Pattern Recognition

Assignment - 1

Bayes Classifier

Hitesh Ramchandani (B15232)
Sagar Gupta (B15233)
Ankit Amrit Raj (B15107)
Group - 8
November 12, 2017

Contents

1 Objective

1. To build Bayes Classifier and classify the following datasets -
   - 2D Dataset - 1 (Artificial)
     - Linearly Separable Dataset
     - Non Linearly Separable Dataset
   - 2D Dataset - 2 (Real World)
2. Plot Decision Region for all pairs of classes.
3. Contour Region Plots for all pairs of classes.
4. Calculate Accuracy, Precision, mean recall, F-measure and Confusion Matrix.

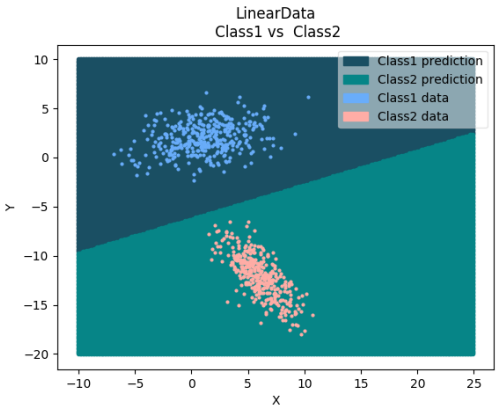
2 Procedure

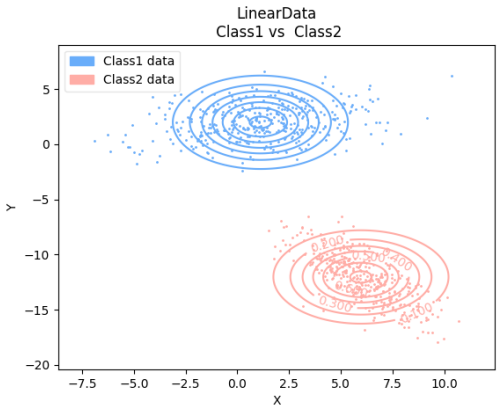
1. Data for each class is partitioned into 75% for training and 25% for testing.
2. The data set for each class is assumed to come from Gaussian distribution.
3. In case 1 ($\Sigma = \sigma^2 I$), mean of covariance matrix for each class was calculated and it's off diagonal terms were assumed to be 0 for further calculations.
4. In case 2 ($\Sigma_i = \Sigma$ for every class), mean of covariance matrix for each class was calculated for further calculations.
5. In case 3 (Σ_i is diagonal matrix), covariance matrix for each class was different and it's off diagonal terms were assumed to be 0 for further calculations.
6. In case 4 (Σ_i is unique), no assumptions were made for further calculations.
7. Based of assumptions, the discriminant function ($g_i(x)$) was calculated for each class and decision region and Contour was plotted.
8. The remaining 25% data was tested for each case and analysis was made.

3 Observations

3.1 Case 1 - $\Sigma = \sigma^2 I$

3.1.1 Linear Data





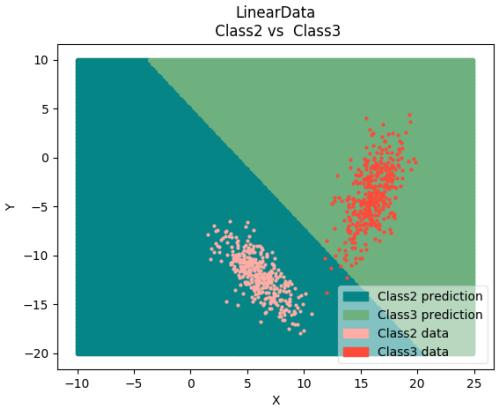
Accuracy = 99.6%

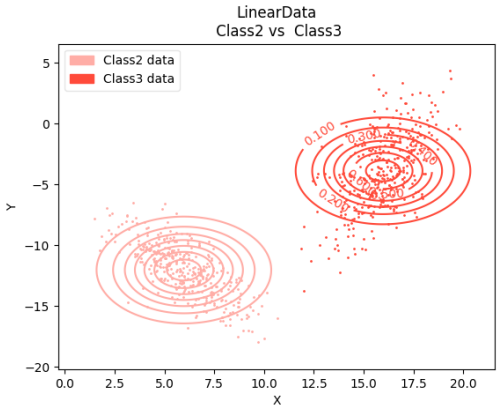
Confusion Matrix

	class1	**class2**
class1	125	0
class2	1	124

Analysis

	class1	**class2**
Precision	1.000	0.992
Recall	0.992	1.000
F-measure	0.996	0.996





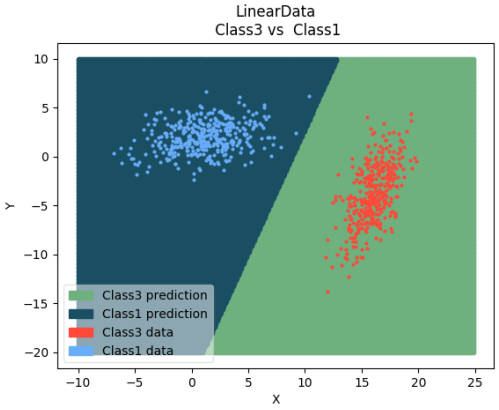
Accuracy = 99.2%

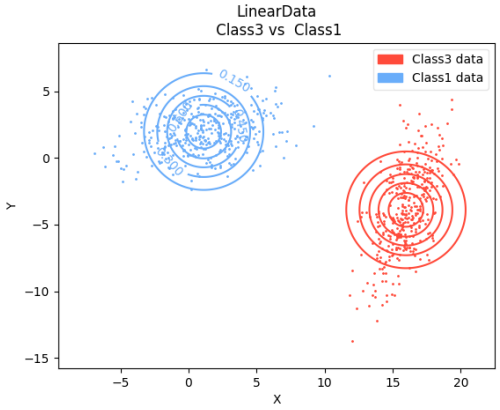
Confusion Matrix

Analysis

	class1	class2
class1	125	0
class2	2	123

	class1	class2
Precision	1.000	0.984
Recall	0.984	1.000
F-measure	0.992	0.992





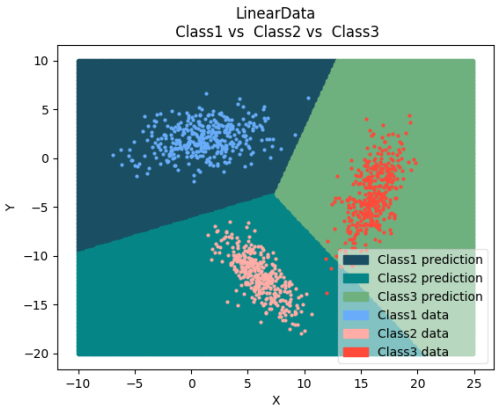
Accuracy = 99.2%

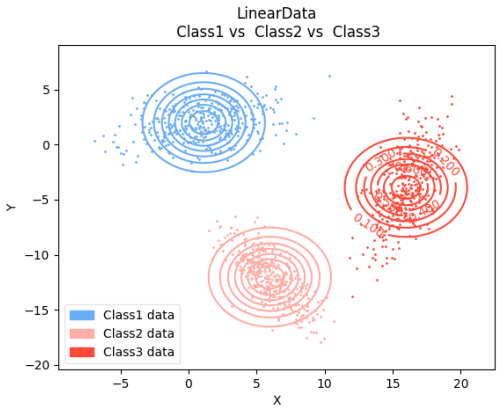
Confusion Matrix

	class1	class2
class1	125	0
class2	2	123

Analysis

	class1	class2
Precision	1.000	0.984
Recall	0.984	1.000
F-measure	0.992	0.992





Accuracy = 98.7%

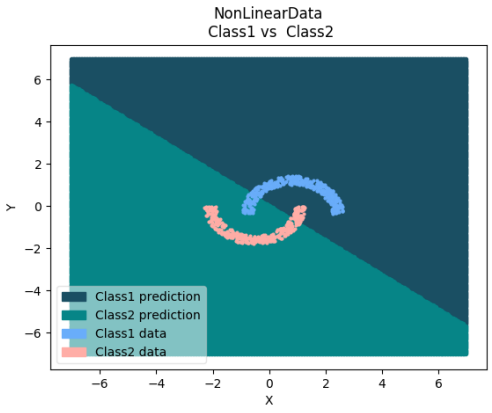
Confusion Matrix

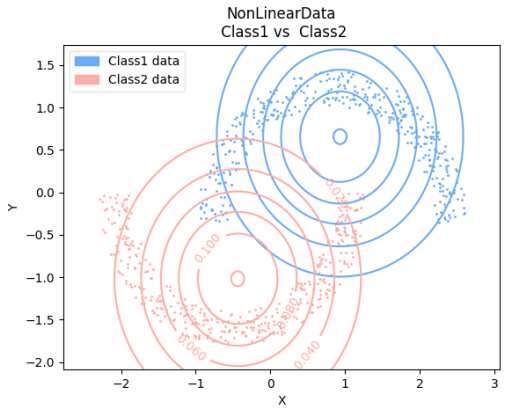
Analysis

	class1	class2	class3
class1	123	0	2
class2	1	124	0
class3	0	2	123

	class1	class2	class3
Precision	0.984	0.992	0.984
Recall	0.992	0.984	0.984
F-measure	0.988	0.988	0.984

