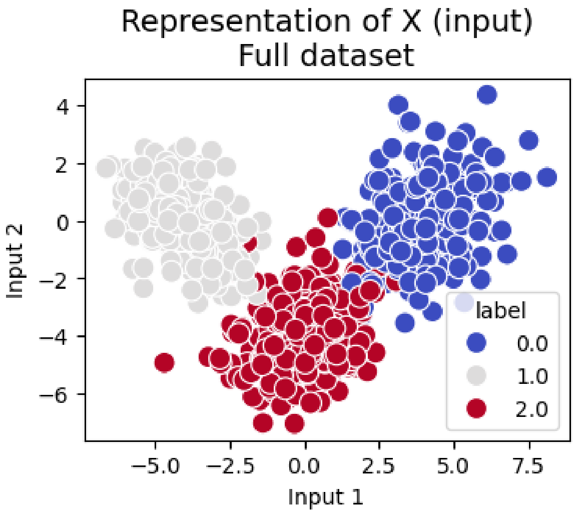
Report Assignment 2 Pattern Recognition & Machine learning

Riccardo Guderzo GE24Z227
Joanna Kolaczek GE24Z229
Miguel Mauer GE24Z022

Task 1: Training dataset 1



1. K-nearest neighbors classifier, for K=1, K=5 and K=9

k=1

Classification accuracy in dataset 1		
Training	Test	Validation
1.00	0.96	0.98

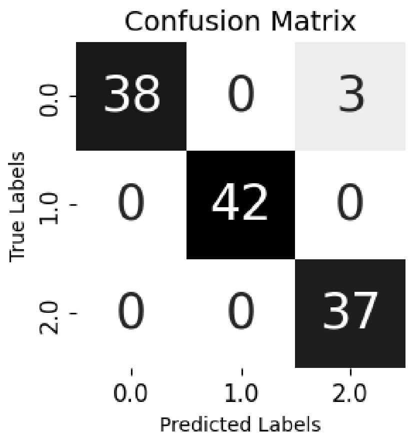
k=5

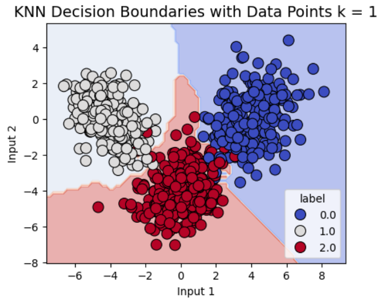
Classification accuracy in dataset 1		
Training	Test	Validation
0.99	0.97	0.98

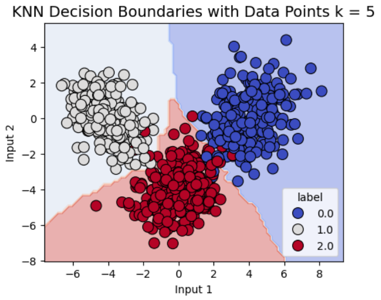
k=9

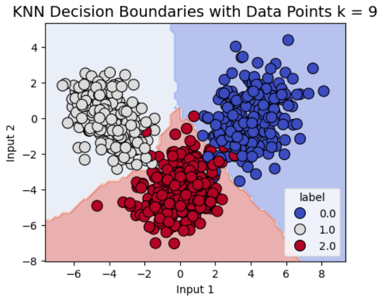
Classification accuracy in dataset 1		
Training	Test	Validation
0.98	0.96	0.98

As the best model configuration we picked the one with k=5. This is what the Confusion Matrix of the model looks like:




KNN Decision Boundaries with Data Points k = 1

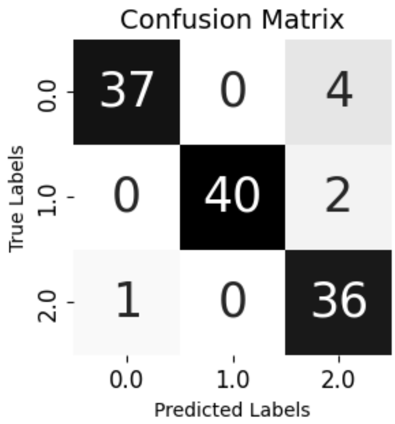

KNN Decision Boundaries with Data Points k = 5

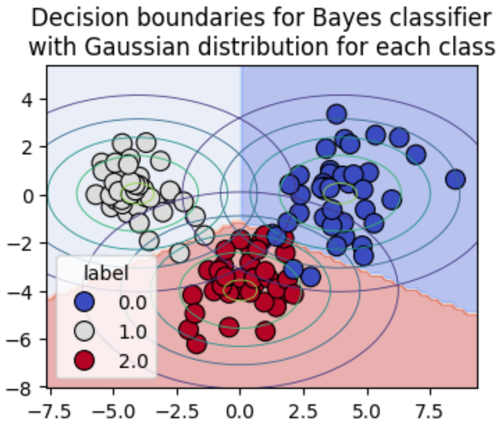

KNN Decision Boundaries with Data Points k = 9

2. Bayes classifier with a Gaussian distribution for every class

a. Covariance matrices for all the classes are the same

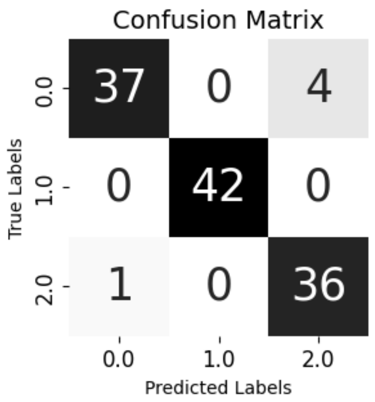
Classification accuracy in dataset 1		
Training	Test	Validation
0.96	0.94	0.94

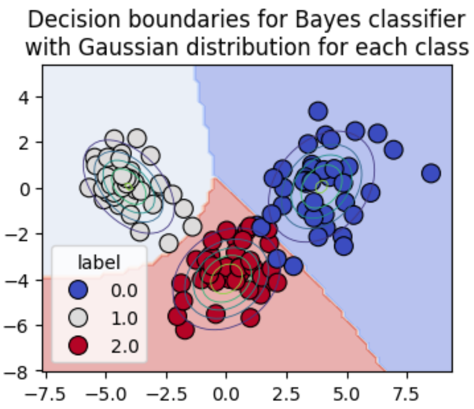



Decision boundaries for Bayes classifier with Gaussian distribution for each class

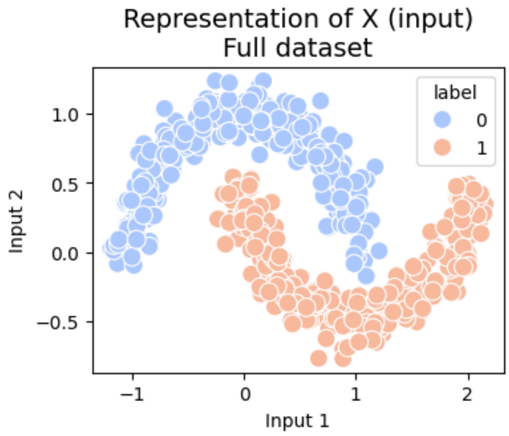
b. Covariance matrices are different

Classification accuracy in dataset 1		
Training	Test	Validation
0.96	0.96	0.96





Task 2: Training dataset 2



In this task we were supposed to classify 2-dimensional data with 2 classes that are non linearly separable. Original dataset with both training and test data is presented in the above plot. In the following points, we're going to compare "KNN-type" (KNN and K-nearest representatives) and "Bayes-type" (Gaussian Bayes and Naive Bayes) cassifiers and try to figure out which one perform the best.

