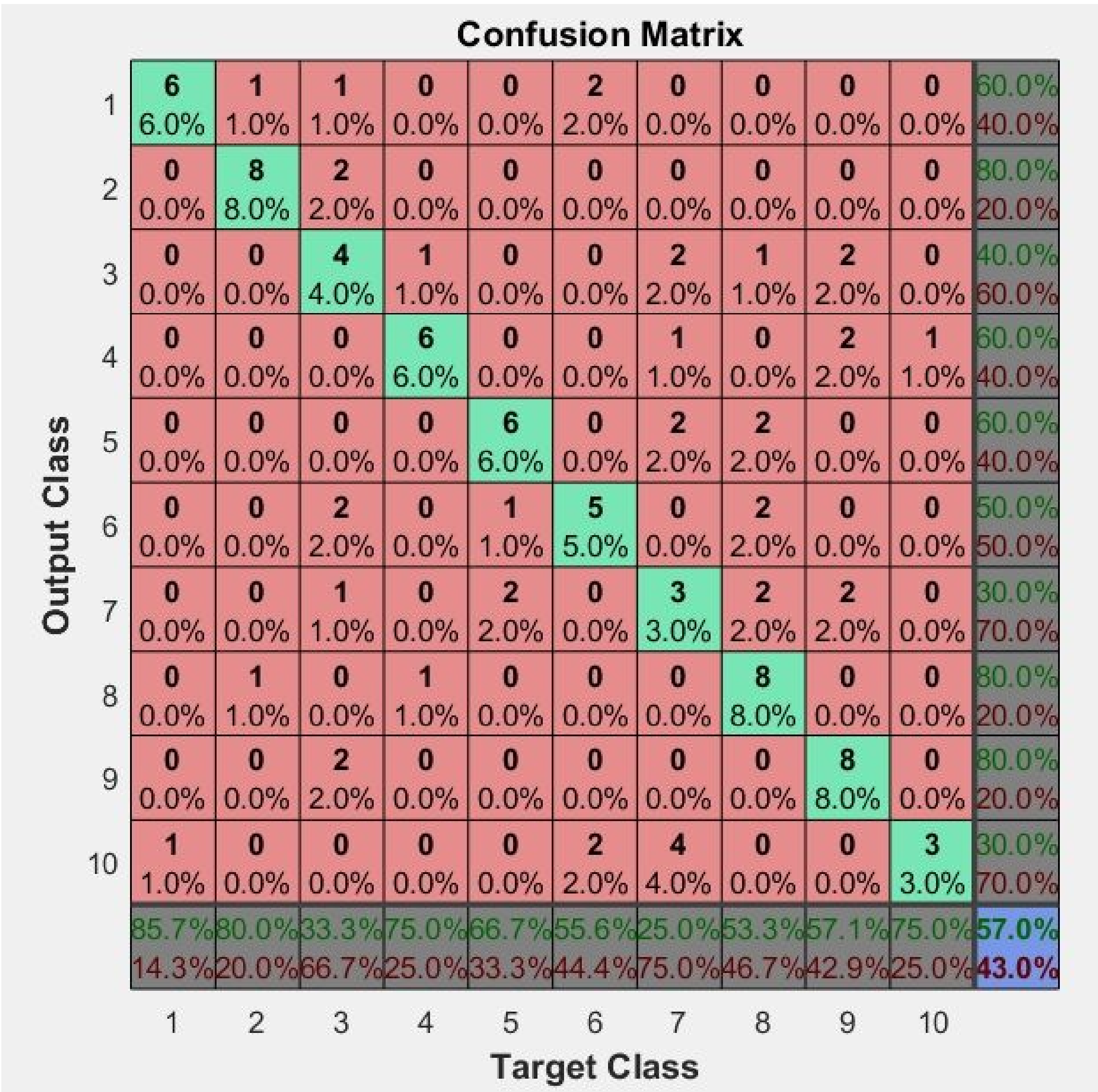
Speaker identification and verification data-set (Varying length Pattern classification)