3.1.2 NonLinear Data





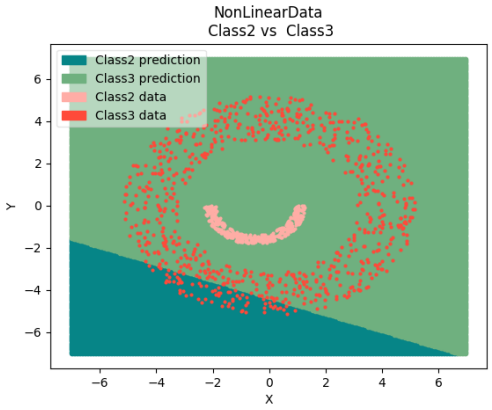
Accuracy = 85.6%

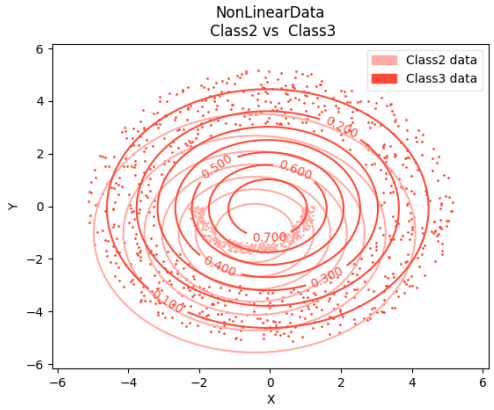
Confusion Matrix

	class1	class2
class1	110	15
class2	21	104

Analysis

	class1	class2
Precision	0.88	0.832
Recall	0.840	0.874
F-measure	0.859	0.852





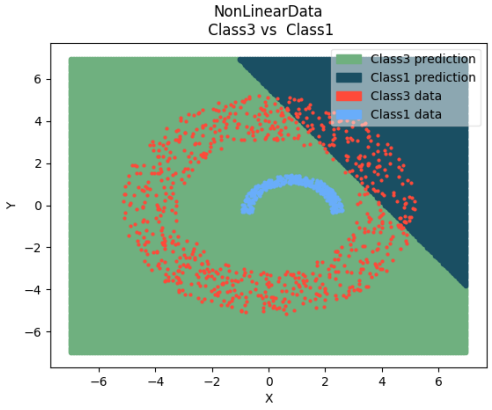
Accuracy = 61.33%

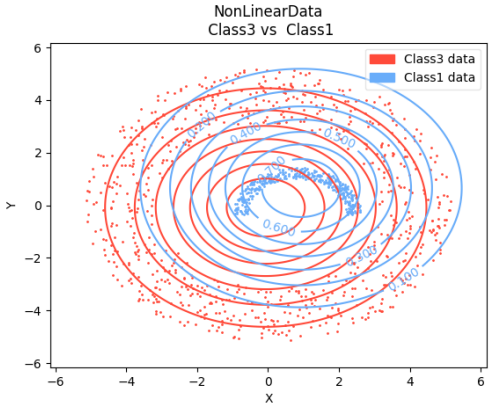
Confusion Matrix

	class1	**class2**
class1	0	125
class2	20	230

Analysis

	class1	**class2**
Precision	0.00	0.92
Recall	0.00	0.647
F-measure	-	0.76





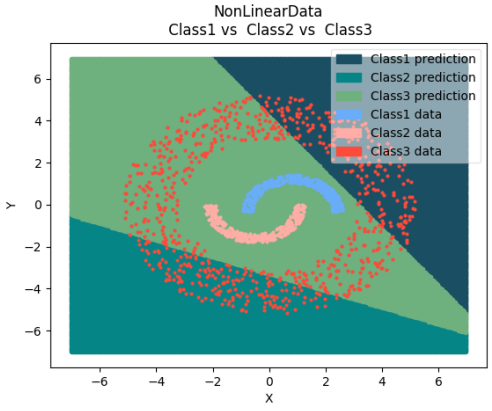
Accuracy = 49.6%

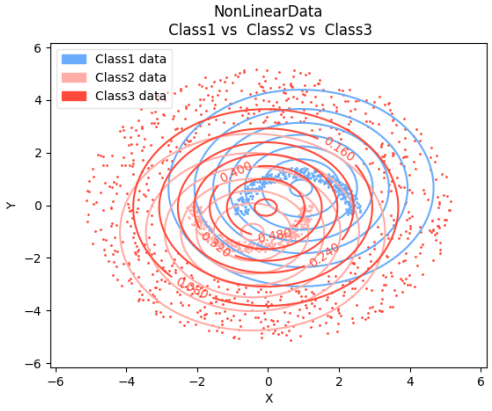
Confusion Matrix

	class1	**class2**
class1	186	64
class2	125	0

Analysis

	class1	**class2**
Precision	0.744	0.00
Recall	0.598	0.00
F-measure	0.663	-





Accuracy = 22.8%

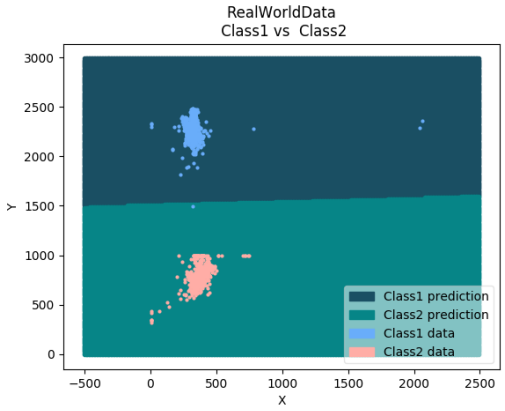
Confusion Matrix

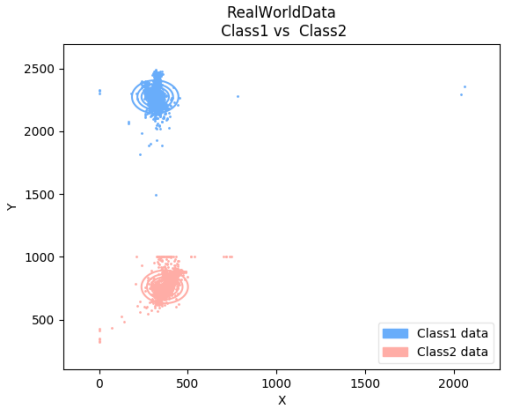
	class1	**class2**	**class3**
class1	0	0	125
class2	0	0	125
class3	90	46	114

Analysis

	class1	**class2**	**class3**
Precision	0.00	0.00	0.456
Recall	0.00	0.00	0.313
F-measure	-	-	0.371

