Programming Assignment - 1

Report Submitted in Partial Fulfillment of the Requirements for the Course

CS401 Introduction to Machine Learning

Submitted by

Group number 10

Rakshitha Kalkura (Roll no. 22CSE1028)

Deeksha Yadav (Roll no. 21CSE1007)

Department of Computer Science and Engineering

National Institute of Technology Goa

November 11, 2024

Contents

1 Introduction

In machine learning and pattern recognition, the goal is to develop algorithms that can automatically classify data into distinct groups based on their inherent features and patterns. This task is essential in various fields, including image recognition, speech processing, and predictive modeling. However, real-world datasets often present different levels of complexity. These can range from simple linear separability, where classes can be easily distinguished by a straight line, to more complex non-linear relationships that require more advanced algorithms. Additionally, datasets often feature overlapping instances, where classes are not clearly separable, posing additional challenges for classification.

To effectively classify data, it is crucial to select algorithms that are well-suited to the specific characteristics of the dataset. The choice of classifier depends on whether the data is linearly separable, non-linearly separable, or contains overlapping instances. Simple algorithms may work well for straightforward tasks, but more complex datasets demand advanced techniques that can capture intricate patterns and boundaries between classes.

This report aims to assess the performance of several machine learning classifiers—specifically the Neural Network (NN) Classifier, the K-Nearest Neighbors (KNN) Classifier, the Reference Template-Based Classifier, and the Bayes Classifier—on datasets with varying characteristics. By examining how each of these algorithms handles linearly separable, non-linearly separable, and overlapping data, the report will provide insights into the most effective classification methods for different types of problems.

2 Objective

To observe and assess the performance of different classifier algorithms on dataset with varied characteristics.

3 Datasets

3.1 Linearly Separable

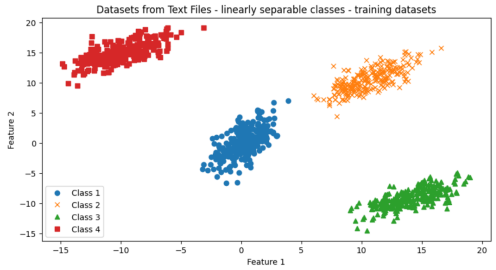


Figure 1: Linearly Separable Training Dataset

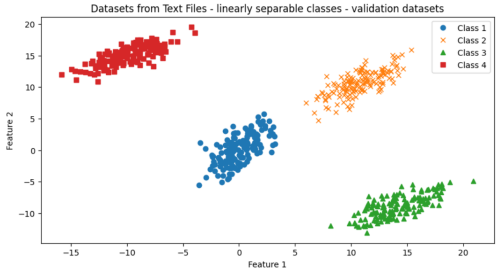


Figure 2: Linearly Separable Validation Dataset

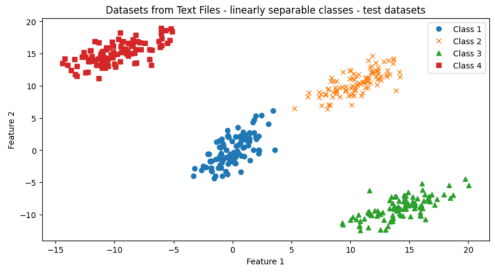


Figure 3: Linearly Separable Testing Dataset

3.2 Non Linearly Separable

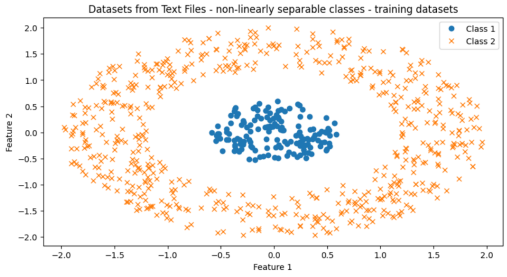


Figure 4: Non Linearly Separable Training Dataset

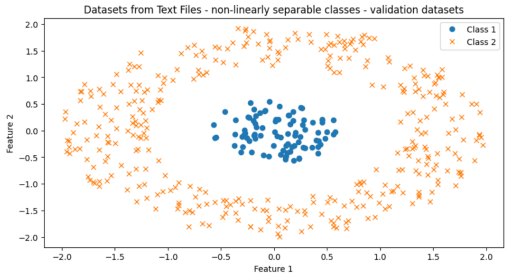


Figure 5: Non Linearly Separable Validation Dataset

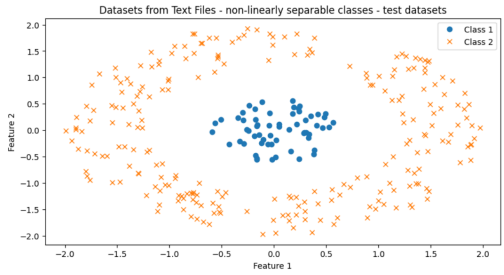


Figure 6: Non Linearly Separable Testing Dataset

3.3 Overlapping

Figure 7: Overlapping Training Dataset

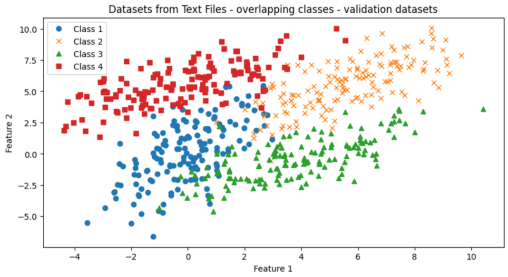


Figure 8: Overlapping Validation Dataset

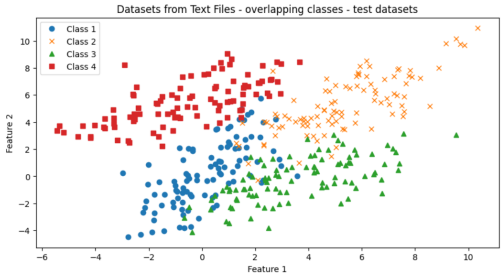


Figure 9: Overlapping Testing Dataset

4 Dataset with Mean

4.1 Linearly Separable Dataset

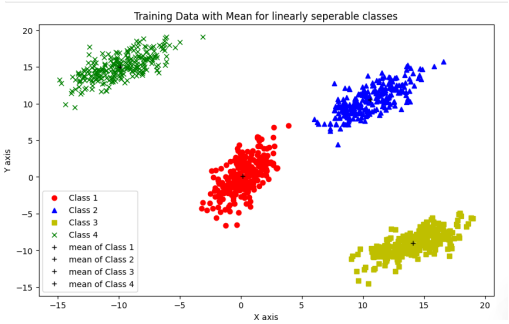


Figure 10: Linearly Separable Dataset with Mean

This dataset contains four classes of data points that are cleanly separable by decision boundaries. It provides a scenario to examine how classifiers perform when relationships between classes are straightforward.

4.2 Non Linearly Separable Dataset

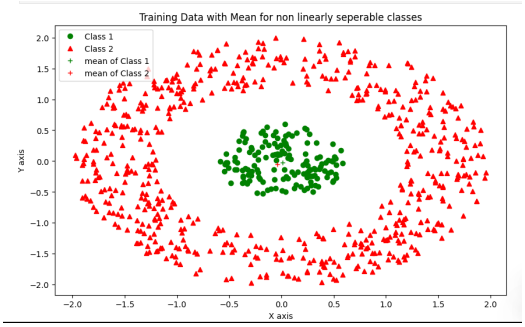


Figure 11: Non Linearly Separable Dataset with mean

This dataset contains two classes of data points that are non-linearly separable. Class 1 (blue circles) is clustered at the center, while Class 2 (orange Xs) surrounds Class 1 in an outer ring. This setup helps to evaluate classifier performance in scenarios requiring complex, non-linear decision boundaries.

4.3 Overlapping Dataset

Figure 12: Overlapping Dataset with Mean

This dataset consists of four overlapping classes of data points, each represented by a unique marker and color, which adds complexity to the separation task. Class 1, shown with blue circles, is primarily centered in the lower-left region, while Class 2, depicted with orange Xs, occupies the upper-right area. Class 3, represented by green triangles, is spread around the middle-lower part of the plot, and Class 4, shown with red squares, is positioned toward the upper-left. The overlap between classes makes it challenging for classifiers, providing an opportunity to evaluate performance in scenarios with less distinct class boundaries.

5 NN Classifier

NN Classifier based on the principle of finding the closest data points in the training data-set to a new, unseen data point and making predictions based on the labels or values associated with those nearest neighbors.