1. K-nearest neighbors classifier, for K=1, K=5 and K=9

k=1

Classification accuracy in dataset 2		
Training	Test	Validation
1.00	1.00	0.99

k=5

Classification accuracy in dataset 2		
Training	Test	Validation
1.00	1.00	1.00

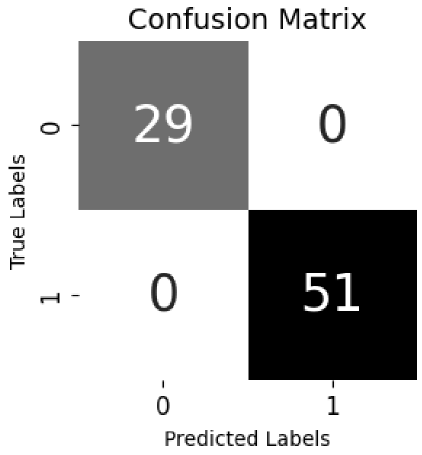
k=9

Classification accuracy in dataset 2

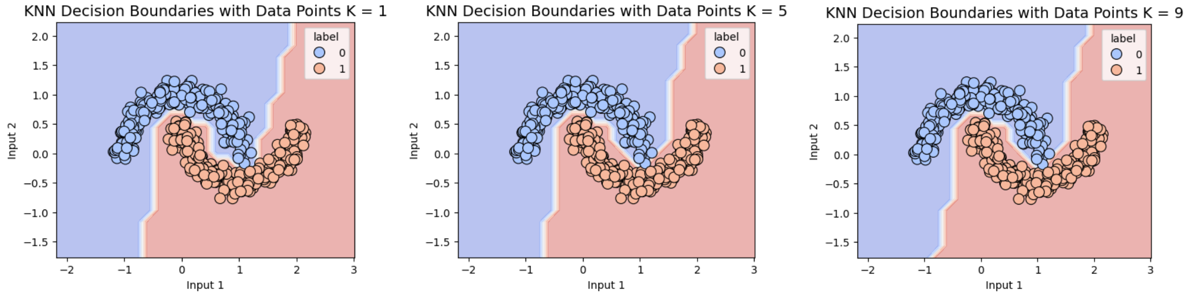
Training	Test	Validation
1.00	1.00	0.99

The KNN classifier was tested with K=1, K=5, and K=9, showing strong performance across all values. For K=1 and K = 9, the model had perfect training and testing accuracy, but slightly worse validation accuracy. For K=5 accuracy 1 was obtained for all datasets and we can observe below classification results as it was chosen as best performing model.

Confusion matrix for k=5



Decision boundaries:


KNN Decision Boundaries with Data Points K = 1

KNN Decision Boundaries with Data Points K = 5

KNN Decision Boundaries with Data Points K = 9

2. K-nearest representatives classifier, for K=1, K=3 and K=5 with 10 representatives per class

k=1

Classification accuracy in dataset 2		
Training	Test	Validation
1.00	1.00	0.99

k=3

Classification accuracy in dataset 2		
Training	Test	Validation

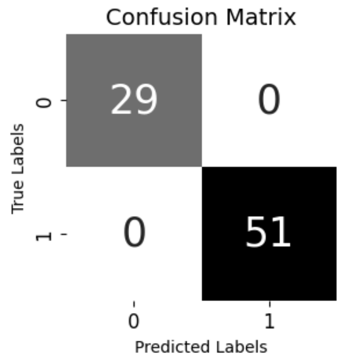
0.99	0.99	0.99

k=5

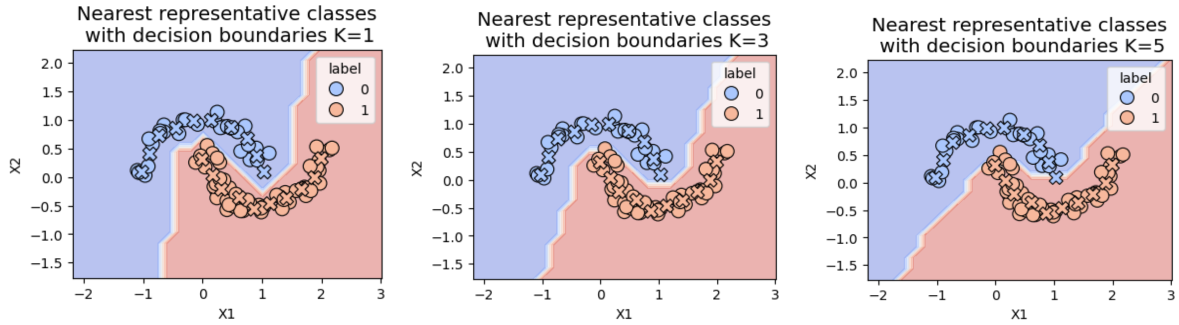
Classification accuracy in dataset 2		
Training	Test	Validation
0.96	0.96	0.97

The K-Nearest Representatives classifier was tested with K=1, K=3, and K=5 using 10 representatives per class, and it shows almost equally good performance. For K=1, both training and test accuracies are 1.00 (see confusion matrix below), with validation accuracy slightly lower at 0.99. This indicates strong performance but may suggest some overfitting. With K=3, training, test, and validation accuracies is 0.99. At K=5, accuracy drops to 0.96 for training and test, with validation at 0.97, indicating a slight decrease in performance.