3.1.3 Real World Data





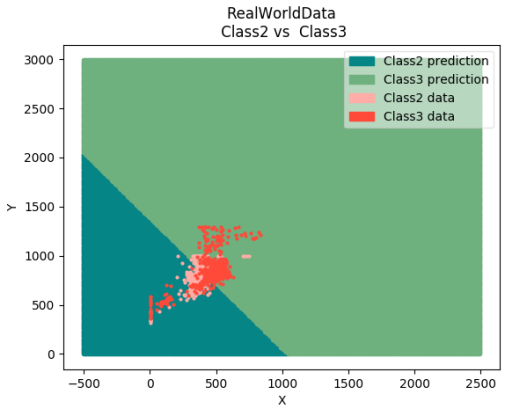
Accuracy = 99.75%

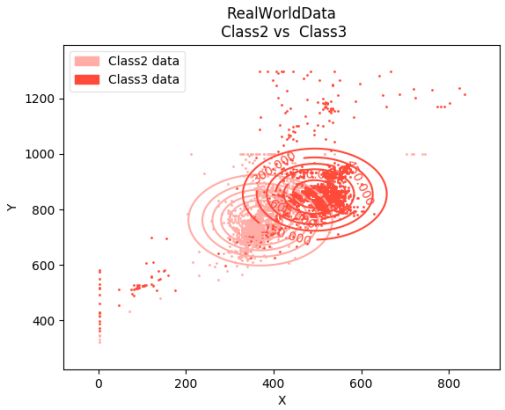
Confusion Matrix

	class1	**class2**
class1	594	3
class2	0	622

Analysis

	class1	**class2**
Precision	0.995	1.000
Recall	1.000	0.995
F-measure	0.997	0.997





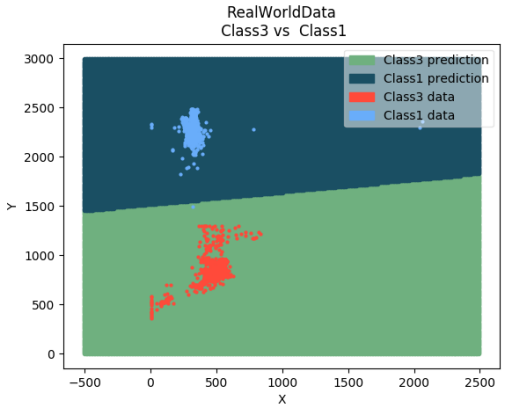
Accuracy = 81.55%

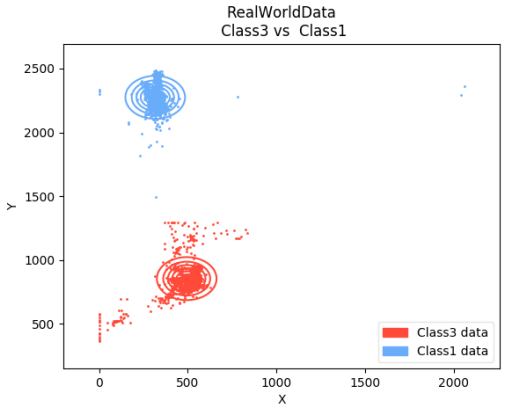
Confusion Matrix

	class1	**class2**
class1	572	50
class2	178	436

Analysis

	class1	**class2**
Precision	0.920	0.710
Recall	0.762	0.897
F-measure	0.834	0.793





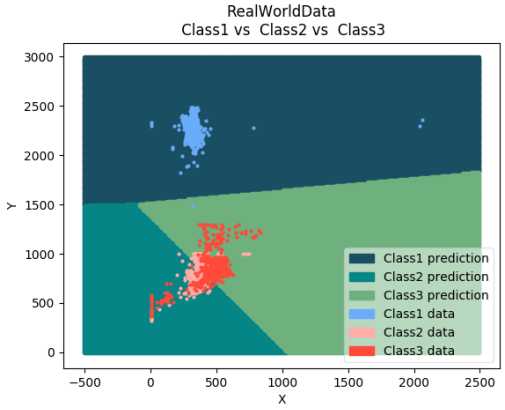
Accuracy = 99.67%

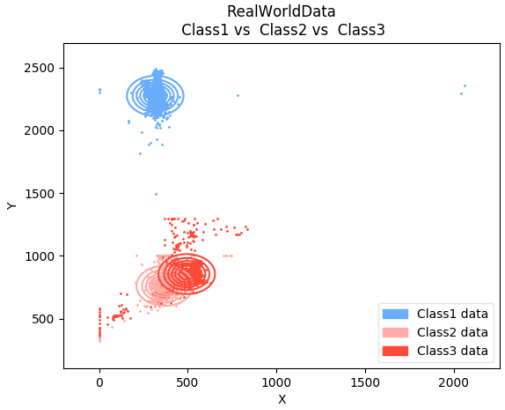
Confusion Matrix

	class1	**class2**
class1	614	0
class2	4	593

Analysis

	class1	**class2**
Precision	1.000	0.993
Recall	0.993	1.000
F-measure	0.997	0.997





Accuracy = 87.34%

Confusion Matrix

	class1	**class2**	**class3**
class1	593	3	1
class2	0	572	50
class3	0	178	436

Analysis

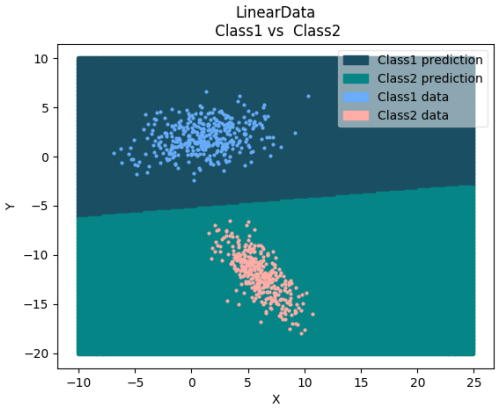
	class1	**class2**	**class3**
Precision	0.993	0.920	0.710
Recall	1.000	0.760	0.895
F-measure	0.997	0.832	0.792

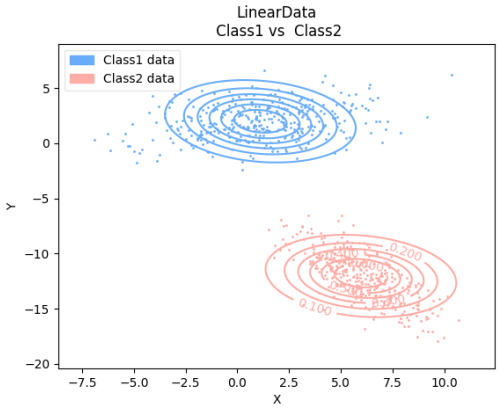
3.1.4 Inferences

1. It can be seen from the plots that the decision surface is linear in nature.
2. It works well for linearly separable data but gives poor results in case of non linearly separable data and real world data.
3. The nature of contour is circular as we have taken $\Sigma_i = \sigma^2 I$

3.2 Case 2 - $\Sigma_i = \Sigma$

3.2.1 Linear Data





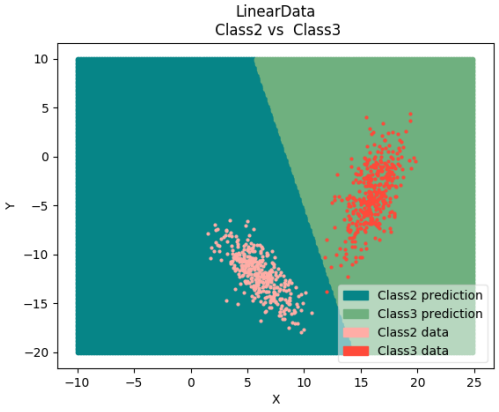
Accuracy = 100.00%

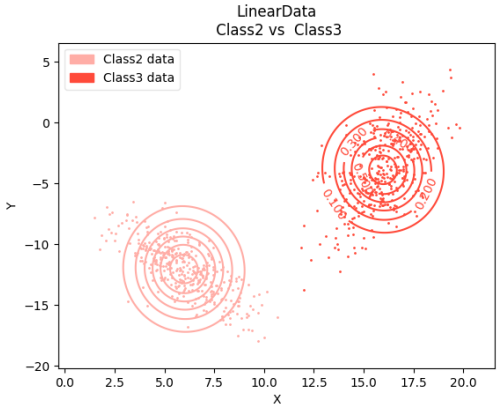
Confusion Matrix

	class1	**class2**
class1	125	0
class2	0	125

Analysis

	class1	**class2**
Precision	1.000	1.000
Recall	1.000	1.000
F-measure	1.000	1.000





It can be seen from the plot that the decision boundary is a straight line.

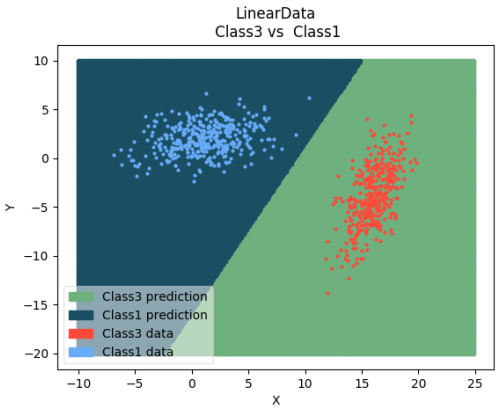
Accuracy = 100.00%

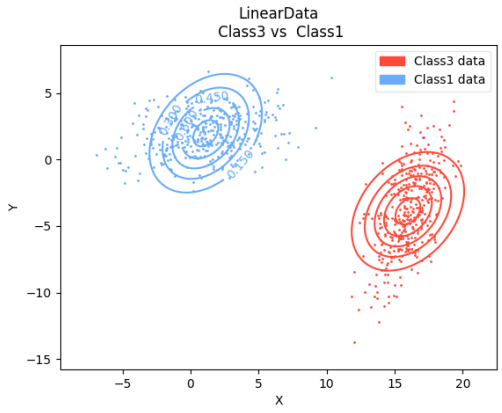
Confusion Matrix

Analysis

	class1	class2
class1	125	0
class2	0	125

	class1	class2
Precision	1.000	1.000
Recall	1.000	1.000
F-measure	1.000	1.000





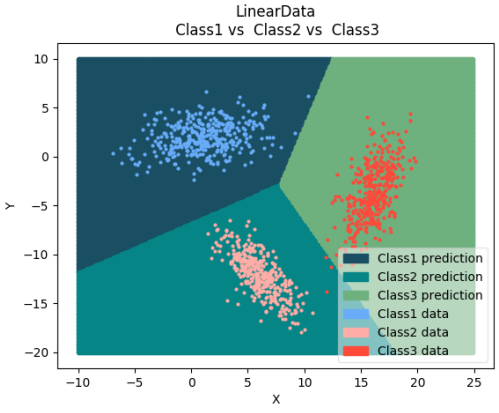
Accuracy = 99.6%

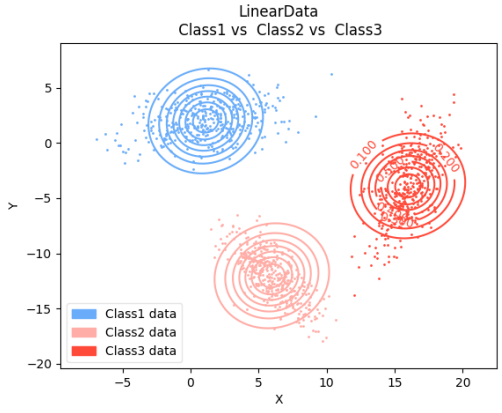
Confusion Matrix

	class1	class2
class1	125	0
class2	1	124

Analysis

	class1	class2
Precision	1.000	0.992
Recall	0.992	1.000
F-measure	0.996	0.996





Accuracy = 98.66%

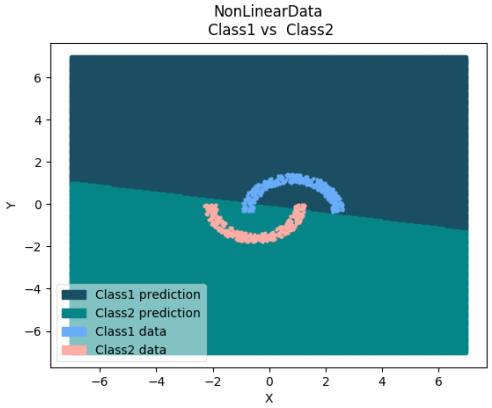
Confusion Matrix

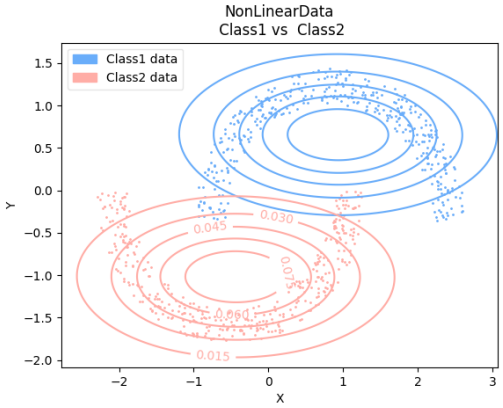
	class1	**class2**	**class3**
class1	123	0	2
class2	1	124	0
class3	0	2	123

Analysis

	class1	**class2**	**class3**
Precision	0.984	0.992	0.984
Recall	0.991	0.984	0.984
F-measure	0.988	0.988	0.984