5.1 Linear Dataset

5.1.1 Metric Summary

Table 1: Classwise recall, precision and F-measure score

Index	Class1	Class2	Class3	Class4
Precision	1.0	1.0	1.0	1.0
Recall	1.0	1.0	1.0	1.0
F1 Score	1.0	1.0	1.0	1.0

Table 2: Mean recall, precision and F-measure score

Index	Value
Accuracy Score	1.0
Mean Precision	1.0
Mean Recall	1.0
Mean F1	1.0

5.1.2 Confusion Matrix

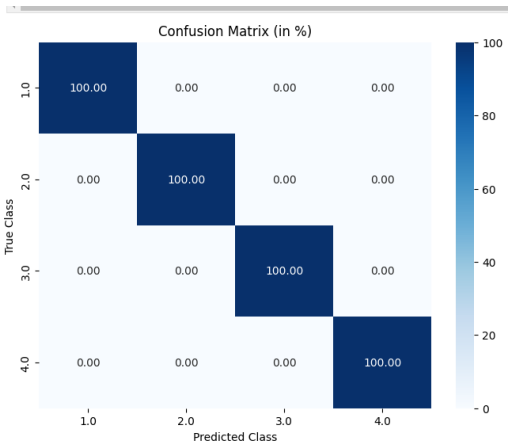


Figure 13: Confusion Matrix

5.1.3 Decision boundary for every pair of class.

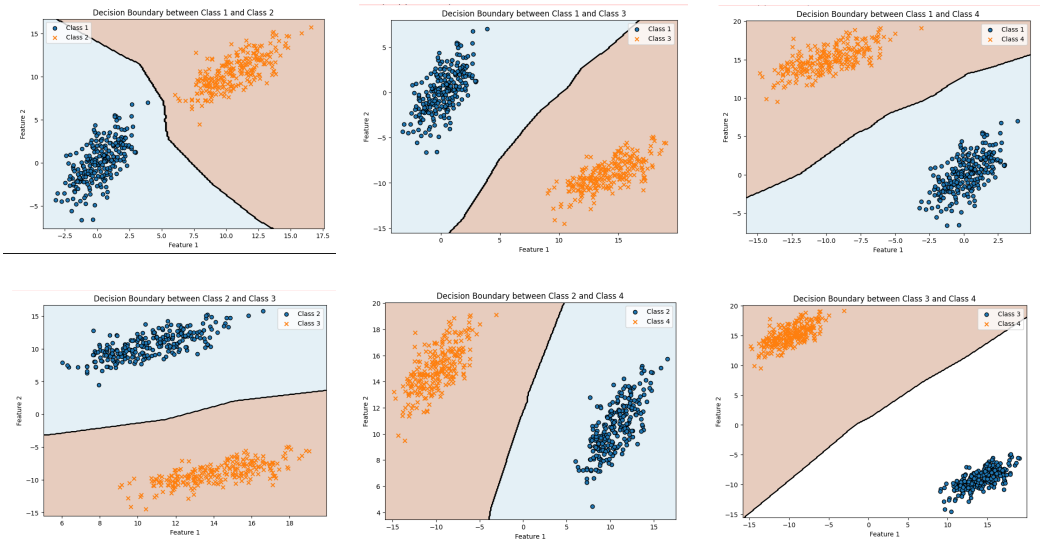


Figure 14: Pairwise Plot

5.1.4 Decision Boundary of Training Data superimposed

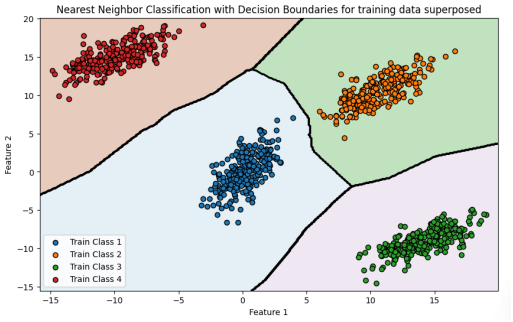


Figure 15: Decision Boundary of Training Data superimposed

5.1.5 Decision Boundary of Testing Data superimposed

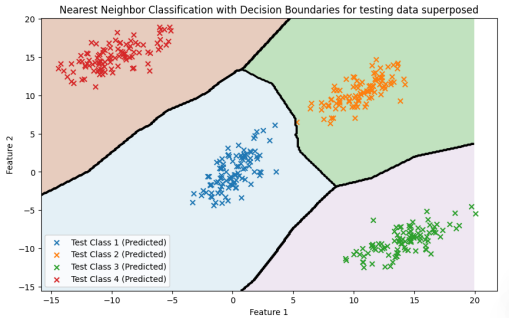


Figure 16: Decision Boundary of Testing Data superimposed

5.1.6 Decision Boundary of Training and Testing Data superimposed

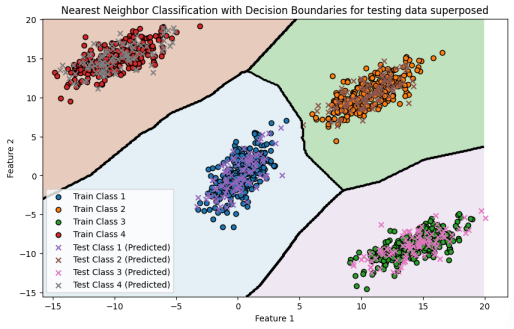


Figure 17: Decision Boundary of Training and Testing Data superimposed

5.2 Non Linear Dataset

5.2.1 Metric Summary

Table 3: Classwise recall, precision and F-measure score

Index	Class1	Class2
Precision	1.0	1.0
Recall	1.0	1.0
F1 Score	1.0	1.0

Table 4: Mean recall, precision and F-measure score

Index	Value
Accuracy Score	1.0
Mean Precision	1.0
Mean Recall	1.0
Mean F1	1.0

5.2.2 Confusion Matrix

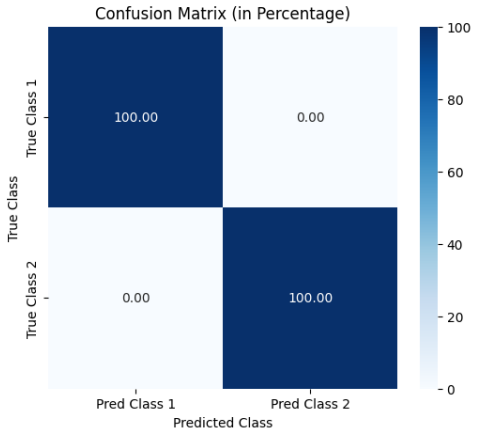


Figure 18: Confusion Matrix

5.2.3 Decision Boundary of Training Data superimposed

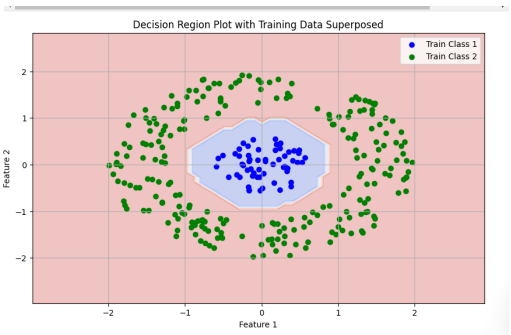


Figure 19: Decision Boundary of Training Data superimposed

5.2.4 Decision Boundary of Testing Data superimposed

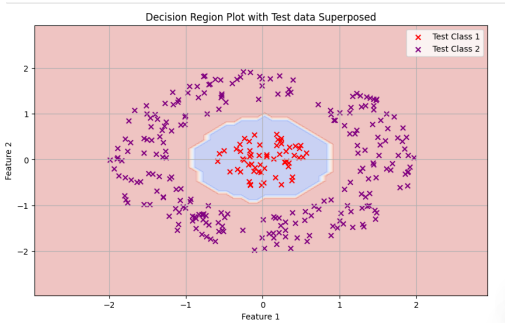


Figure 20: Decision Boundary of Testing Data superimposed

5.2.5 Decision Boundary of Training and Testing Data superimposed

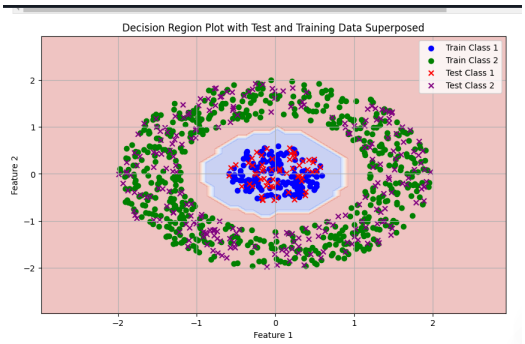


Figure 21: Decision Boundary of Training and Testing Data superimposed

5.3 Overlapping Dataset

5.3.1 Metric Summary

Table 5: Classwise recall, precision and F-measure score

Index	Class1	Class2	Class3	Class4
Precision	0.736842	0.933333	0.879121	0.923810
Recall	0.8400	0.8400	0.8000	0.9700
F1 Score	0.785047	0.884211	0.837696	0.946341

Table 6: Mean recall, precision and F-measure score

Index	Value
Accuracy Score	0.8625
Mean Precision	0.868276
Mean Recall	0.8625
Mean F1	0.863324