Confusion matrix for k=1



Decision boundaries (X are class representatives)

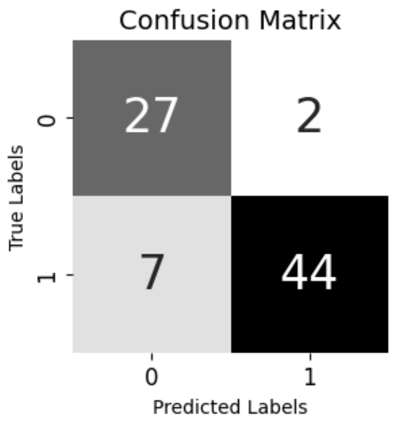


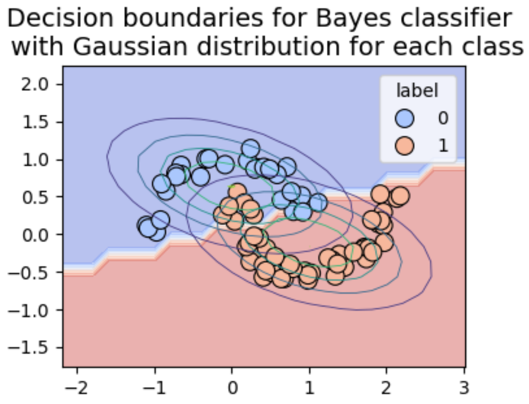
Compared to the KNN classifier, the K-Nearest Representatives method shows slightly lower accuracies overall but allows for generalization with a smaller amount of data. This can be beneficial when the initial dataset is very large, as it reduces computational complexity. However, it is important to select an appropriate number of representatives to be sure that the original class distribution is well represented.

3. Bayes classifier with a Gaussian distribution for every class

a. Covariance matrices for all the classes are the same

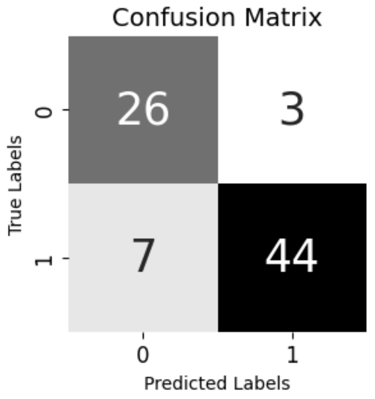
Classification accuracy in dataset 2		
Training	Test	Validation
0.90	0.89	0.86

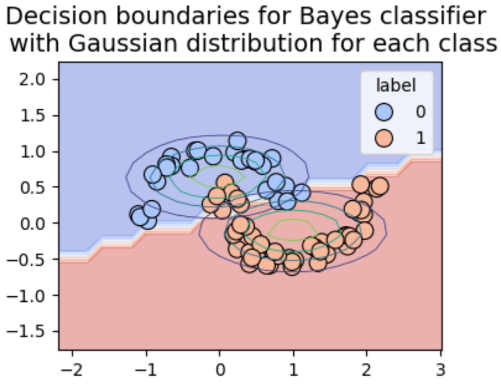




b. Covariance matrices are different

Classification accuracy in dataset 2		
Training	Test	Validation
0.89	0.88	0.86



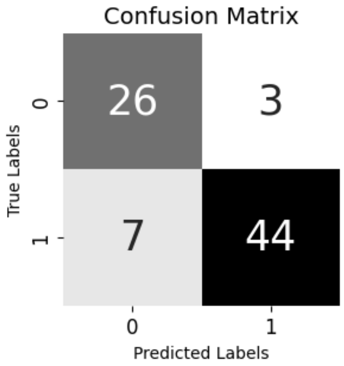


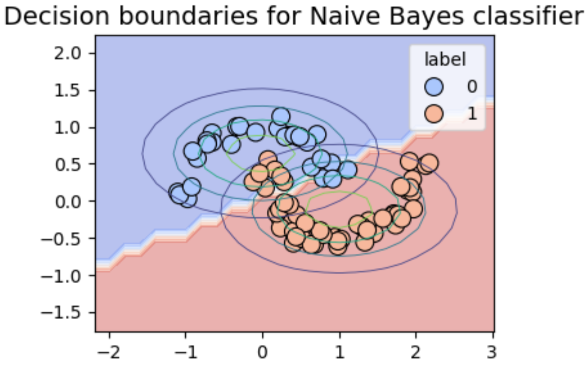
The Bayes classifier with a Gaussian distribution was evaluated in two cases: (a) using the same covariance matrix for all classes, and (b) allowing different covariance matrices for each class. In both cases, the classification accuracies are very similar. For case (a), the model achieved 0.90 training accuracy, 0.89 test accuracy, and 0.86 validation accuracy. In case (b), with different covariance matrices, the training accuracy was 0.89, test accuracy 0.88, and validation accuracy 0.86. This suggests that allowing different covariance matrices in this case does not significantly impact the model's performance, with both approaches returning comparable results.