3.2.2 NonLinear Data





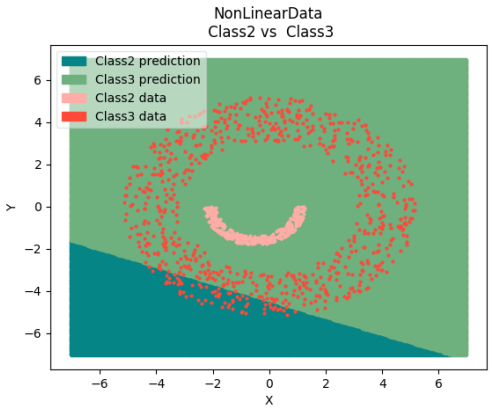
Accuracy = 95.6%

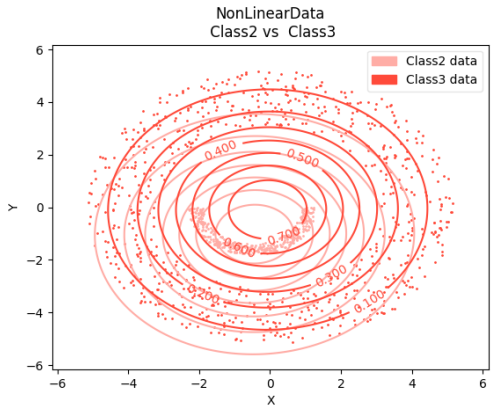
Confusion Matrix

	class1	**class2**
class1	121	4
class2	7	118

Analysis

	class1	**class2**
Precision	0.968	0.944
Recall	0.945	0.967
F-measure	0.955	0.955





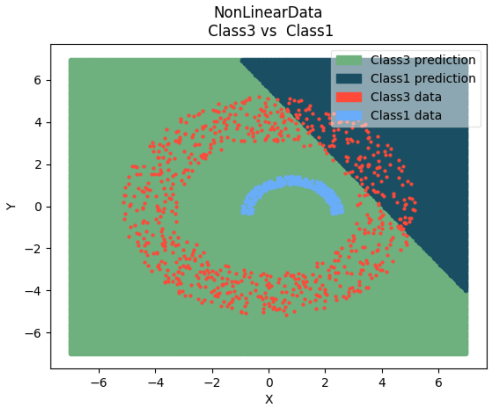
Accuracy = 61.33%

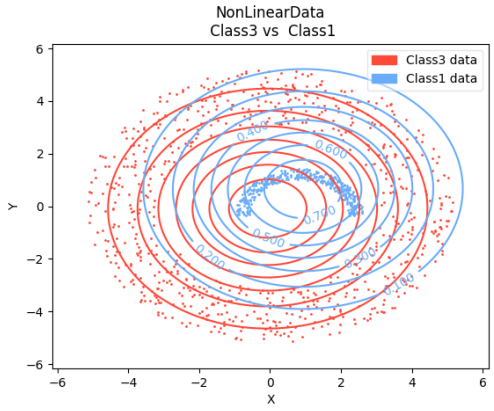
Confusion Matrix

	class1	**class2**
class1	0	124
class2	20	230

Analysis

	class1	**class2**
Precision	0.000	0.920
Recall	0.000	0.647
F-measure	-	0.760





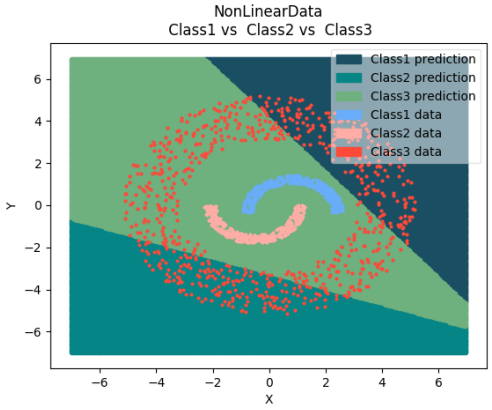
Accuracy = 49.33%

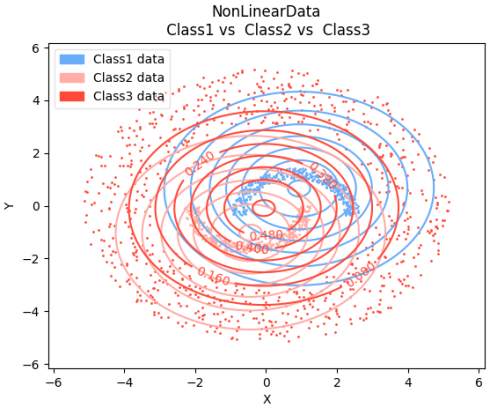
Confusion Matrix

	class1	**class2**
class1	185	65
class2	125	0

Analysis

	class1	**class2**
Precision	0.740	0.000
Recall	0.596	0.000
F-measure	0.661	0.996





Accuracy = 22.40%

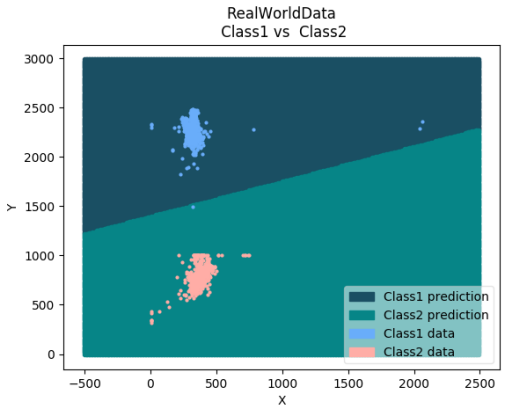
Confusion Matrix

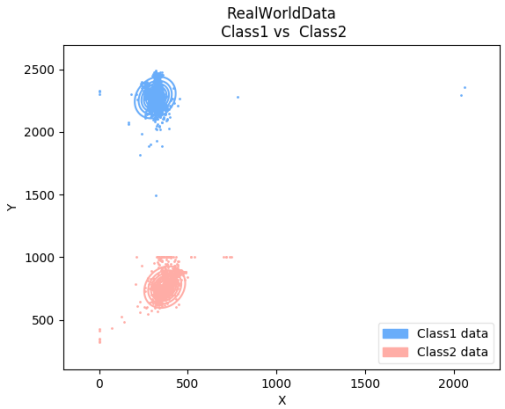
	class1	**class2**	**class3**
class1	0	0	125
class2	0	0	125
class3	89	49	112

Analysis

	class1	**class2**	**class3**
Precision	0.000	0.000	0.448
Recall	0.000	0.000	0.309
F-measure	-	-	0.366