5.3.2 Confusion Matrix

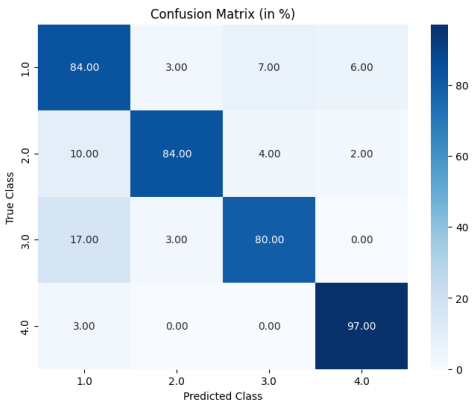


Figure 22: Confusion Matrix

5.3.3 Decision boundary for every pair of class.

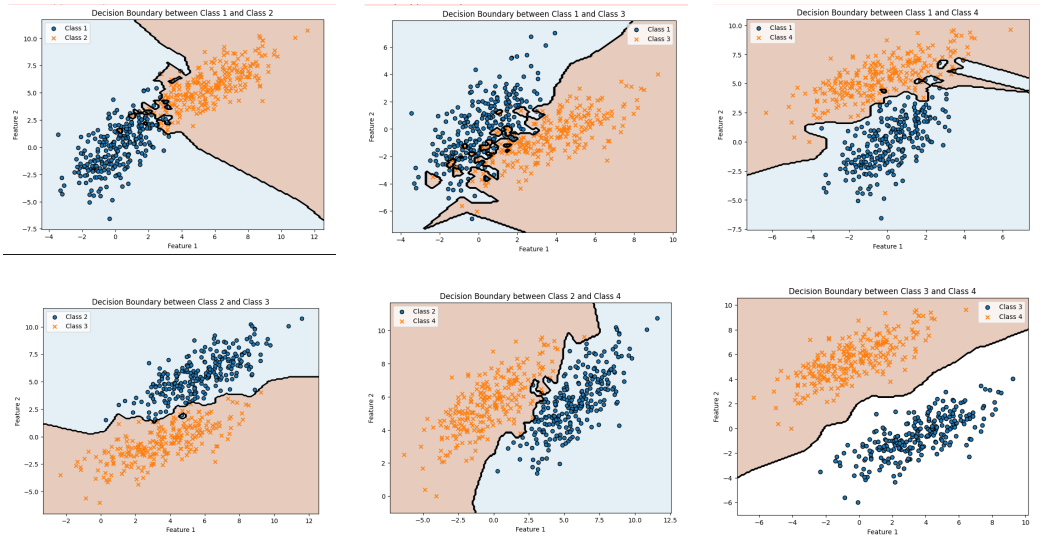


Figure 23: Pairwise Plot

5.3.4 Decision Boundary of Training Data superimposed

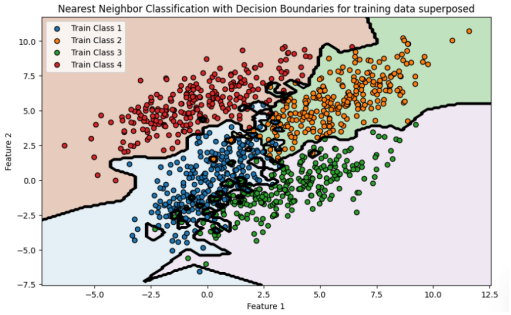


Figure 24: Decision Boundary of Training Data superimposed

5.3.5 Decision Boundary of Testing Data superimposed

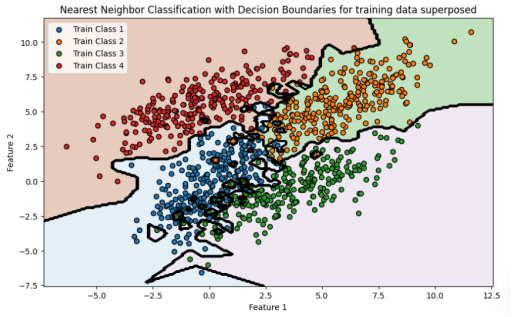


Figure 25: Decision Boundary of Testing Data superimposed

5.3.6 Decision Boundary of Training and Testing Data superimposed

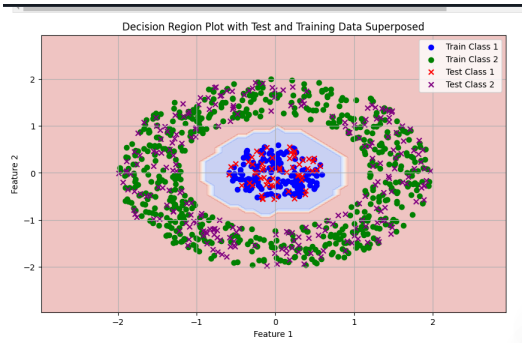


Figure 26: Decision Boundary of Training and Testing Data superimposed

6 K-Nearest Neighbour Classifier

In KNN, an object is classified or its value is predicted based on the majority class or the average of its k-nearest neighbors in a feature space, where "k" is a user-defined parameter representing the number of neighboring data points considered. It relies on the assumption that similar data points are likely to have similar outcomes.

6.1 Linear Dataset

After plotting the Elbow curve for the KNN algorithm applied on the linear dataset, we figured out that the optimal value of K = 1. That concludes that the best KNN algorithm will give most accurate prediction when the neared neighbour is chosen to be 1 which is nothing but nn-classifier.

6.1.1 Accuracy Value for different K values

```
Accuracy for k=1: 1.0000
Accuracy for k=3: 1.0000
Accuracy for k=5: 1.0000
Accuracy for k=7: 1.0000
Accuracy for k=9: 1.0000
Accuracy for k=11: 1.0000
Accuracy for k=15: 1.0000
Best k value: 1 with accuracy: 1.0000
Test Accuracy for best k=1: 1.0000
```

Figure 27: Accuracy Values

6.1.2 Elbow Curve

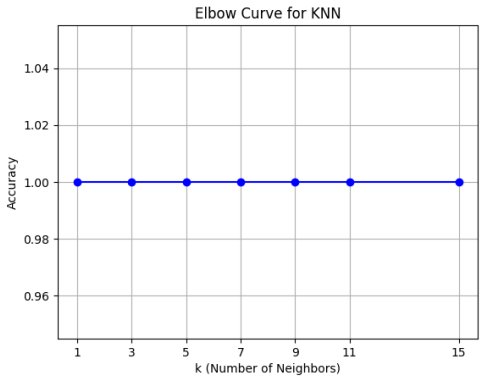


Figure 28: Elbow curve

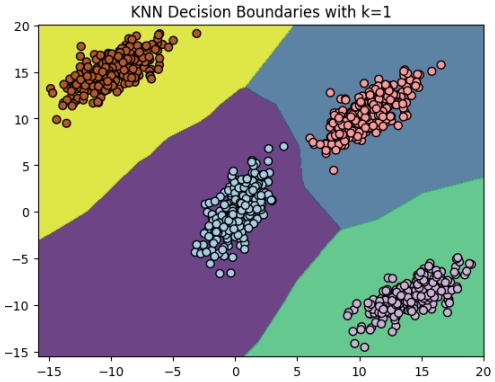


Figure 29: Decision Boundary for K=1

For linearly separable classes, the KNN classifier with k = 1 (nearest neighbor classifier) performs optimally and is equivalent to the nearest neighbor classifier.

6.2 Non Linear Dataset

Similar to section 3.1, the optimal value of K for Non-Linear dataset is coming out to be 1. This concludes that the best nn-classifier will be suitable for the Non-linear dataset

6.2.1 Accuracy Value for different K values

```
Accuracy for k=1: 1.0000
Accuracy for k=3: 1.0000
Accuracy for k=5: 1.0000
Accuracy for k=7: 1.0000
Accuracy for k=9: 1.0000
Accuracy for k=11: 1.0000
Accuracy for k=15: 1.0000
Best k value: 1 with accuracy: 1.0000
Test Accuracy for best k=1: 1.0000
```

Figure 30: Accuracy Values

6.2.2 Elbow Curve

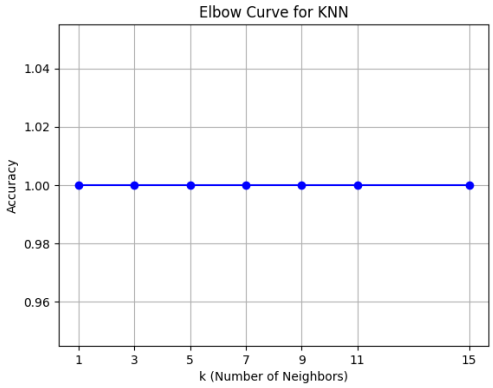


Figure 31: Elbow Curve

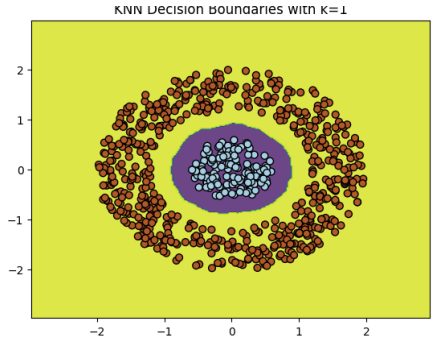


Figure 32: Decision Boundary for K=1

KNN with K=1 and the nearest neighbor classifier both focus on the nearest training data point for classification. This leads to identical performance characteristics on non-linearly separable datasets, with both methods being highly sensitive to local variations and noise in the data. Consequently, they can achieve similar accuracies, especially when there is no clear, linear separation between classes.