4. Naive-Bayes classifier with a Gaussian distribution for every class:

a. Covariance matrices for all the classes are the same

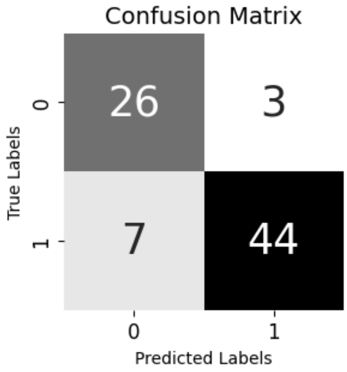
Classification accuracy in dataset 2		
Training	Test	Validation
0.88	0.88	0.83

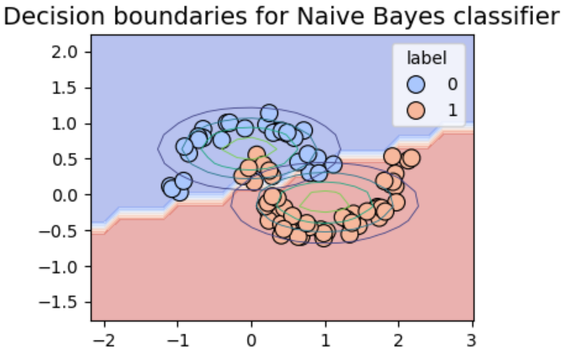




b. Covariance matrices are different

Classification accuracy in dataset 2		
Training	Test	Validation
0.88	0.88	0.86





The Naive-Bayes classifier was tested under two conditions: (a) with the same covariance matrix for all classes, and (b) with different covariance matrices for each class. In both cases, the training and test accuracies are identical at 0.88. However, the validation accuracy differs slightly, with case (a) achieving 0.83 and case (b) achieving 0.86. This indicates that in this case allowing different covariance matrices improves the model's validation.

In general, “KNN-type” classifiers performed better in this task because they classify points based on distance to neighbors, allowing them to capture non-linear patterns in the data. “Bayes-type” classifiers rely on assumptions of Gaussian distributions for each class, which may not fit non-linearly separable data well and lead to lower performance.

Task 3: Training dataset 3

1. K-nearest neighbors classifier, for K=1, K=9 and K=15

k=1

Classification accuracy in dataset 3		
Training	Test	Validation
	0.50	0.52

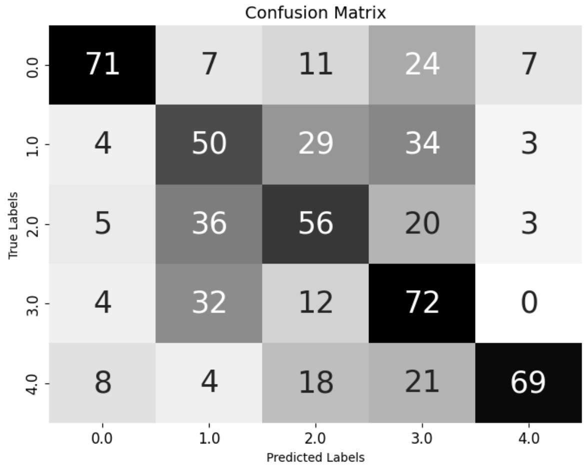
k=9

Classification accuracy in dataset 3		
Training	Test	Validation
	0.53	0.60

k=15

Classification accuracy in dataset 3		
Training	Test	Validation
	0.53	0.58

Using k=9 the validation data shows higher accuracy. Here is the Confusion Matrix for k=9:



	0.0	1.0	2.0	3.0	4.0
0.0	71	7	11	24	7
1.0	4	50	29	34	3
2.0	5	36	56	20	3
3.0	4	32	12	72	0
4.0	8	4	18	21	69