3.2.3 Real World Data





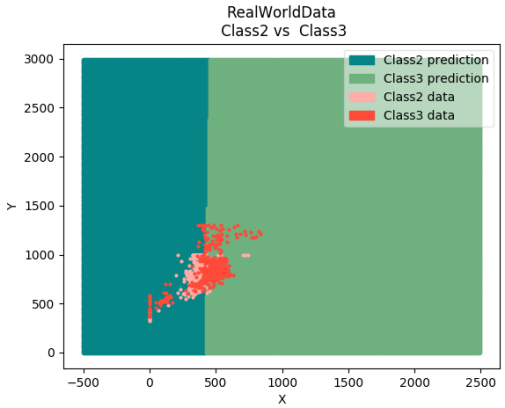
Accuracy = 99.67%

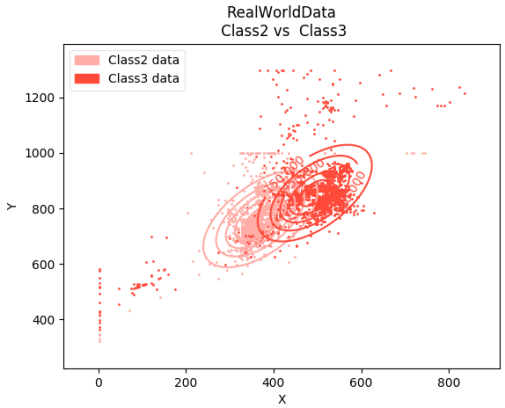
Confusion Matrix

Analysis

	class1	class2
class1	593	4
class2	0	622

	class1	class2
Precision	0.993	1.000
Recall	1.000	0.993
F-measure	0.996	0.996





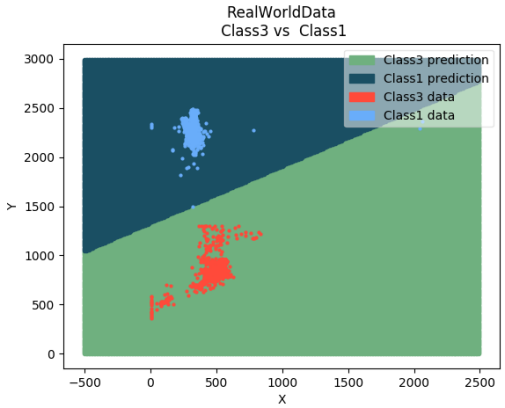
Accuracy = 81.60%

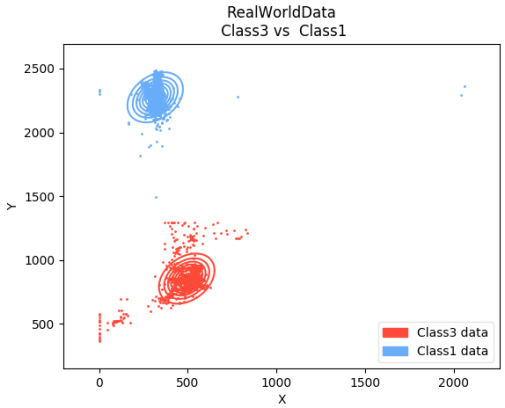
Confusion Matrix

	class1	class2
class1	601	21
class2	206	408

Analysis

	class1	class2
Precision	0.966	0.664
Recall	0.744	0.951
F-measure	0.841	0.782





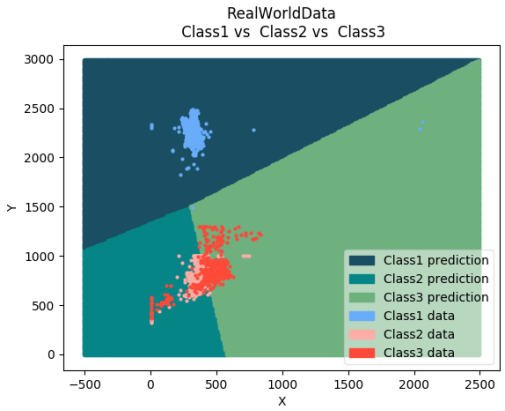
Accuracy = 98.80%

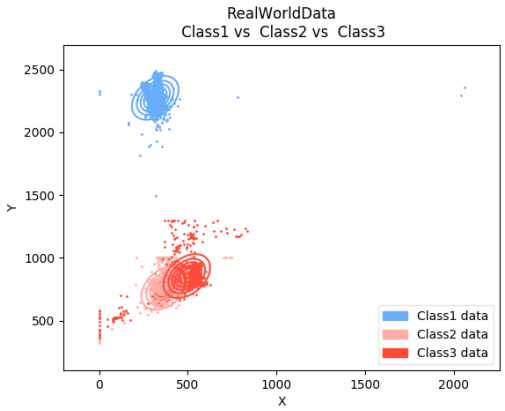
Confusion Matrix

	class1	**class2**
class1	614	0
class2	14	583

Analysis

	class1	**class2**
Precision	1.000	0.976
Recall	0.977	1.000
F-measure	0.988	0.988





Accuracy = 87.10%

Confusion Matrix

	class1	**class2**	**class3**
class1	581	3	13
class2	0	592	30
class3	0	191	423

Analysis

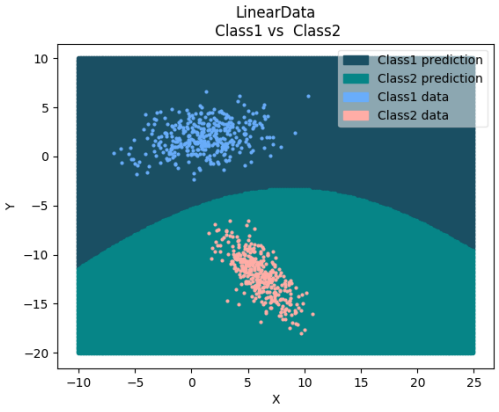
	class1	**class2**	**class3**
Precision	0.973	0.951	0.688
Recall	1.000	0.753	0.907
F-measure	0.986	0.841	0.783

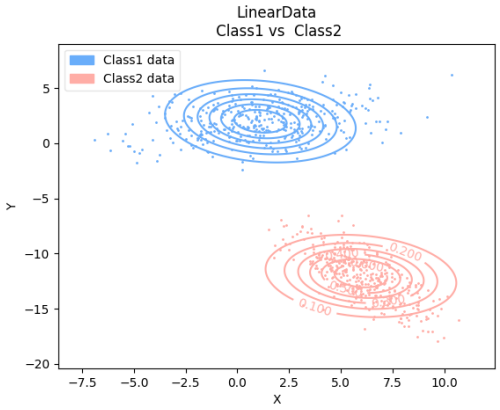
3.2.4 Inferences

1. It can be seen from the plots that the decision surface is linear in nature.
2. It works well for linearly separable data but gives poor results in case of non linearly separable data and real world data.
3. The nature of contour is elliptical as we have taken $\Sigma_i = \Sigma$.
4. To find same Σ_i, we took mean of the $\Sigma_i's$ of the three classes.

3.3 Case 3 - Σ_i is a diagonal matrix

3.3.1 Linear Data





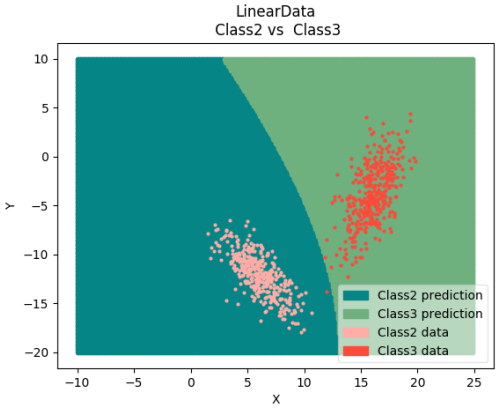
Accuracy = 100.00%

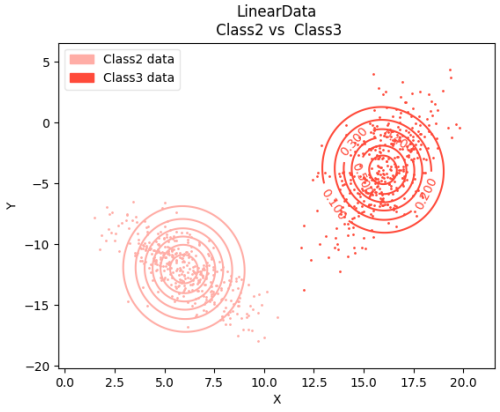
Confusion Matrix

	class1	**class2**
class1	125	0
class2	0	125

Analysis

	class1	**class2**
Precision	1.000	1.000
Recall	1.000	1.000
F-measure	1.000	1.000





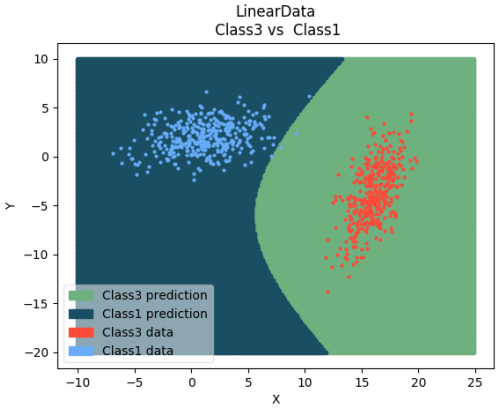
Accuracy = 100.00%

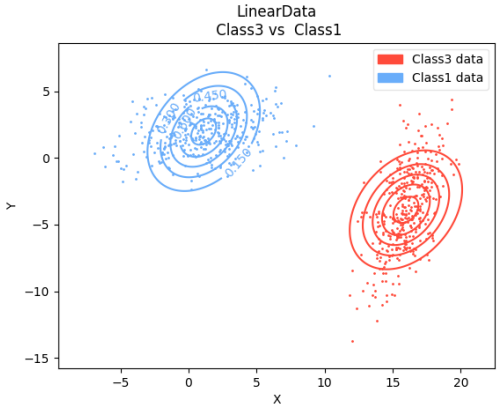
Confusion Matrix

	class1	**class2**
class1	125	0
class2	0	125

Analysis

	class1	**class2**
Precision	1.000	1.000
Recall	1.000	1.000
F-measure	1.000	1.000





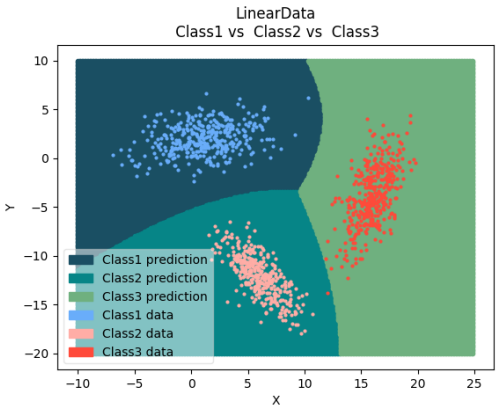
Accuracy = 99.2%

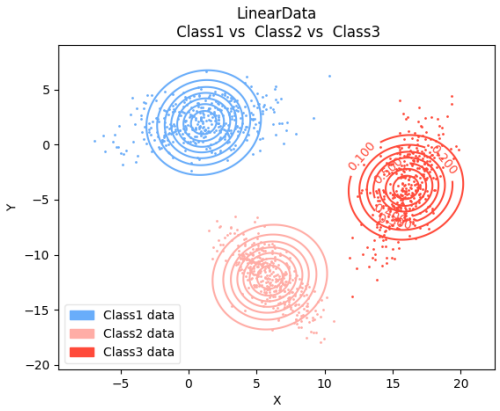
Confusion Matrix

	class1	**class2**
class1	125	0
class2	2	123

Analysis

	class1	**class2**
Precision	1.000	0.984
Recall	0.984	1.000
F-measure	0.992	0.992





Accuracy = 100.00%

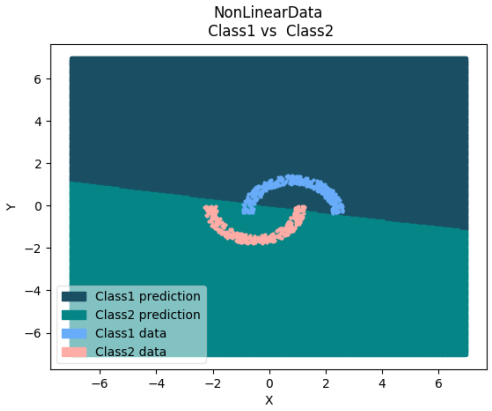
Confusion Matrix

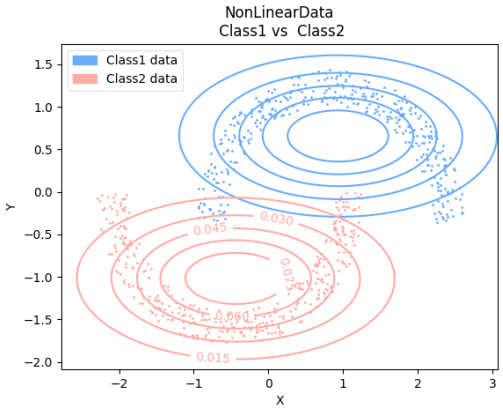
	class1	**class2**	**class3**
class1	125	0	0
class2	0	125	0
class3	0	0	125

Analysis

	class1	**class2**	**class3**
Precision	1.000	1.000	1.000
Recall	1.000	1.000	1.000
F-measure	1.000	1.000	1.000

