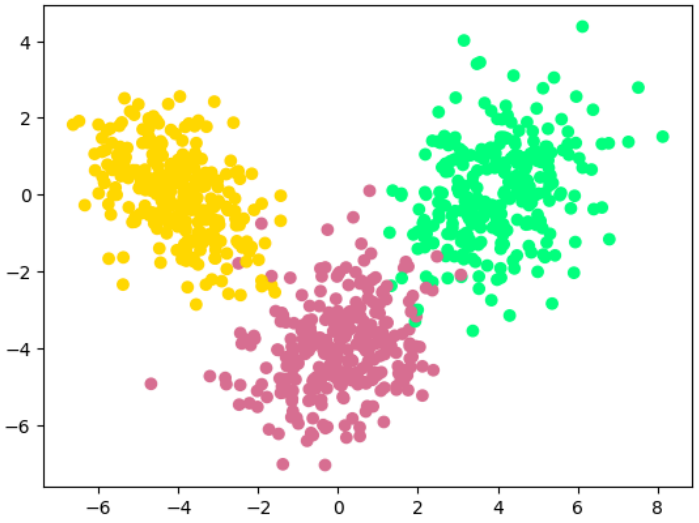
Report Assignment 4 Pattern Recognition & Machine learning

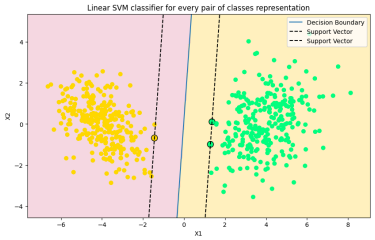
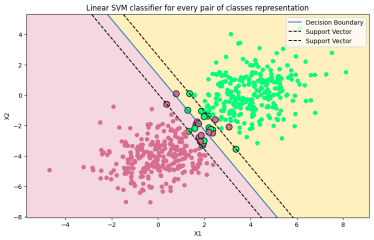
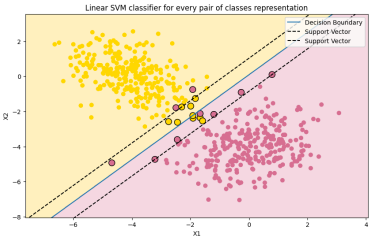
Riccardo Guderzo GE24Z227
Joanna Kolaczek GE24Z229
Miguel Mauer GE24Z022

Task 1: Classifier for Dataset 1 using linear kernel based SVMs

Since we have 3 classes in our dataset, we need to perform linear kernel based SVM for each pair of the classes.



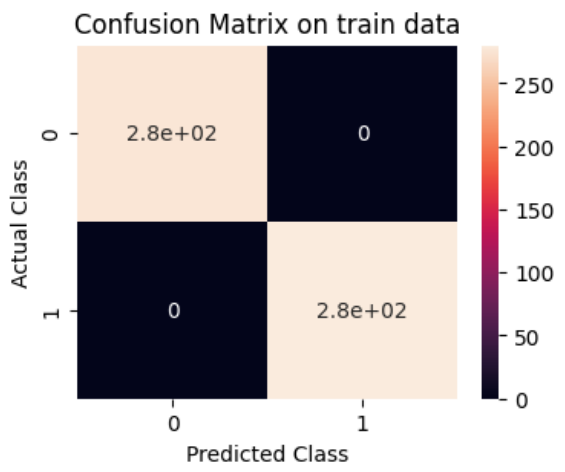
Linear kernel based SVM for:

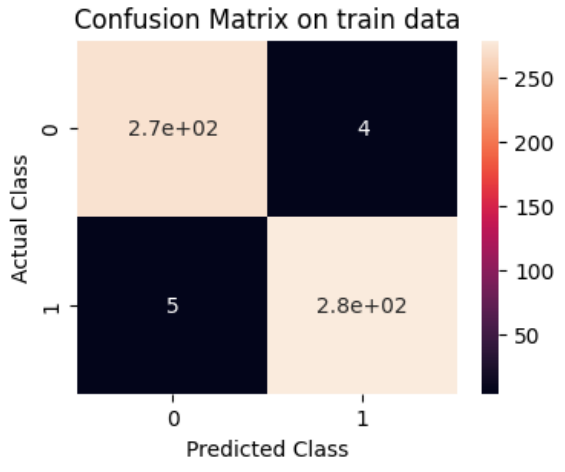
Classes 0 & 1	Classes 0 & 2	Classes 1 & 2
		