- Classifier : maximum probability as the identity of the person

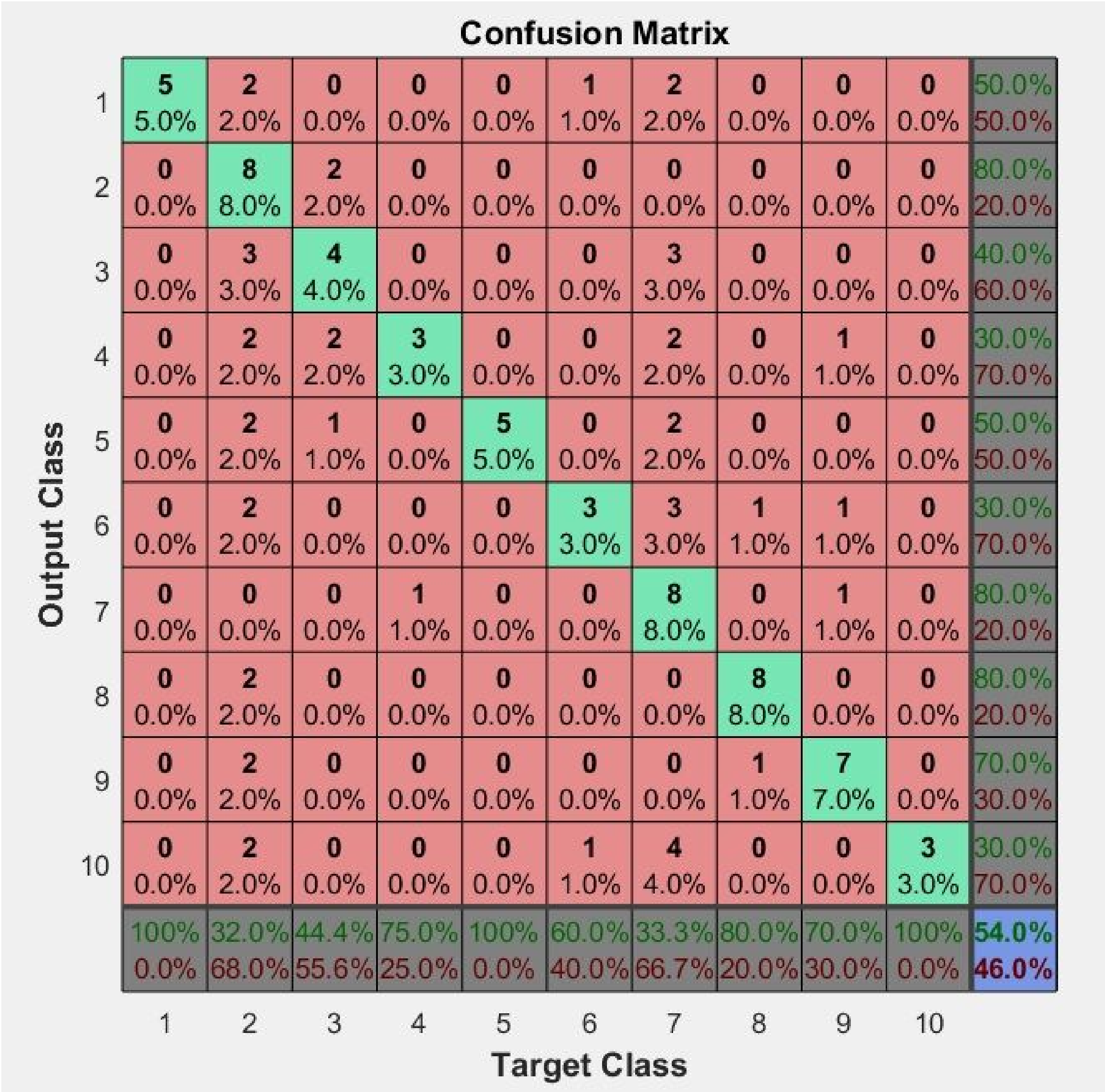
$$p(X \mid \lambda_m) = \prod p(x_t \mid \lambda_m) = \prod \sum \pi_{mq} N(x_t \mid \mu_{mq}, C_{mq})$$