6.3 Overlapping Dataset

The optimal value of K turns out to be 9

6.3.1 Accuracy Value for different K values

```
Accuracy for k=1: 0.8667
Accuracy for k=3: 0.8800
Accuracy for k=5: 0.8956
Accuracy for k=7: 0.8933
Accuracy for k=9: 0.8978
Accuracy for k=11: 0.8956
Accuracy for k=15: 0.8978
Accuracy for k=17: 0.8911
Accuracy for k=19: 0.8889
Accuracy for k=21: 0.8911
Accuracy for k=23: 0.8911
Best k value: 9 with accuracy: 0.8978
Test Accuracy for best k=9: 0.9220
```

Figure 33: Accuracy Values

6.3.2 Elbow Curve

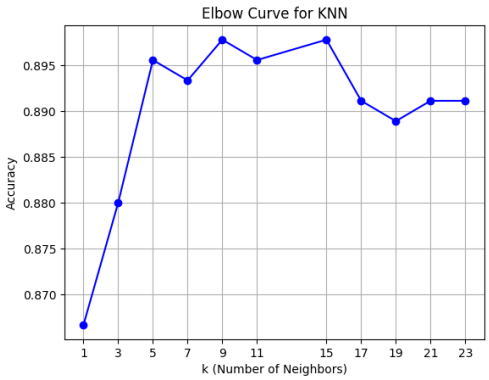


Figure 34: Elbow Curve

6.4 Decision Boundary with K=9

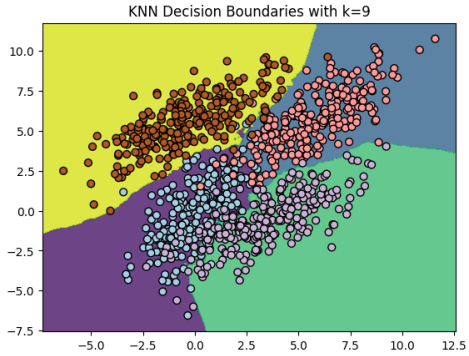


Figure 35: Decision Boundary with K=9

6.4.1 Metric Summary

Table 7: Classwise recall, precision and F-measure score

Index	Class1	Class2	Class3	Class4
Precision	0.777778	0.946237	0.880435	0.925234
Recall	0.84	0.88	0.81	0.99
F1 Score	0.807692	0.911917	0.843750	0.956522

Table 8: Mean recall, precision and F-measure score

Index	Value
Accuracy Score	0.88
Mean Precision	0.882421
Mean Recall	0.88
Mean F1	0.879970

6.4.2 Confusion Matrix

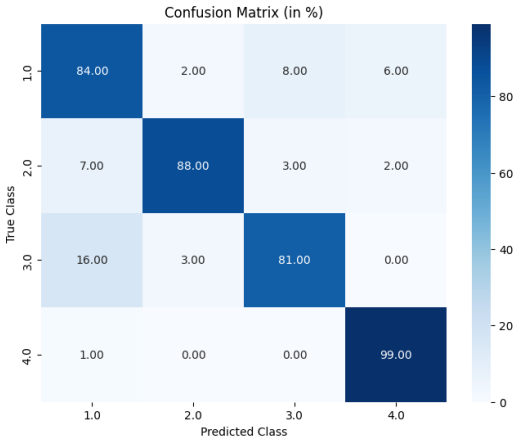


Figure 36: Confusion Matrix

6.4.3 Decision boundary for every pair of class.

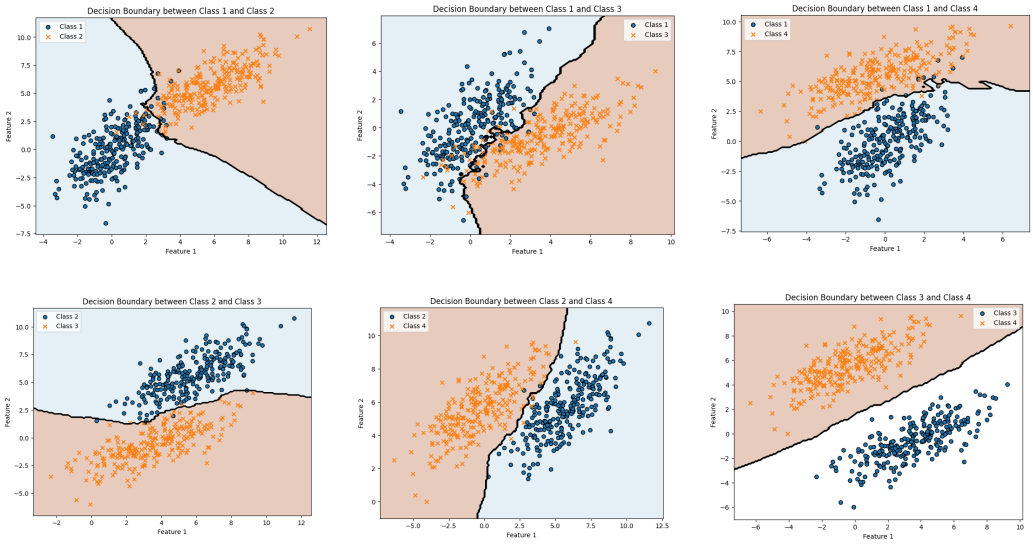


Figure 37: Pairwise Plot

6.4.4 Decision Boundary of Training Data superimposed

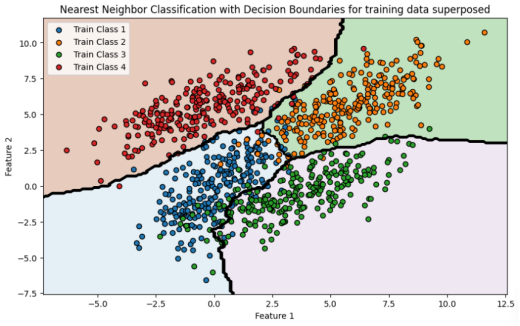


Figure 38: Decision Boundary for Training Dataset

6.4.5 Decision Boundary of Testing Data superimposed

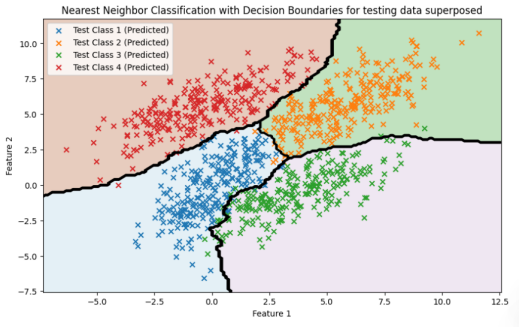


Figure 39: Decision Boundary for Testing Dataset

7 Mean Vector as Reference Template Based Classifier

The Mean Vector as a Reference Template-Based Classifier calculates the Euclidean distance between a data point and the mean distribution of a particular class. It then assigns the data point to the class for which the Euclidean distance is minimized, indicating the class that is the closest match to the data point in the feature space. This method leverages the central

tendency of each class to make classification decisions, choosing the class whose mean is most similar to the input data point.