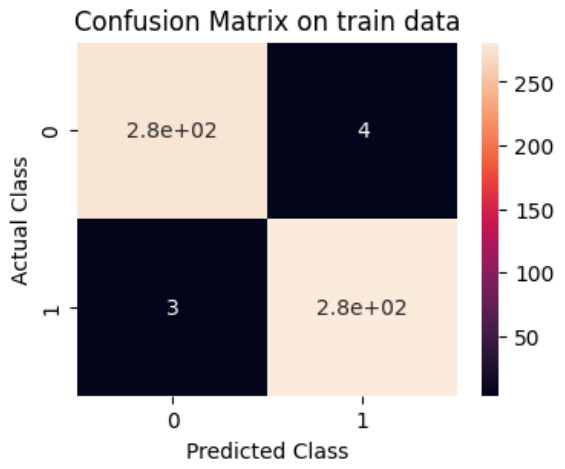
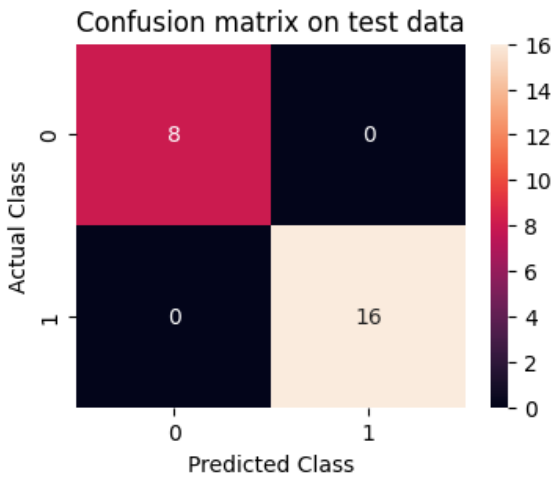
Training, testing and validation accuracies:

Classes 0 & 1			Classes 0 & 2			Classes 1 & 2		
Training	Test	Validation	Training	Test	Validation	Training	Test	Validation
100.00%	100.00%	100.00%	98.39%	95.83%	97.96%	98.76%	100.00%	100.00%

Confusion matrices for training and test data:

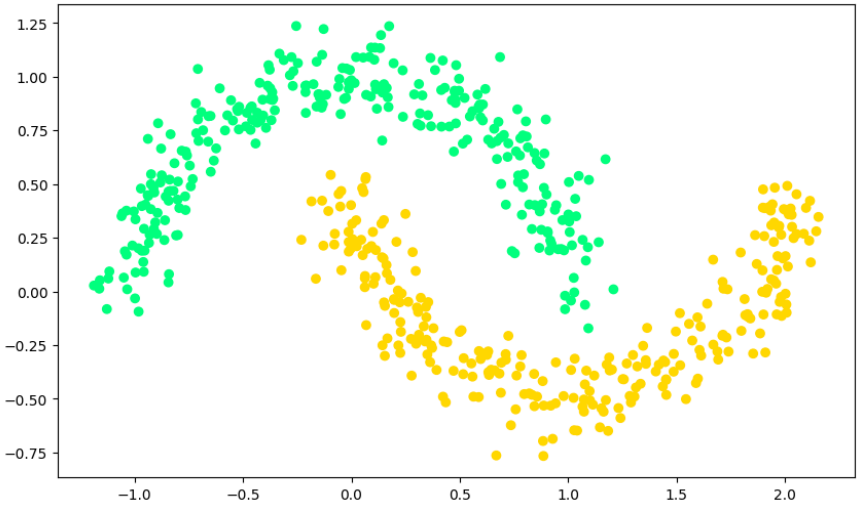
Classes 0 & 1	
Training	Test
	

Classes 0 & 2	
Training	Test
	

Classes 1 & 2	
Training	Test
	

Task 2: Classifier for Dataset 2

Nonlinearly separable data for 2 classes:



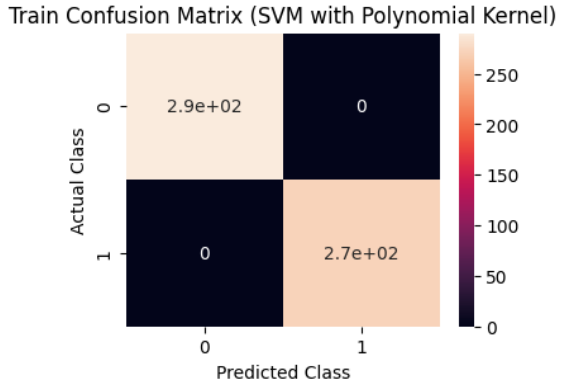
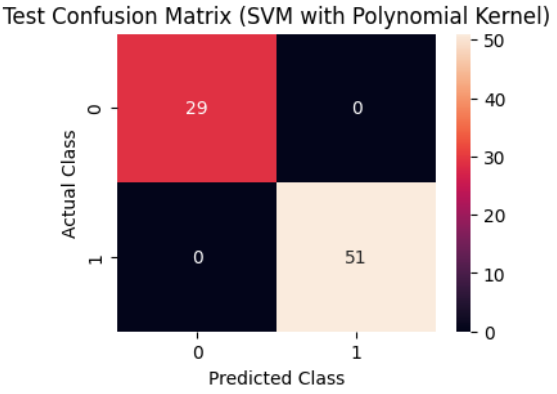
(a) polynomial kernel based SVM

We've chosen the degrees 1,2,3 & 5 for polynomial kernel based SVM and we checked the training, test and validation accuracies for C=1, C=10, C=100. The highest accuracies were obtained for the degree 5. For C=10 and C=100 we got the best accuracy overall.

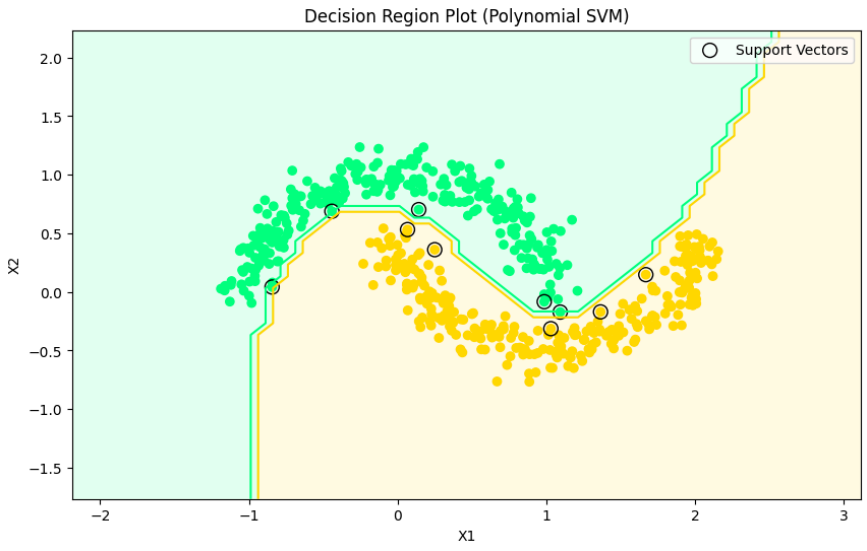
Training, testing and validation accuracies:

degree	C=1			C=10			C=100		
	Training	Test	Validation	Training	Test	Validation	Training	Test	Validation
1	86.63%	87.50%	84.91%	88.24%	87.50%	86.79%	88.41%	87.50%	86.79%
2	88.24%	87.50%	86.16%	88.41%	88.75%	86.79%	88.59%	88.75%	86.79%
3	89.30%	91.25%	86.79%	95.37%	97.50%	94.97%	99.64%	98.75%	98.11%
5	99.82%	100.00%	99.37%	100.00%	100.00%	99.37%	100.00%	100.00%	99.37%