<u>Verification:</u>

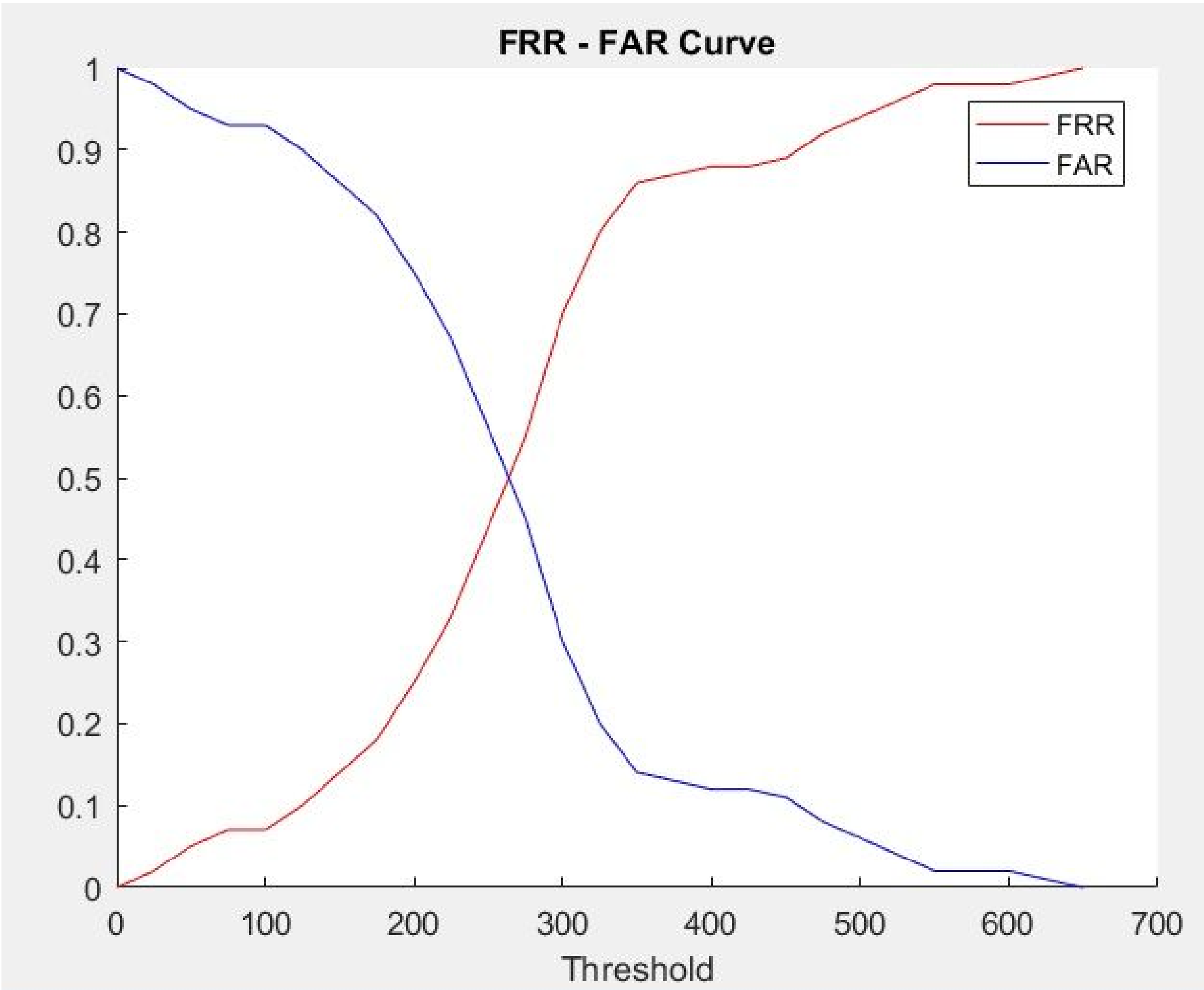
Bayes Classifier:



Bayes Classifier (GMM)



Identification:



Hidden Markov Models :

- It is a Markov Model in which system being studied is assumed to be a Markov process with hidden states .
- The hidden nature of the HMM are due to the unknown nature of the state sequence that the model passes through which is the difference between the simpler Markov models where the states are directly visible to the observer .

Word Recognition Problem :

- Hyper-parameter : No.of clusters and No. of States
- The Results shown below are for No. of states = 5(or even 6,9)
- No. of clusters = 5

(Validation data consisting 101.htk to 130.htk)