7.1 Linear Dataset

7.1.1 Metric Summary

Table 9: Classwise recall, precision and F-measure score

Index	Class1	Class2	Class3	Class4
Precision	1.0	1.0	1.0	1.0
Recall	1.0	1.0	1.0	1.0
F1 Score	1.0	1.0	1.0	1.0

Table 10: Mean recall, precision and F-measure score

Index	Value
Accuracy Score	100%
Mean Precision	1.0
Mean Recall	1.0
Mean F1	1.0

7.1.2 Confusion Matrix

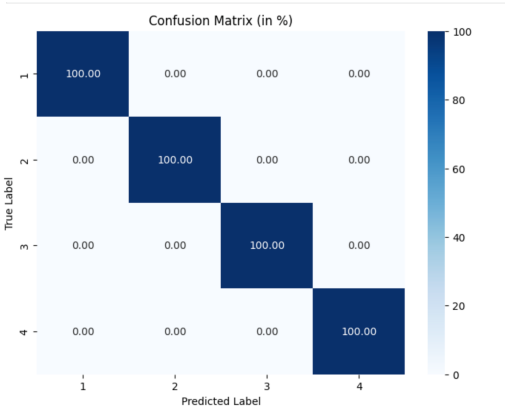


Figure 40: Confusion Matrix

7.1.3 Decision boundary for every pair of class.

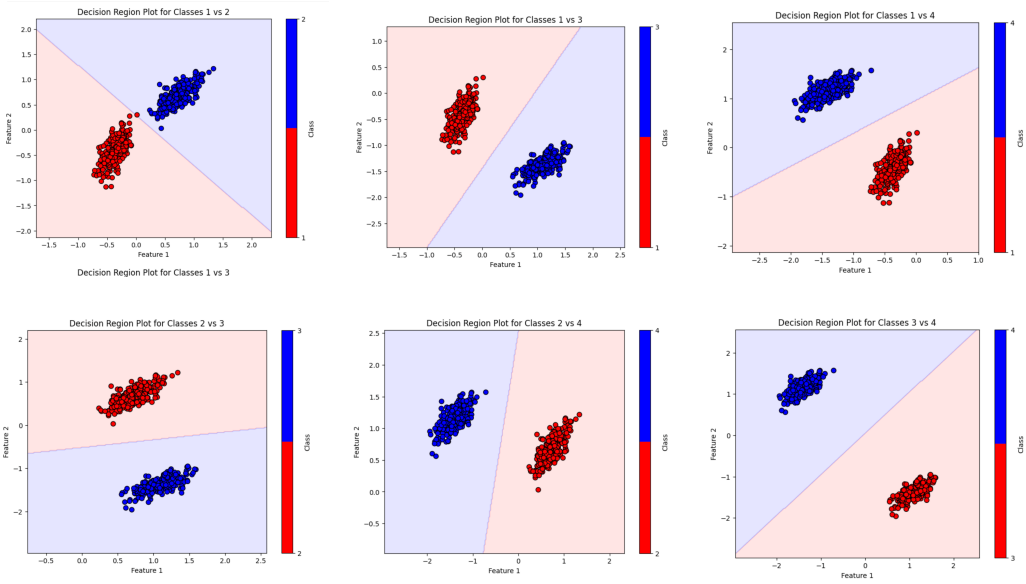


Figure 41: Pairwise Plot

7.1.4 Decision Boundary of Training Data superimposed

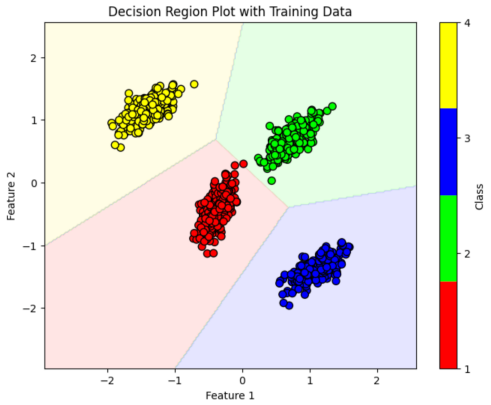


Figure 42: Decision Boundary for Training Dataset

7.1.5 Decision Boundary of Testing Data superimposed

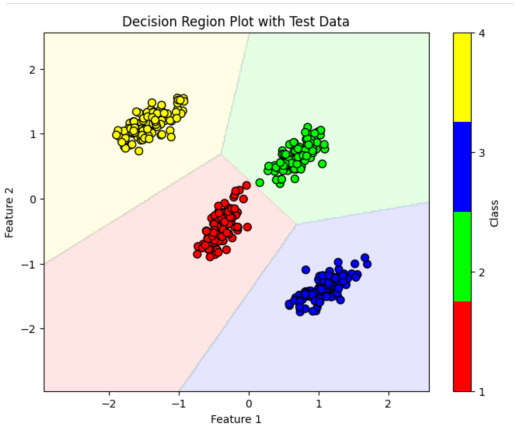


Figure 43: Decision Boundary for Testing Dataset

7.2 Non Linear Dataset

7.2.1 Metric Summary

Table 11: Classwise recall, precision and F-measure score

Index	Class1	Class2
Precision	0.212329	0.811688
Recall	0.516667	0.520833
F1 Score	0.300971	0.634518

Table 12: Mean recall, precision and F-measure score

Index	Value
Accuracy Score	52.00%
Mean Precision	0.512009
Mean Recall	0.518750
Mean F1	0.467744

7.2.2 Confusion Matrix

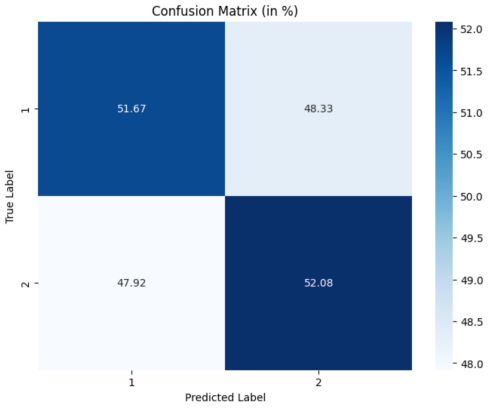


Figure 44: Confusion Matrix

7.2.3 Decision Boundary of Training Data superimposed

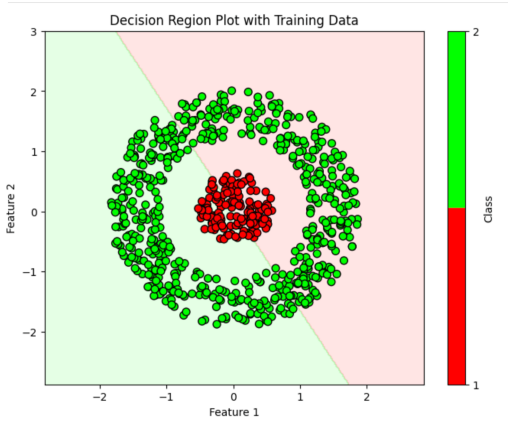


Figure 45: Decision Boundary for Training Dataset

7.2.4 Decision Boundary of Testing Data superimposed

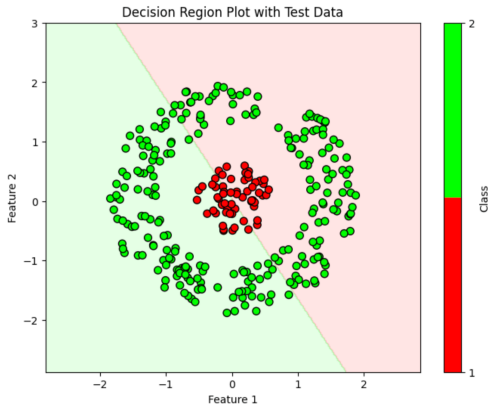


Figure 46: Decision Boundary for Testing Dataset

7.3 Overlapping Dataset

7.3.1 Metric Summary

Table 13: Classwise recall, precision and F-measure score

Index	Class1	Class2	Class3	Class4
Precision	0.750000	0.858586	0.823529	0.861111
Recall	0.8100	0.8500	0.7000	0.9300
F1 Score	0.778846	0.854271	0.756757	0.894231

Table 14: Mean recall, precision and F-measure score

Index	Value
Accuracy Score	82.25%
Mean Precision	0.823307
Mean Recall	0.8225
Mean F1	0.821026