3.3.2 NonLinear Data





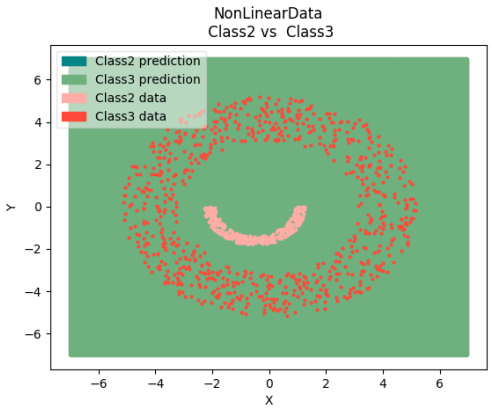
Accuracy = 96.4%

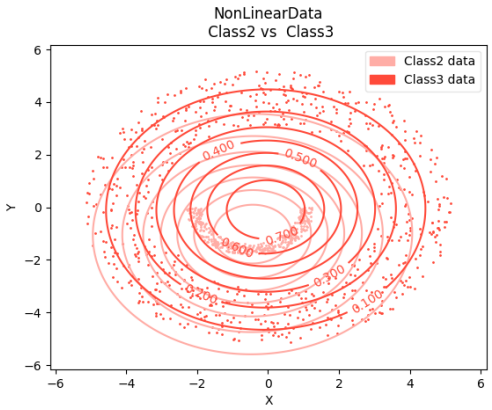
Confusion Matrix

	class1	**class2**
class1	120	5
class2	4	121

Analysis

	class1	**class2**
Precision	0.960	0.968
Recall	0.967	0.960
F-measure	0.963	0.964





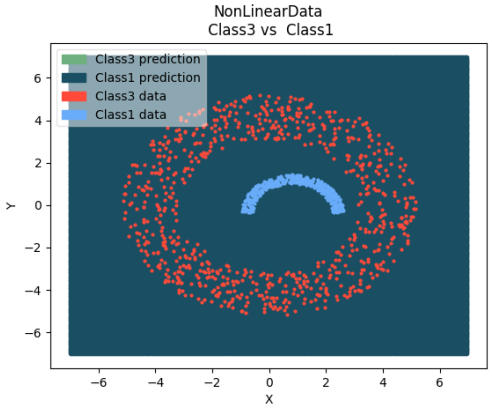
Accuracy = 66.66%

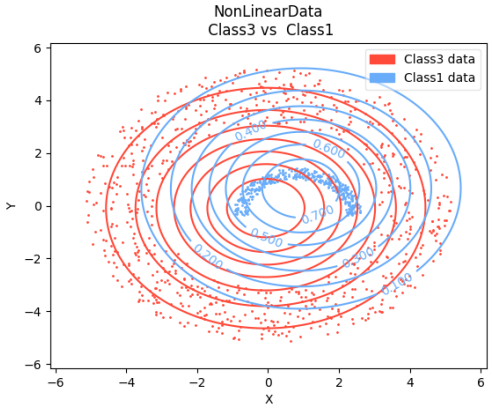
Confusion Matrix

	class1	**class2**
class1	0	125
class2	0	250

Analysis

	class1	**class2**
Precision	0.000	1.000
Recall	-	0.666
F-measure	-	0.800





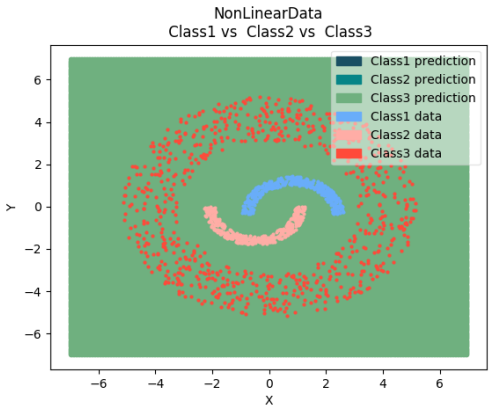
Accuracy = 33.33%

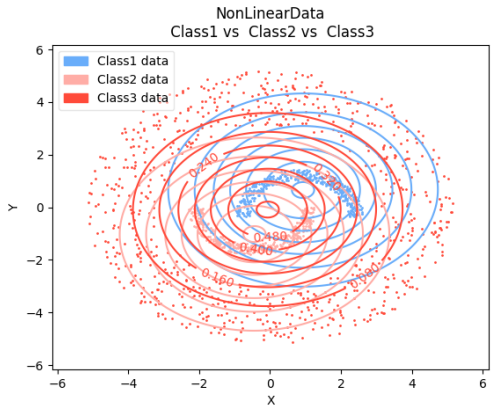
Confusion Matrix

	class1	**class2**
class1	0	250
class2	0	125

Analysis

	class1	**class2**
Precision	0.000	1.000
Recall	-	0.333
F-measure	-	0.500





Accuracy = 50.00%

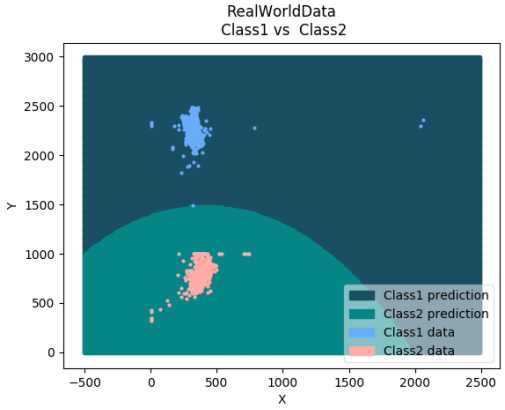
Confusion Matrix

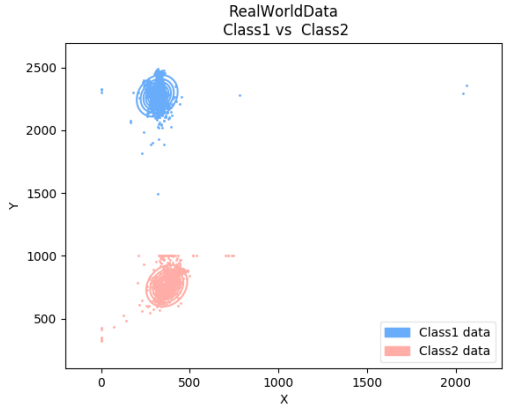
	class1	**class2**	**class3**
class1	0	0	125
class2	0	0	125
class3	0	0	250

Analysis

	class1	**class2**	**class3**
Precision	0.000	0.000	1
Recall	-	-	0.500
F-measure	-	-	0.666

3.3.3 RealWorld Data





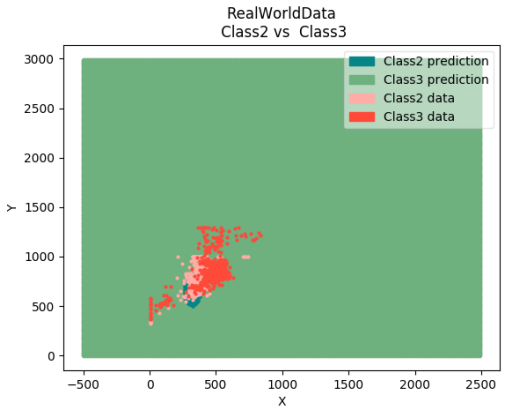
Accuracy = 99.75%

Confusion Matrix

	class1	**class2**
class1	594	3
class2	0	622

Analysis

	class1	**class2**
Precision	0.994	1.000
Recall	1.000	0.994
F-measure	0.997	0.997



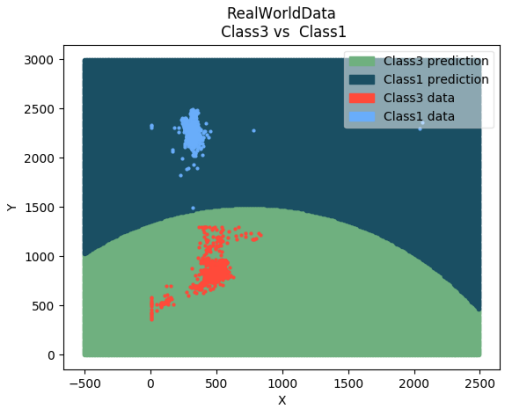
Accuracy = 87.70%

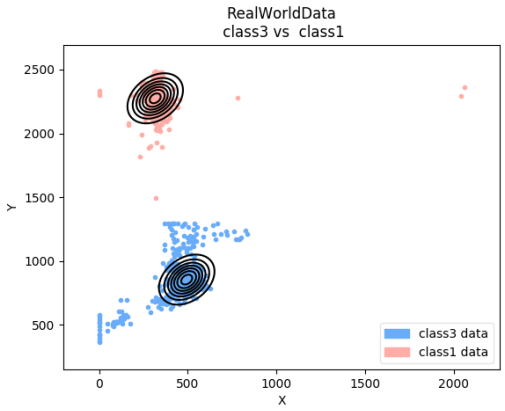
Confusion Matrix

	class1	**class2**
class1	500	122
class2	30	584

Analysis

	class1	**class2**
Precision	0.804	0.951
Recall	0.943	0.827
F-measure	0.868	0.884





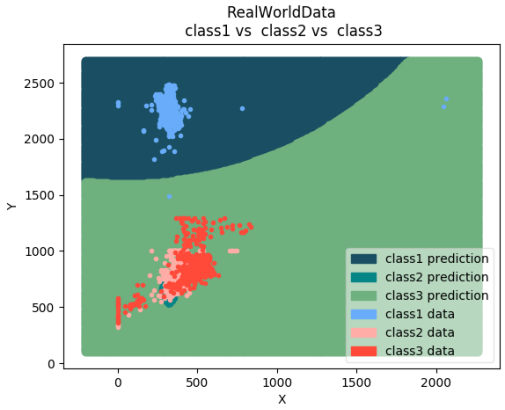
Accuracy = 99.75%

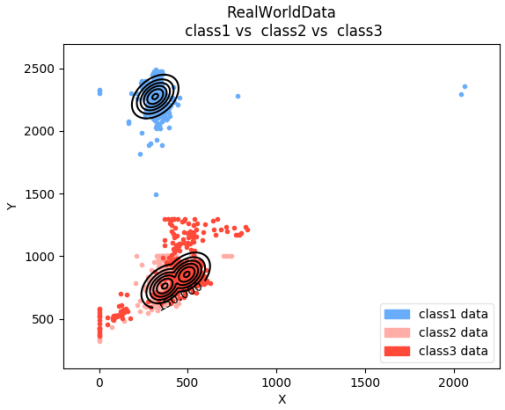
Confusion Matrix

	class1	**class2**
class1	614	0
class2	3	594

Analysis

	class1	**class2**
Precision	1.000	0.995
Recall	0.995	1.000
F-measure	0.997	0.997





Accuracy = 90.80%

Confusion Matrix

	class1	**class2**	**class3**
class1	580	2	15
class2	0	500	122
class3	0	30	584

Analysis

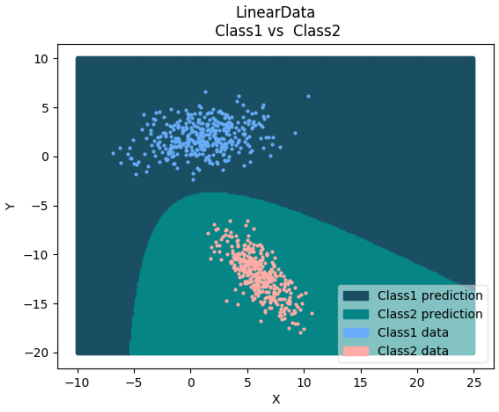
	class1	**class2**	**class3**
Precision	0.971	0.804	0.951
Recall	1.000	0.939	0.809
F-measure	0.985	0.866	0.875