Confusion matrices for training and test data for the best performing model:

Degree = 5, C = 10	
Training	Test
	

Decision region plot for polynomial based SVM with degree=5 and C=10:

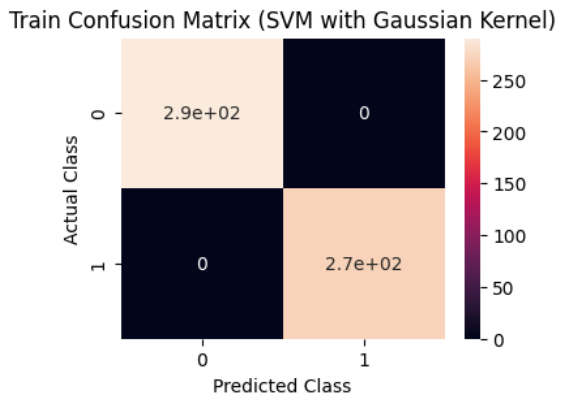
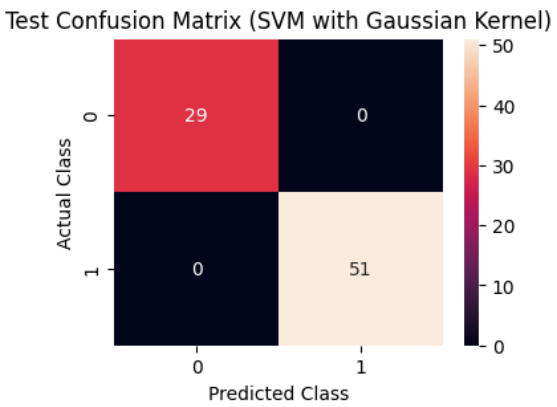


(b) Gaussian kernel based SVM

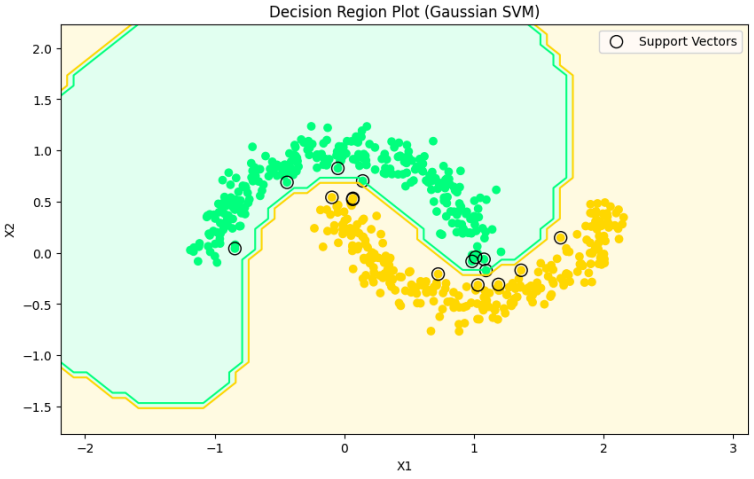
We've chosen the width 1,0.01,0.001 & 0.0001 for gaussian kernel based SVM and we checked the training, test and validation accuracies for C=1, C=10, C=100. The highest accuracies were obtained for the width 1. For C=10 and C=100 we got the best accuracy overall.

width	C=1			C=10			C=100		
	Training	Test	Validation	Training	Test	Validation	Training	Test	Validation
1	99.82%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
0.01	84.13%	87.5%	87.5%	87.16%	87.5%	87.5%	88.23%	87.5%	87.5%
0.001	78.78%	78.75%	78.75%	84.31%	87.5%	87.5%	87.34%	87.5%	87.5%
0.0001	51.69%	36.25%	36.25%	78.78%	78.75%	78.75%	84.31%	87.5%	87.5%