7.3.2 Confusion Matrix

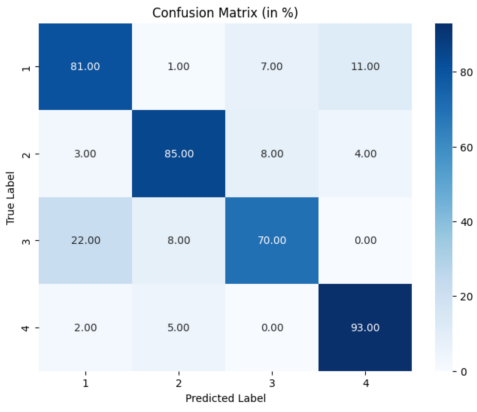


Figure 47: Confusion Matrix

7.3.3 Decision boundary for every pair of class.

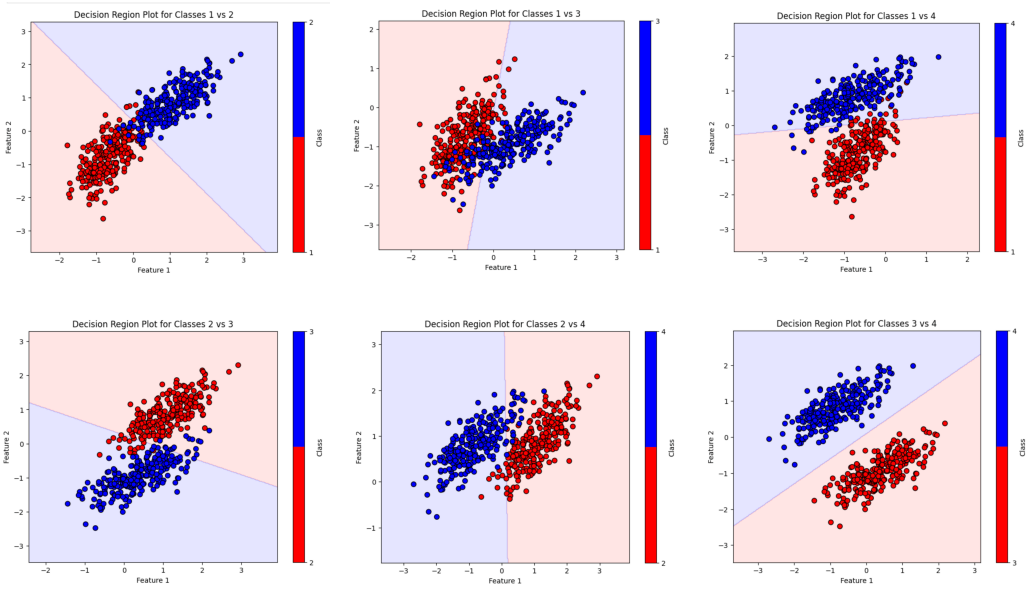


Figure 48: Pairwise Plot

7.3.4 Decision Boundary of Training Data superimposed

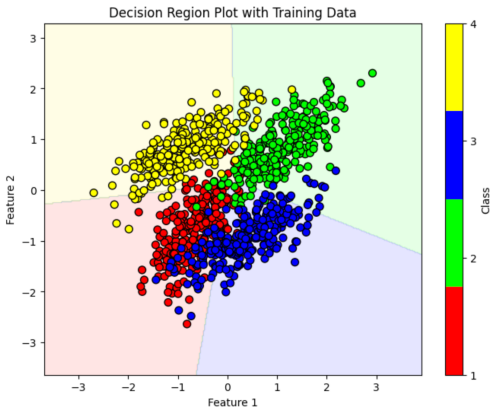


Figure 49: Decision Boundary for Training Dataset

7.3.5 Decision Boundary of Testing Data superimposed

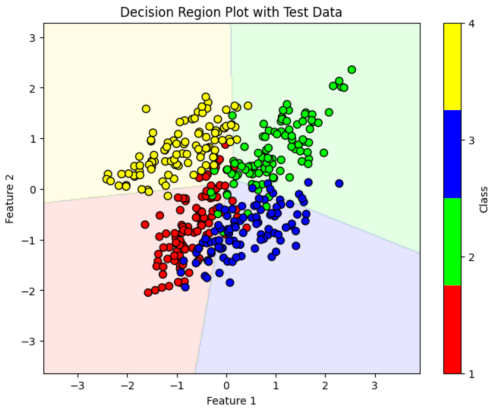


Figure 50: Decision Boundary for Testing Dataset

8 Meanvector and covariance matrix as reference template for a class

The Mean Vector as a Reference Template-Based Classifier calculates the Mahalodian distance between a data point and the mean distribution of a particular class. It then assigns the data point to the class for which the Mahalodian distance is minimized.

8.1 Linear Dataset

8.1.1 Metric Summary

Table 15: Classwise recall, precision and F-measure score

Index	Class1	Class2	Class3	Class4
Precision	1.0	1.0	1.0	1.0
Recall	1.0	1.0	1.0	1.0
F1 Score	1.0	1.0	1.0	1.0

Table 16: Mean recall, precision and F-measure score

Index	Value
Accuracy Score	100.00%
Mean Precision	1.0
Mean Recall	1.0
Mean F1	1.0

8.1.2 Confusion Matrix

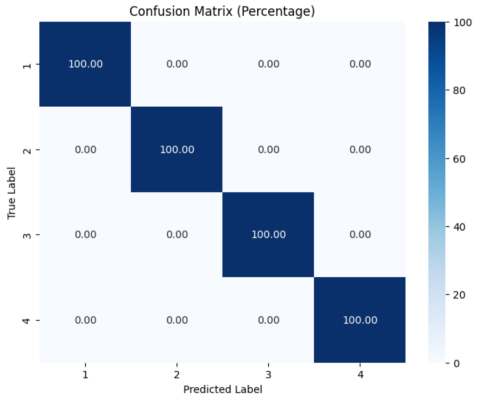


Figure 51: Confusion Matrix

8.1.3 Decision boundary for every pair of class.

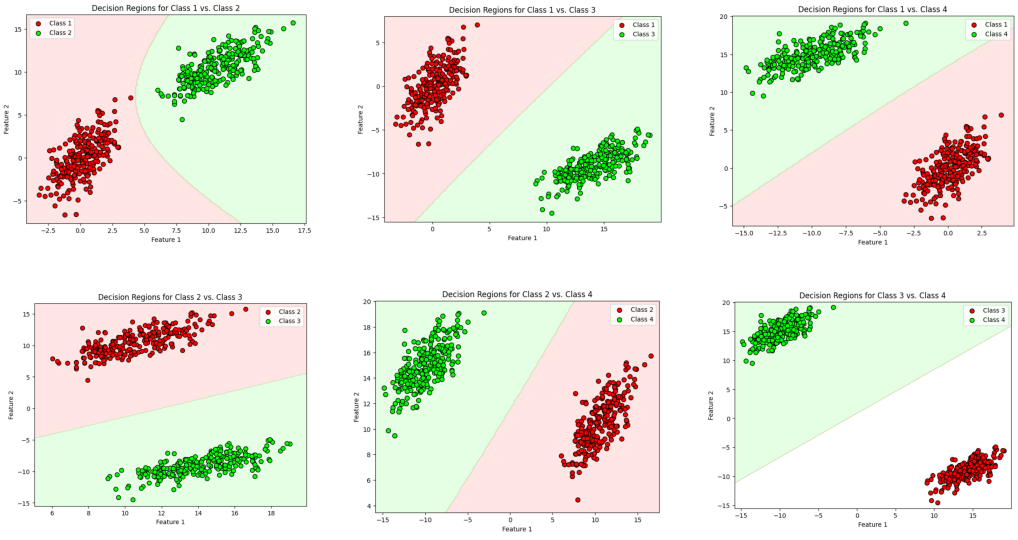


Figure 52: Pairwise Plot

8.1.4 Decision Boundary of Training Data superimposed

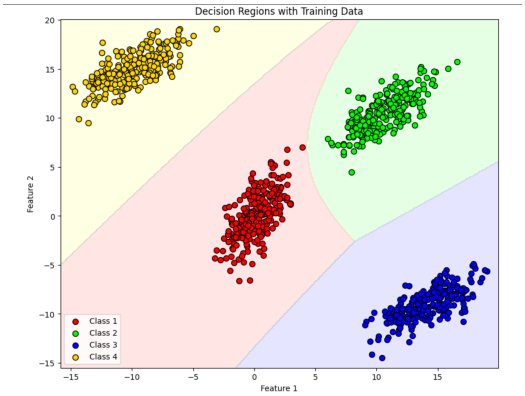


Figure 53: Decision Boundary for Training Dataset

8.1.5 Decision Boundary of Testing Data superimposed

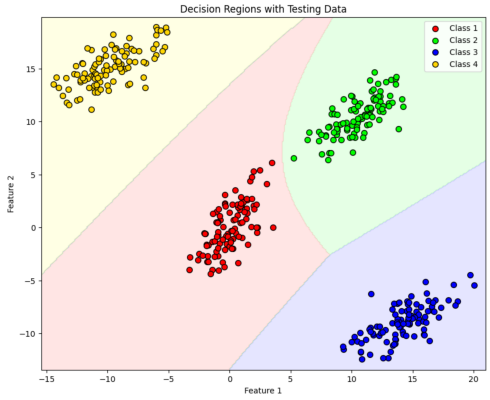


Figure 54: Decision Boundary for Testing Dataset

8.2 Non Linear Dataset

8.2.1 Metric Summary

Table 17: Classwise recall, precision and F-measure score

Index	Class1	Class2
Precision	0.00	0.8
Recall	0.0	1.0
F1 Score	0.000000	0.888889

Table 18: Mean recall, precision and F-measure score

Index	Value
Accuracy Score	0.80
Mean Precision	0.4
Mean Recall	0.5
Mean F1	0.444444