2. K-nearest representatives classifier, for K=1, K=5 and K=9 with 10 representatives per class

Using the same M=10 for all of the k values:
k=1

Classification accuracy in dataset 3		
Training	Test	Validation
0.67	0.63	0.66

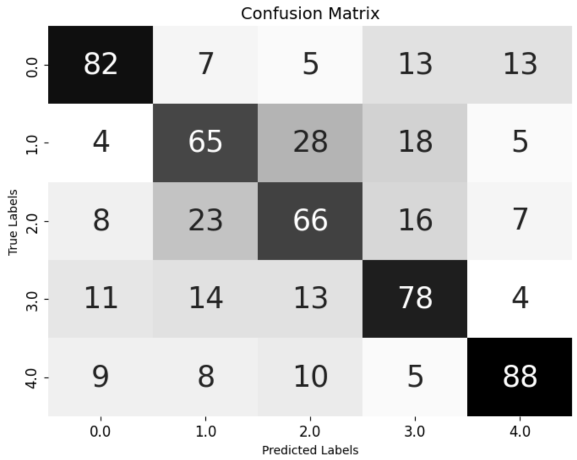
k=5

Classification accuracy in dataset 3		
Training	Test	Validation
0.52	0.51	0.49

k=9

Classification accuracy in dataset 3		
Training	Test	Validation
0.47	0.46	0.46

The confusion matrix for the best model configuration M=10, k=1:



	0.0	1.0	2.0	3.0	4.0
0.0	82	7	5	13	13
1.0	4	65	28	18	5
2.0	8	23	66	16	7
3.0	11	14	13	78	4
4.0	9	8	10	5	88

3. Bayes classifier with a Gaussian distribution for every class

Covariance matrices for all the classes are the same:

Classification accuracy in dataset 3		
Training	Test	Validation
0.52	0.52	0.55

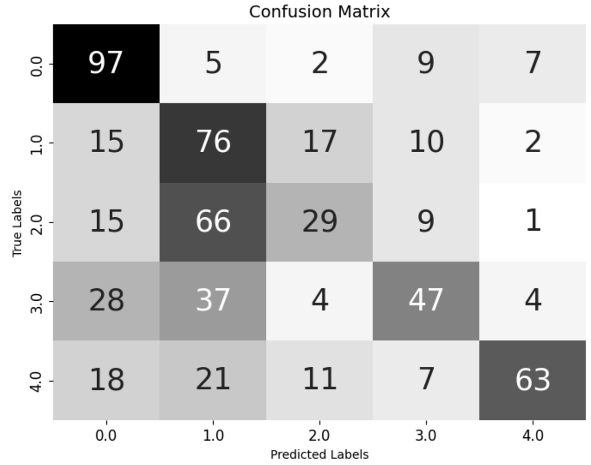
Covariance matrices for all the classes are different:

Classification accuracy in dataset 3		
Training	Test	Validation

0.56	0.59	0.59

For the Bayes classifier with a Gaussian distribution, the model's performance improves when the covariance matrices are allowed to differ for each class. When the covariance matrices are the same across all classes, performance drops to 0.52 for both training and test accuracy, with 0.55 validation accuracy.

Confusion matrix for Bayes classifier with Gaussian distribution for every class, when covariance matrices are different for each class:



4. Naive-Bayes classifier with a Gaussian distribution for every class

Covariance matrices for all the classes are the same:

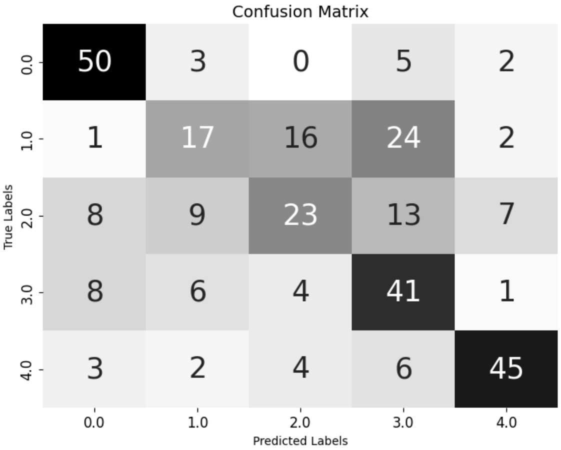
Classification accuracy in dataset 3		
Training	Test	Validation
0.51	0.50	0.58

Covariance matrices for all the classes are different:

Classification accuracy in dataset 3		
Training	Test	Validation
0.53	0.52	0.59

Naive-Bayes classifier like Bayes classifier with a Gaussian distribution also performs better when the covariance matrices differ for each class. With a training accuracy of 0.51, test accuracy of 0.50, and validation accuracy of 0.58 for the same covariance matrices. When the covariance matrices are different, the accuracies increase to 0.53 for training, 0.52 for test, and 0.59 for validation.

Confusion matrix for Naive-Bayes classifier, when covariance matrixes are different for each class:



	0.0	1.0	2.0	3.0	4.0
0.0	50	3	0	5	2
1.0	1	17	16	24	2
2.0	8	9	23	13	7
3.0	8	6	4	41	1
4.0	3	2	4	6	45