Confusion matrices for training and test data for the best performing model:

Width = 1, C = 10	
Training	Test
	

Decision region plot for Gaussian based SVM with width=5 and C=10:



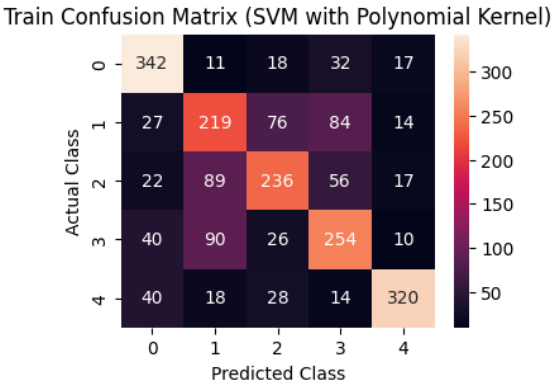
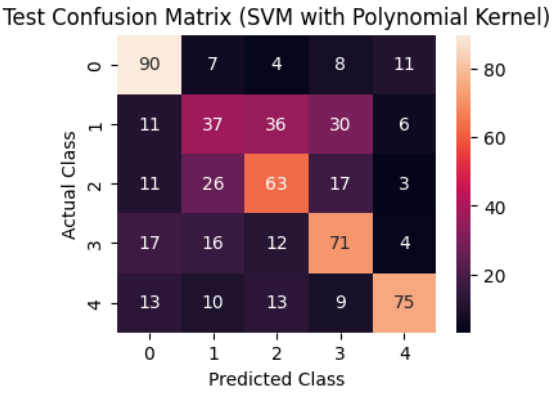
Task 3: Classifier for Dataset 3

(a) polynomial kernel based SVMs

Training, testing and validation accuracies:

degree	C=1			C=10			C=100		
	Training	Test	Validation	Training	Test	Validation	Training	Test	Validation
1	55.24%	51.33%	56.67%	59.57%	53.00%	55.67%	59.86%	53.67%	56.00%
2	59.76%	54.83%	56.67%	60.71%	53.50%	54.67%	62.19%	53.83%	53.67%
3	61.10%	54.17%	56.00%	65.29%	56.00%	58.00%	80.90%	55.83%	56.67**%**
5	90.05%	54.83%	56.33%	99.76%	51.67%	56.33%	100.00%	51.17%	55.33%

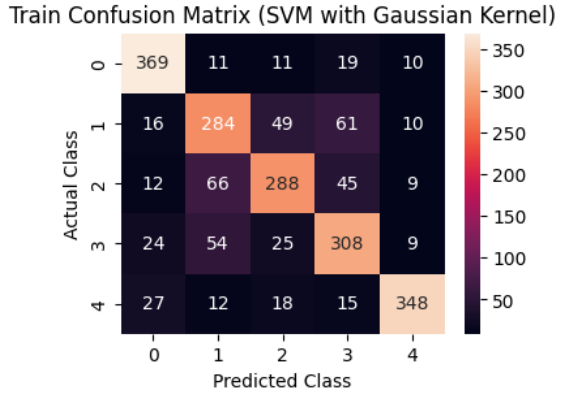
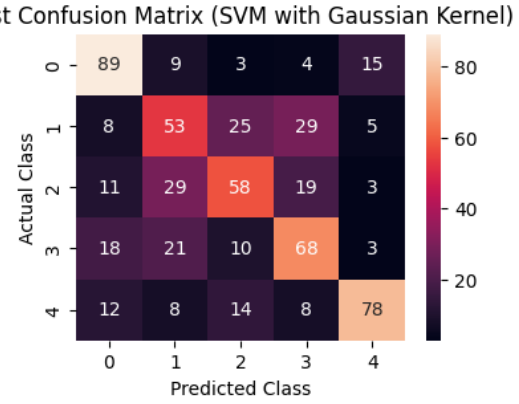
Confusion matrices for training and test data for the best performing model:

Degree = 3, C = 10	
Training	Test
	

(b) Gaussian kernel based SVMs

width	C=1			C=10			C=100		
	Training	Test	Validation	Training	Test	Validation	Training	Test	Validation
1	63.81%	57.17%	61.0%	76.01%	57.67%	60.0%	96.81%	55.33%	56.33%
0.01	51.23%	48.83%	53.67%	57.52%	52.33%	58.0%	59.71%	54.67%	56.67%
0.001	49.90%	46.67%	51.33%	51.05%	48.83%	53.33%	57.75%	52.33%	58.0%
0.0001	49.90%	46.67%	51.33%	49.90%	46.67%	51.33	51.05%	48.83%	53.33%

Training, testing and validation accuracies:Confusion matrices for training and test data for the best performing model:

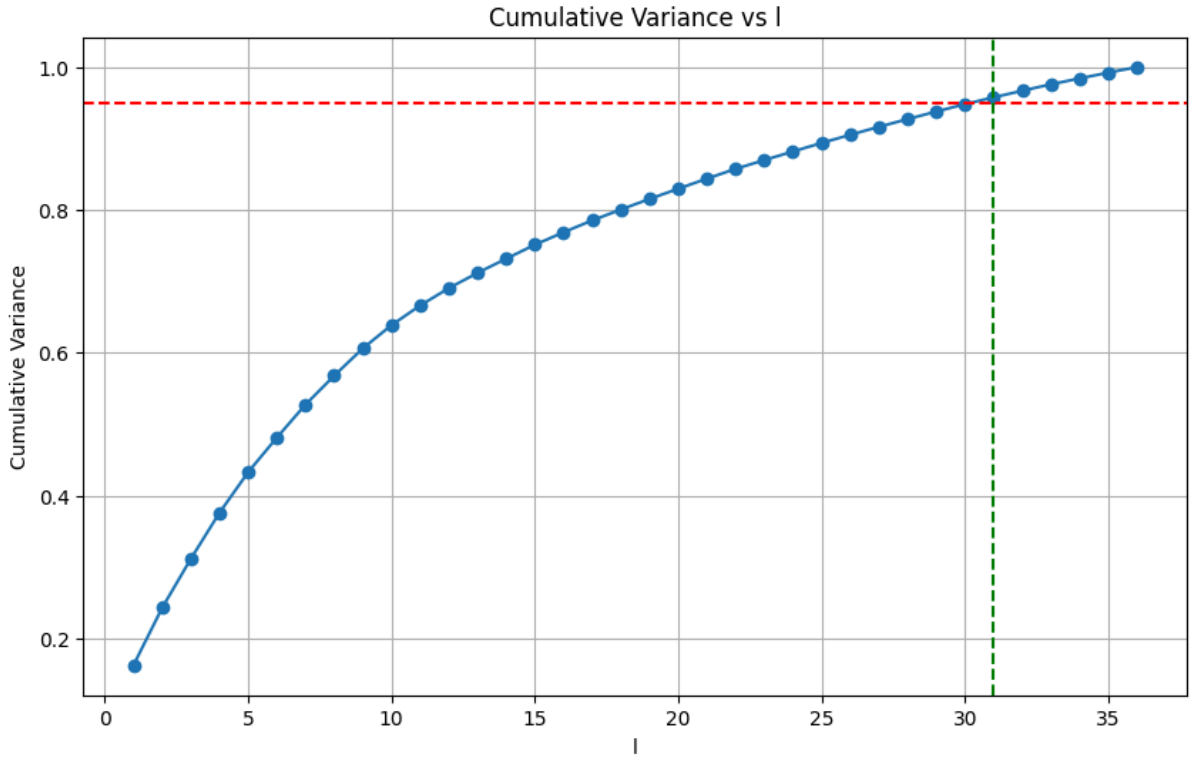
<table>
<tr><td colspan="2">Width = 1, C = 10</td></tr>
<tr><td>Training</td><td>Test</td></tr>
<tr><td> 
</td><td> 
</td></tr>
</table>