8.2.2 Confusion Matrix

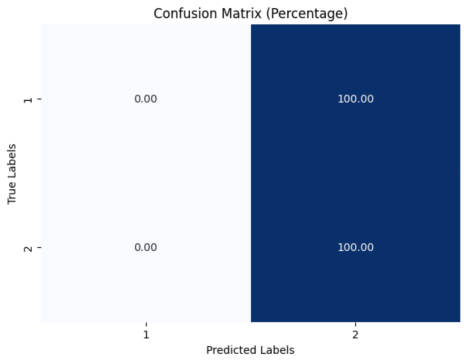


Figure 55: Confusion Matrix

8.2.3 Validation and Accuracy

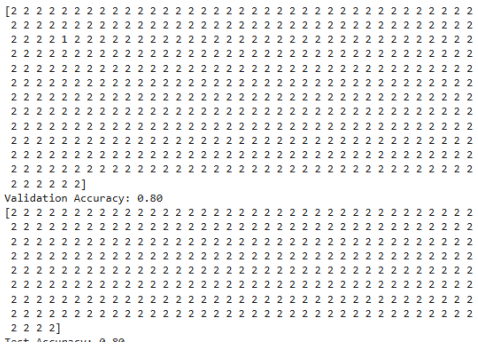


Figure 56: Validation and Accuracy of Training Dataset

8.2.4 Decision Boundary of Training Data superimposed

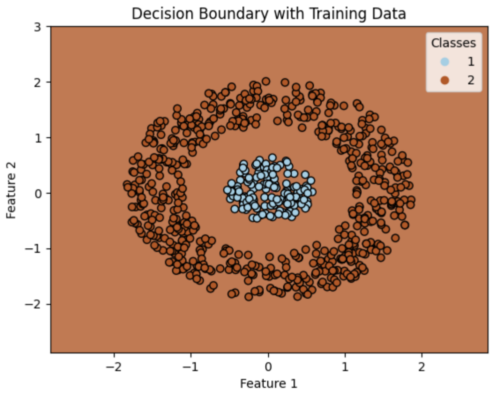


Figure 57: Decision Boundary for Training Dataset

8.2.5 Validation and Accuracy

```
Validation Predictions: [2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2]
Validation Accuracy: 0.80
Test Predictions: [2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2]
Test Accuracy: 0.80
```

Figure 58: Validation and Accuracy of Testing Dataset

8.2.6 Decision Boundary of Testing Data superimposed

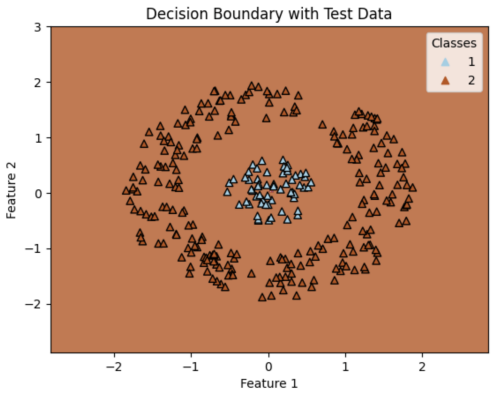


Figure 59: Decision Boundary for Testing Dataset

Conclusion: Issue with Reference Template-based Classifier and Concentric Ring Data When working with a non-linearly separable dataset where the classes form concentric rings (with class 1 forming the innermost ring and class 2 forming the outer ring), I faced an issue where all the test samples were classified into class 2 using a Reference Template-based classifier that utilized both the mean vector and covariance matrix as templates. Despite having over 100 test samples labeled as class 1, they were all misclassified as class 2.

Additionally, I couldn't visualize the decision boundary for this approach, which also contributed to the challenge.

Potential Reasons for the Issue: Dominance of the Covariance Matrix: The covariance matrix plays a key role in shaping the decision boundary of a Reference Template classifier. When both the mean and covariance matrix are used, the classifier models the distribution of the data assuming a Gaussian (normal) distribution for each class. In a concentric ring structure, the covariance matrix may struggle to capture the circular nature of the data. This is because the covariance matrix typically assumes elliptical contours, which does not suit the ring-like configuration of the data. As a result, the classifier may incorrectly classify all samples as class 2, failing to recognize

the ring structure.

Non-linearly Separable Data: The concentric rings configuration, where class 1 is the inner ring and class 2 is the outer ring, poses a challenge for a model that assumes linear separability. The covariance matrix in the Reference Template classifier may fail to model this distribution correctly, leading to all test samples being classified as class 2. The Mahalanobis distance used in the template classifier might not adequately capture the non-linear structure of the data, leading to misclassifications.

Degeneracy or Instability in the Covariance Matrix: If the covariance matrix is nearly singular (due to small data variance or poor conditioning), it might cause instability, further contributing to the misclassification of samples. Regularization of the covariance matrix is typically required in such cases, but it might still struggle to deal with non-linearly separable data like concentric rings.

Using Only the Mean Vector: When I used only the mean vector as the template, the classifier was much simpler and treated the classes as if they were spherical in nature. This worked better for concentric rings since the mean vector naturally defines a decision boundary that might better capture the structure of the data. The decision boundary in this case was more circular, which could explain why a decision plot was successfully generated, even though all test samples were still misclassified as class 2.

Conclusion and Next Steps: The primary reason for all the test samples being classified into class 2 is likely the failure of the covariance matrix to capture the concentric ring structure of the data. The covariance matrix assumes an elliptical spread for each class, which does not suit ring-shaped distributions.

8.3 Overlapping Dataset

8.3.1 Metric Summary

Table 19: Classwise recall, precision and F-measure score

Index	Class1	Class2	Class3	Class4
Precision	0.796460	0.956522	0.882979	0.970297
Recall	0.9000	0.8800	0.8300	0.9800
F1 Score	0.845070	0.916667	0.855670	0.975124

Table 20: Mean recall, precision and F-measure score

Index	Value
Accuracy Score	89.75%
Mean Precision	0.901564
Mean Recall	0.8975
Mean F1	0.898133

8.3.2 Confusion Matrix

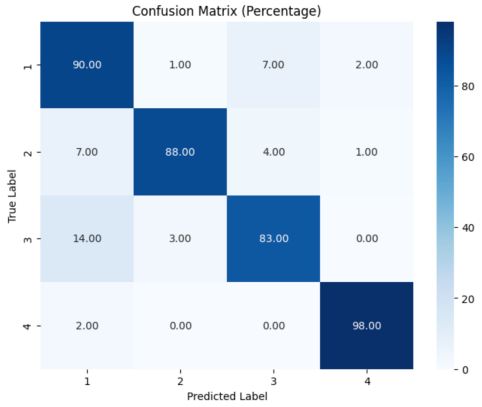


Figure 60: Confusion Matrix

8.3.3 Decision boundary for every pair of class.

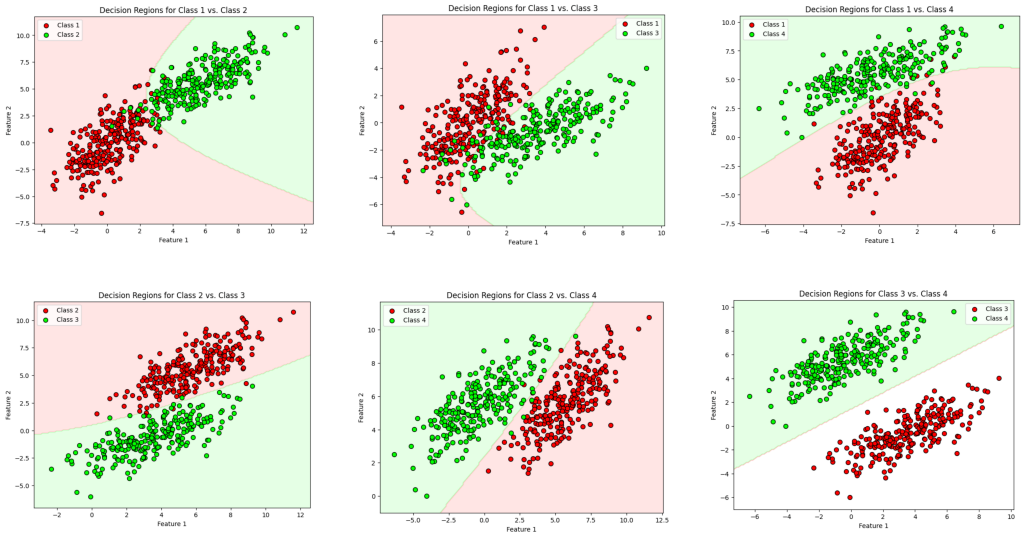


Figure 61: Pairwise Plot

8.3.4 Decision Boundary of Training Data superimposed

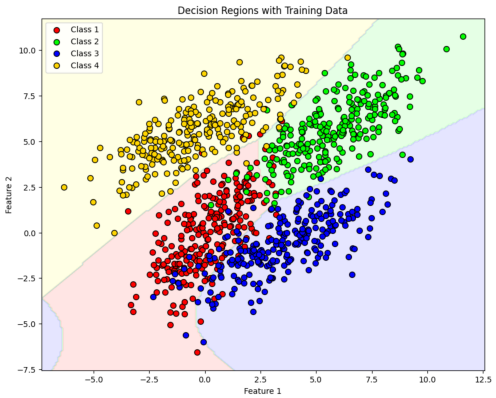


Figure 62: Decision Boundary for Training Dataset

8.3.5 Decision Boundary of Testing Data superimposed

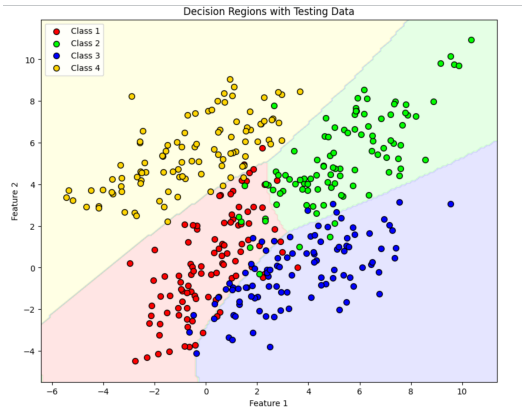


Figure 63: Decision Boundary for Testing Dataset

9 Bayesian Classifier-Unimodal Gaussian Density

The Bayesian Classifier is a fundamental concept in machine learning, based on Bayes' Theorem1, every pair of classified features is independent of each other. In a Unimodal Gaussian Density context, the Bayesian Classifier makes some specific assumptions. It assumes that the likelihood function, p(x—t), follows a Gaussian or normal distribution. The term"Unimodal" refers to the fact that there is only one mode, or peak, in the Gaussian distribution. In other words, it assumes that the data points are distributed around a single value, rather than multiple values.