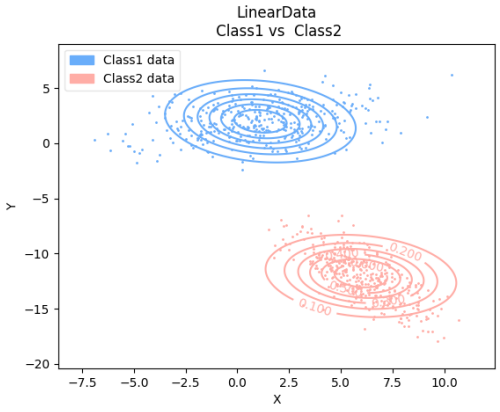
3.3.4 Inferences

1. It can be seen from the plots that the decision surface is non - linear in nature.
2. It works well for linearly separable data as well as non-linearly separable data.
3. The nature of contour is elliptical and is different for each class but oriented horizontally.

3.4 Case 4 - Σ_i is unique

3.4.1 Linear Data





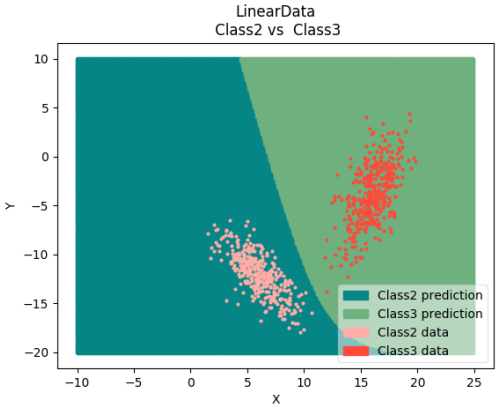
Accuracy = 100.00%

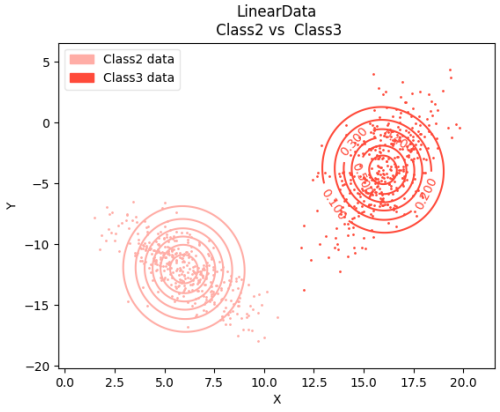
Confusion Matrix

	class1	**class2**
class1	125	0
class2	0	125

Analysis

	class1	**class2**
Precision	1.000	1.000
Recall	1.000	1.000
F-measure	1.000	1.000





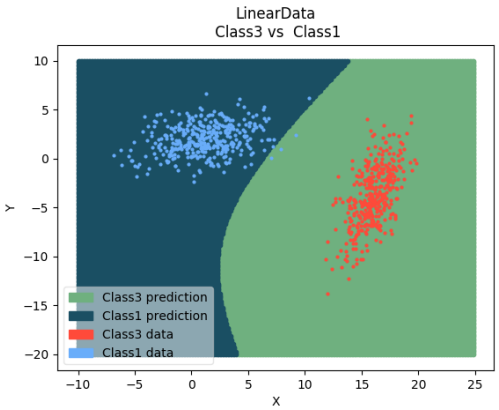
Accuracy = 100.00%

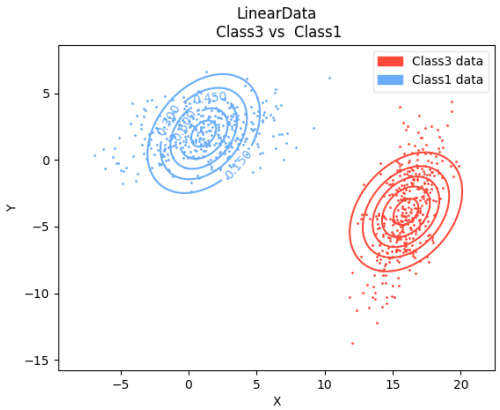
Confusion Matrix

	class1	**class2**
class1	125	0
class2	0	125

Analysis

	class1	**class2**
Precision	1.000	1.000
Recall	1.000	1.000
F-measure	1.000	1.000





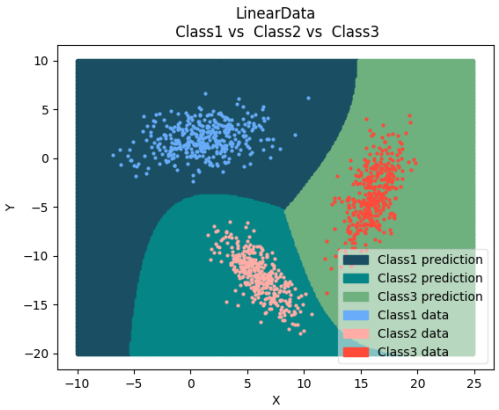
Accuracy = 99.2%

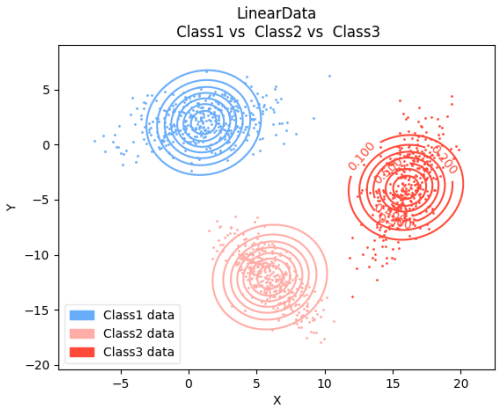
Confusion Matrix

	class1	**class2**
class1	125	0
class2	2	123

Analysis

	class1	**class2**
Precision	1.000	0.984
Recall	0.984	1.000
F-measure	0.992	0.992





Accuracy = 100.00%

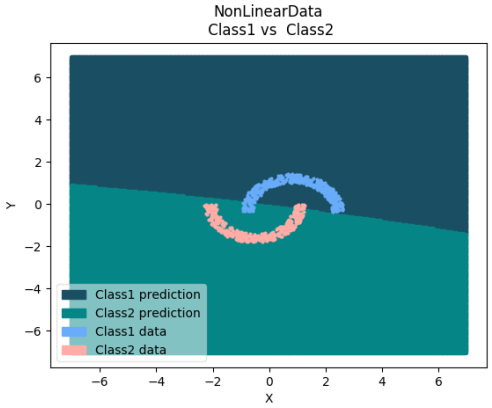
Confusion Matrix

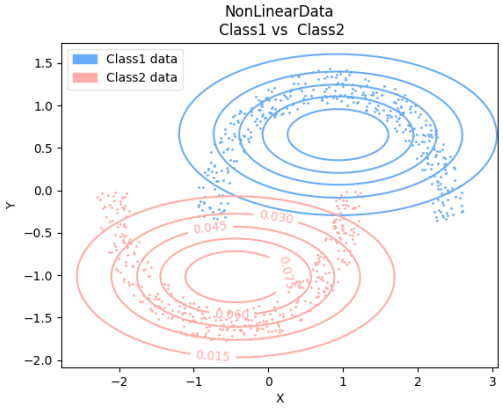
	class1	**class2**	**class3**
class1	125	0	0
class2	0	125	0
class3	0	0	125

Analysis

	class1	**class2**	**class3**
Precision	1.000	1.000	1.000
Recall	1.000	1.000	1.000
F-measure	1.000	1.000	1.000