Percentage of bounded and unbounded support vectors out of the training examples, for the best performing models (Polynomial: C=10 degree = 5, Gaussian: C=10, width = 1):

Kernel	Percentage of Bounded SV	Percentage of Unbounded SV
Polynomial	7.5%	0%
Gaussian	3%	3.7%

Task 4: Perform PCA for Dataset 3 and identify a suitable value of the reduced dimension l

Principal Component Analysis was applied to Dataset 3 to find a suitable reduced dimension l for better data representation. The smallest l needed to retain at least 90% of the original variance is 31. This method reduces dimensionality while preserving key information, improving model efficiency.



Task 5: Classifier for Dataset 3 with reduced dimension representation obtained in Exercise 4, as the input to the classifier. The classification models are GMM, MLFFNN, SVM:

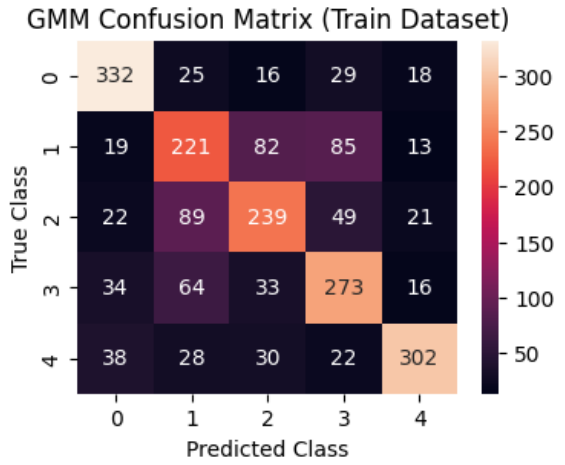
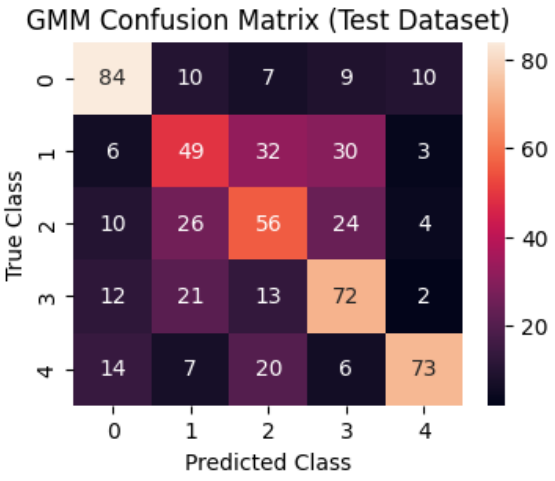
From exercise 3 the best performing SVM was the one with Gaussian kernel, width=2 and C=10. It was chosen for comparison with GMM and MLFFNN in this task. For Gaussian Mixture Model the number of components is 3 and a diagonal matrix was chosen. The Neural Network was designed as follows:

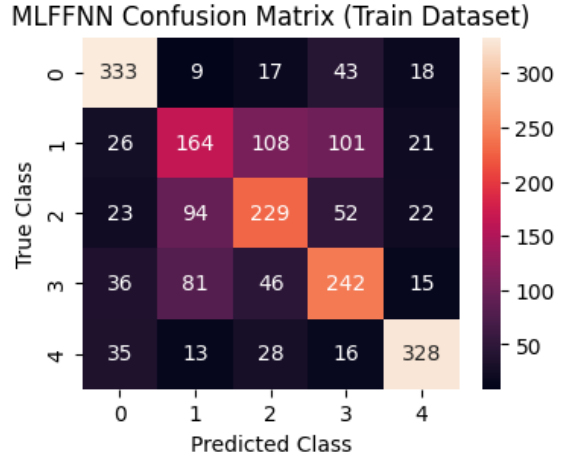
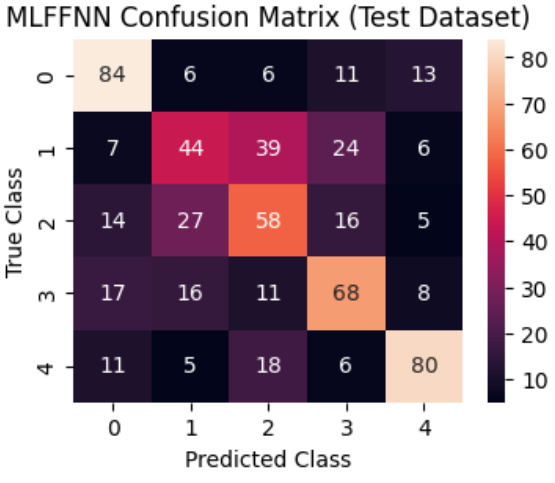
Layer (type)	**Output Shape**	**Param #**
Linear-1	[-1, 1, 25]	925
Tanh-2	[-1, 1, 25]	0
Linear-3	[-1, 1, 15]	390
Tanh-4	[-1, 1, 15]	0
Linear-5	[-1, 1, 5]	80
Total parameters: 1395		

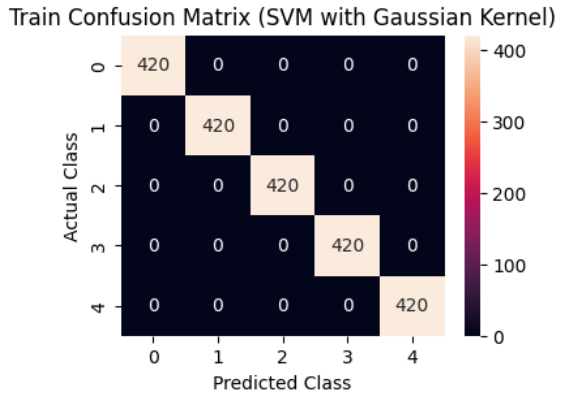
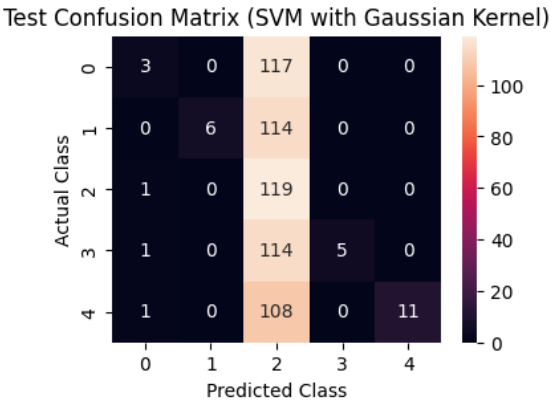
Training, testing and validation accuracies for each model:

GMM			MLFFNN			SVM		
Training	Test	Validation	Training	Test	Validation	Training	Test	Validation
65.10%	55.67%	58.33%	61.71%	55.67%	57.33%	100.00%	24.00%	24.00%

Confusion matrices:

GMM	
Training	Test
	

MLFFNN	
Training	Test
	

Gaussian kernel SVM with gamma = 1, C=10	
Training	Test
	

From above confusion matrices and accuracies we can see that both GMM and MLFFNN for chosen parameters and construction give similar classification output. When it comes to SVM that gave the best result in previous task, in this case after reducing the dimensionality it performs very poorly on testing data, that may be caused by overfitting on the testing data. Chosen values of C=10, gamma = 1 might be too aggressive for reduced dataset, leading to high model complexity. When gamma is high, the decision boundary can become overly sensitive to individual data points.

Percentage of bounded and unbounded support vectors out of the training examples, for the Gaussian kernel based model with width = 1 and C = 10.

Kernel	Percentage of Bounded SV	Percentage of Unbounded SV
Gaussian	20%	0%