hmm1 corresponding to word five
hmm2 corresponding to word two
hmm3 corresponding to word three
hmm4 corresponding to word four

```
priti@priti-HP-Notebook: ~/Desktop/htk_tut
 FATAL ERROR - Terminating program HVite
priti@priti-HP-Notebook:~/Desktop/htk_tut$ HVite -C config_HEREST -S valid_ALL.s
cp -H hmmALL  -i result.mlf  -w wNet  s-dict s-list
priti@priti-HP-Notebook:~/Desktop/htk_tut$ HResults -p -I Valid_GT.mlf  s-list r
esult.mlf
====================== HTK Results Analysis =======================
  Date: Thu Oct 27 15:02:16 2016
  Ref : Valid_GT.mlf
  Rec : result.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=100.00 [H=120, S=0, N=120]
WORD: %Corr=100.00, Acc=100.00 [H=120, D=0, S=0, I=0, N=120]
------------------------ Confusion Matrix -------------------------
       h   h   h   h
       m   m   m   m
       m   m   m   m
       1   2   3   4  Del [ %c / %e]
hmm1  30   0   0   0    0
hmm2   0  30   0   0    0
hmm3   0   0  30   0    0
hmm4   0   0   0  30    0
Ins    0   0   0   0
===================================================================
priti@priti-HP-Notebook:~/Desktop/htk_tut$
```

The results on the test data
(test data consisting from 131.htk to 160.htk)

1. priti@priti-HP-Notebook: ~/Desktop/htk_tut

```
===================================================================
priti@priti-HP-Notebook:~/Desktop/htk_tut$ HVite -C config_HEREST -S TEST_ALL.sc
p -H hmmALL  -i result.mlf  -w wNet  s-dict s-list
priti@priti-HP-Notebook:~/Desktop/htk_tut$ HResults -p -I TEST_GT.mlf  s-list re
sult.mlf
====================== HTK Results Analysis =======================
  Date: Thu Oct 27 15:03:18 2016
  Ref : TEST_GT.mlf
  Rec : result.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=100.00 [H=120, S=0, N=120]
WORD: %Corr=100.00, Acc=100.00 [H=120, D=0, S=0, I=0, N=120]
------------------------ Confusion Matrix -------------------------
       h   h   h   h
       m   m   m   m
       m   m   m   m
       1   2   3   4  Del [ %c / %e]
hmm1  30   0   0   0    0
hmm2   0  30   0   0    0
hmm3   0   0  30   0    0
hmm4   0   0   0  30    0
Ins    0   0   0   0
===================================================================
priti@priti-HP-Notebook:~/Desktop/htk_tut$
```

Trajectory Dataset
Number of States : 5
Number of Clusters : 5

Validation Data Results:

```
priti@priti-HP-Notebook: ~/Desktop/htk_tut
Ins     0   0   0
===================================================================
priti@priti-HP-Notebook:~/Desktop/htk_tut$ HVite -C config_HEREST -S validALL.sc
p -H initModels/hmmALL  -i result.mlf  -w wNet  s-dict s-list
priti@priti-HP-Notebook:~/Desktop/htk_tut$ HResults -p -I VALID_GT.mlf  s-list r
esult.mlf
====================== HTK Results Analysis =======================
  Date: Thu Oct 27 18:08:47 2016
  Ref : VALID_GT.mlf
  Rec : result.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=100.00 [H=225, S=0, N=225]
WORD: %Corr=100.00, Acc=100.00 [H=225, D=0, S=0, I=0, N=225]
------------------------ Confusion Matrix -------------------------
       h   h   h
       m   m   m
       m   m   m
       1   2   3  Del [ %c / %e]
hmm1  75   0   0    0
hmm2   0  75   0    0
hmm3   0   0  75    0
Ins    0   0   0
===================================================================
priti@priti-HP-Notebook:~/Desktop/htk_tut$
```

Test Data results:

```
priti@priti-HP-Notebook: ~/Desktop/htk_tut
===================================================================
priti@priti-HP-Notebook:~/Desktop/htk_tut$ HParse s-gram wNet
priti@priti-HP-Notebook:~/Desktop/htk_tut$ HVite -C config_HEREST -S testALL.scp
 -H initModels/hmmALL  -i result.mlf  -w wNet  s-dict s-list
priti@priti-HP-Notebook:~/Desktop/htk_tut$ HResults -p -I TEST_GT.mlf  s-list re
sult.mlf
====================== HTK Results Analysis =======================
  Date: Thu Oct 27 18:03:31 2016
  Ref : TEST_GT.mlf
  Rec : result.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=100.00 [H=225, S=0, N=225]
WORD: %Corr=100.00, Acc=100.00 [H=225, D=0, S=0, I=0, N=225]
------------------------ Confusion Matrix -------------------------
       h   h   h
       m   m   m
       m   m   m
       1   2   3  Del [ %c / %e]
hmm1  75   0   0    0
hmm2   0  75   0    0
hmm3   0   0  75    0
Ins    0   0   0
===================================================================
priti@priti-HP-Notebook:~/Desktop/htk_tut$
```