3.4.2 NonLinear Data





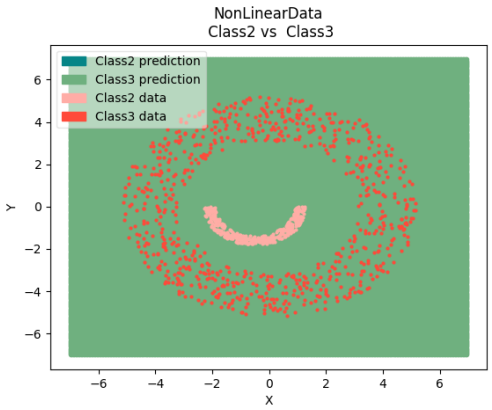
Accuracy = 96.4%

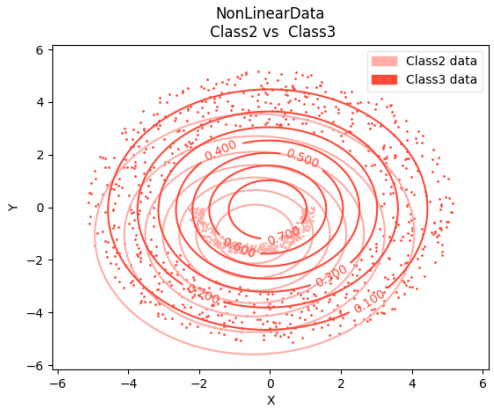
Confusion Matrix

	class1	**class2**
class1	120	5
class2	4	121

Analysis

	class1	**class2**
Precision	0.960	0.968
Recall	0.967	0.960
F-measure	0.963	0.964





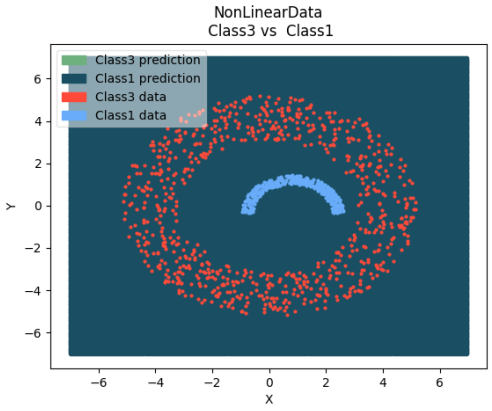
Accuracy = 66.66%

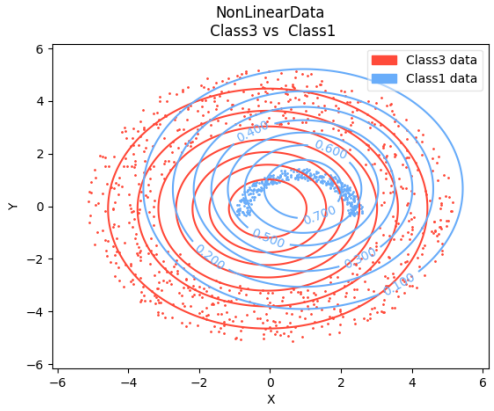
Confusion Matrix

	class1	**class2**
class1	0	125
class2	0	250

Analysis

	class1	**class2**
Precision	0.000	1.000
Recall	-	0.666
F-measure	-	0.800





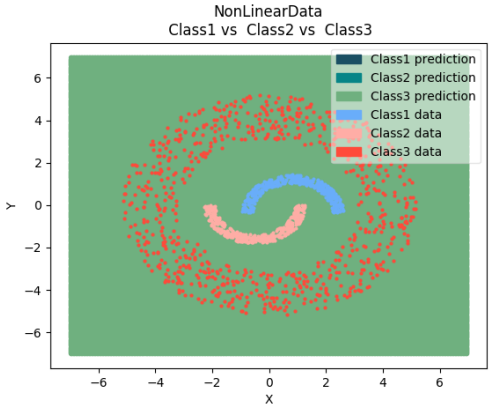
Accuracy = 33.33%

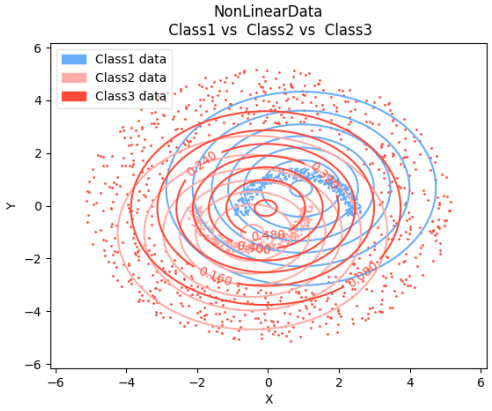
Confusion Matrix

	class1	**class2**
class1	0	250
class2	0	125

Analysis

	class1	**class2**
Precision	0.000	1.000
Recall	-	0.333
F-measure	-	0.500





Accuracy = 50.00%

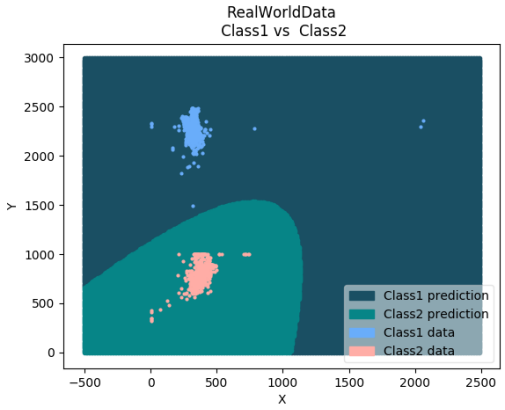
Confusion Matrix

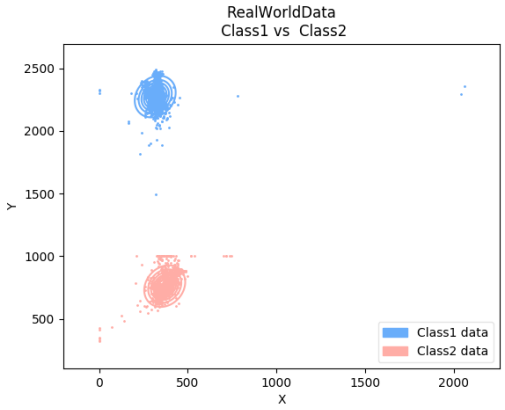
	class1	**class2**	**class3**
class1	0	0	125
class2	0	0	125
class3	0	0	250

Analysis

	class1	**class2**	**class3**
Precision	0.000	0.000	1
Recall	-	-	0.500
F-measure	-	-	0.666

3.4.3 RealWorld Data





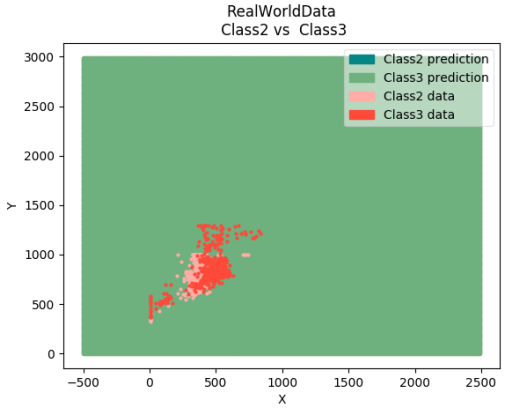
Accuracy = 99.75%

Confusion Matrix

	class1	**class2**
class1	594	3
class2	0	622

Analysis

	class1	**class2**
Precision	0.994	1.000
Recall	1.000	0.994
F-measure	0.997	0.997



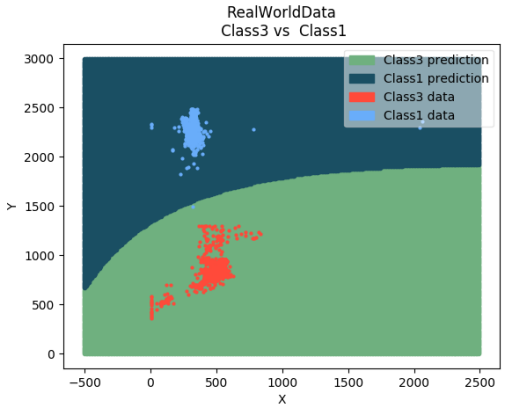
Accuracy = 77.10%

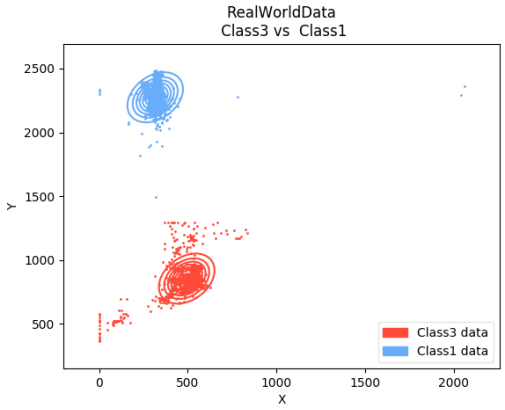
Confusion Matrix

Analysis

	class1	class2
class1	350	272
class2	11	603

	class1	class2
Precision	0.562	0.982
Recall	0.969	0.689
F-measure	0.712	0.809





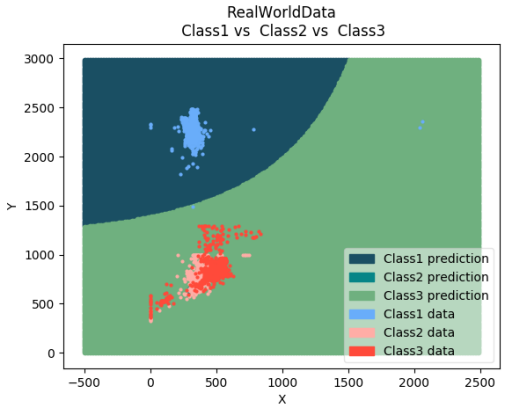
Accuracy = 99.66%

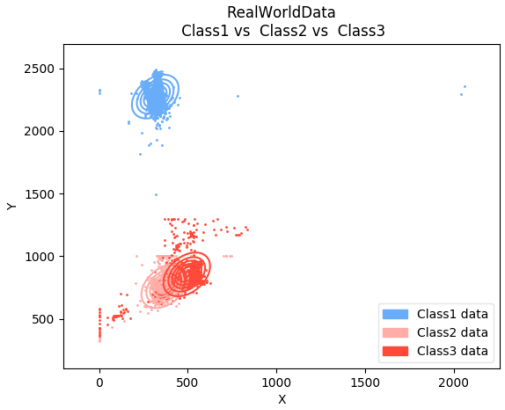
Confusion Matrix

	class1	class2
class1	614	0
class2	4	593

Analysis

	class1	class2
Precision	1.000	0.993
Recall	0.993	1.000
F-measure	0.996	0.996





Accuracy = 83.68%

Confusion Matrix

	class1	**class2**	**class3**
class1	581	2	14
class2	0	350	272
class3	0	11	603

Analysis

	class1	**class2**	**class3**
Precision	0.973	0.562	0.982
Recall	1.000	0.964	0.678
F-measure	0.986	0.710	0.802

3.4.4 Inferences

1. It can be seen from the plots that the decision surface is non - linear in nature.
2. It works well for linearly separable data and good (but not as good as linearly separable data) for non linearly separable data.
3. The nature of contour is elliptical and different for every class and also oriented independently in any directions.

4 Conclusion

1. Bayes Classifier works well for Linearly separable data in all cases with high accuracy but fails for Non linearly separable data with poor accuracy.
2. In case of Real World Data, the data is overlapping and thus results in lesser accuracy.
3. In first two cases, the decision boundary comes out to be straight line whereas in last two cases, the decision boundary is quadratic due to different covariance matrix chosen.