Case1 : Covariance matrix for all the classes is the same and is 2I Case2 : Full covariance matrix for all the classes and is same for all the classes a. Same covariance matrix for all the classes may be obtained by taking average of covariance matrices of all the classes b. Same covariance matrix for all the classes by computing the covariance matrix of training data of all the classes combined. Case3 : The covariance matrix is diagonal and is different for each class Case4 : The covariance matrix is diagonal and is different for

each class

9.1 Linearly Separable Dataset Case 1

Validation Accuracy: 100.00Test Accuracy: 100.00

9.1.1 Metric Analysis

	Class	Precision	Recall	F-Measure
0	Accuracy	1.0	NaN	NaN
1	1.0	1.0	1.0	1.0
2	2.0	1.0	1.0	1.0
3	3.0	1.0	1.0	1.0
4	4.0	1.0	1.0	1.0
5	Mean	1.0	1.0	1.0

Figure 64: Precision, Recall, F1 Score, Accuracy

9.1.2 Confusion Matrix

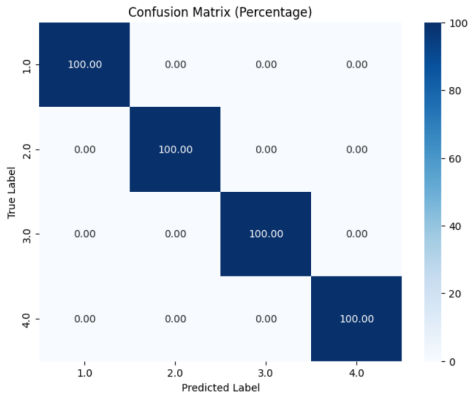


Figure 65: Confusion Matrix

9.1.3 Decision boundary for every pair of class.

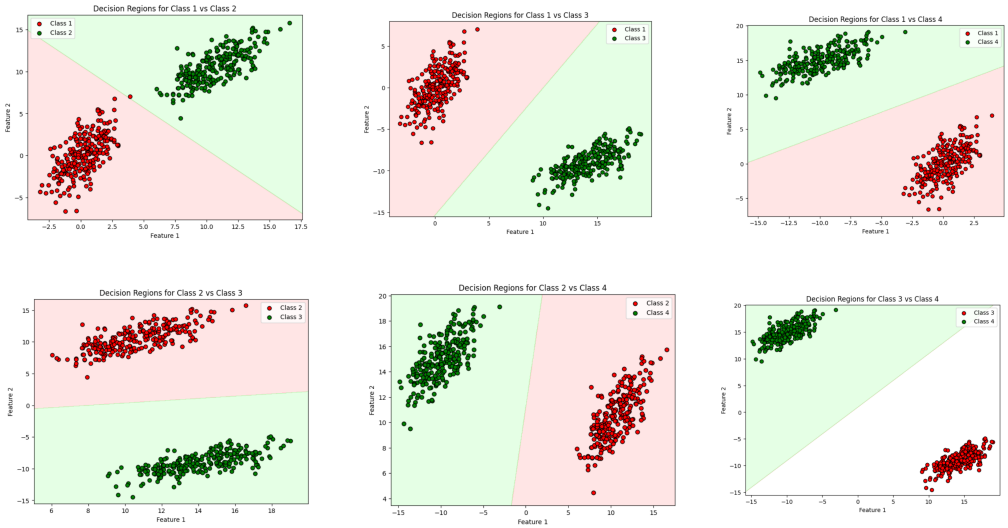


Figure 66: Pairwise Plot

9.1.4 Decision Boundary of Training Data superimposed

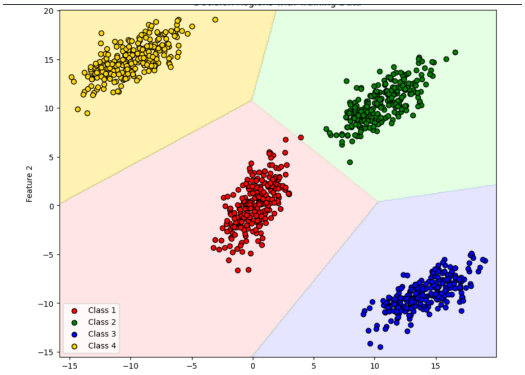


Figure 67: Decision Boundary for Training Dataset

9.1.5 Decision Boundary of Testing Data superimposed

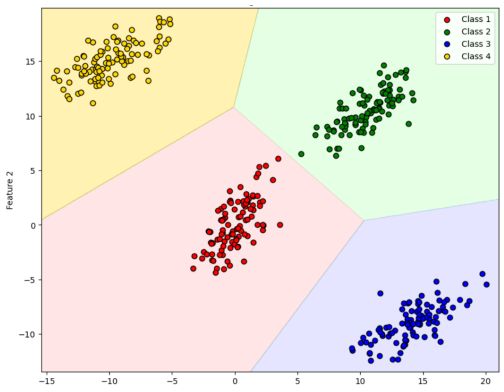


Figure 68: Decision Boundary for Testing Dataset

9.2 Linearly Separable Dataset Case 2

9.2.1 Metric Analysis

Evaluation Results with Averaged Covariance Matrix:

	Class	Precision	Recall	F-measure
0	1	1.0	1.0	1.0
1	2	1.0	1.0	1.0
2	3	1.0	1.0	1.0
3	4	1.0	1.0	1.0
4	Mean	1.0	1.0	1.0

Evaluation Results with Combined Covariance Matrix:

	Class	Precision	Recall	F-measure
0	1	1.0	1.0	1.0
1	2	1.0	1.0	1.0
2	3	1.0	1.0	1.0
3	4	1.0	1.0	1.0
4	Mean	1.0	1.0	1.0

Figure 69: Precision, Recall, F1 Score, Accuracy

9.2.2 Avg Confusion Matrix

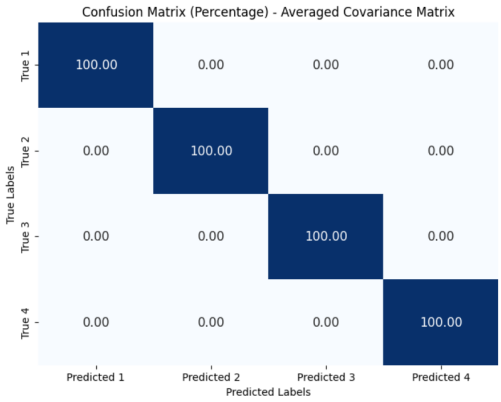


Figure 70: Avg Confusion Matrix

9.2.3 Combined Confusion Matrix

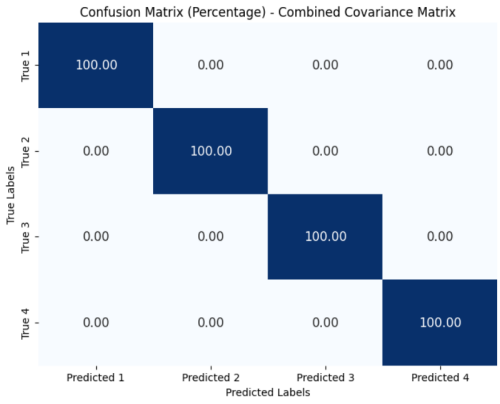


Figure 71: Combined Confusion Matrix

9.2.4 Covariance Matrix Training Dataset

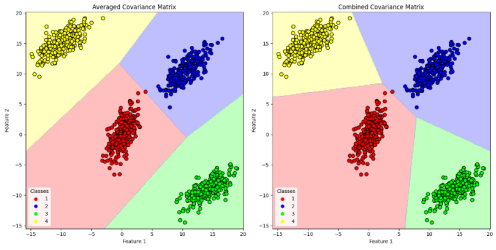


Figure 72: Covariance Matrix Training Dataset

9.2.5 Covariance Matrix Testing Dataset

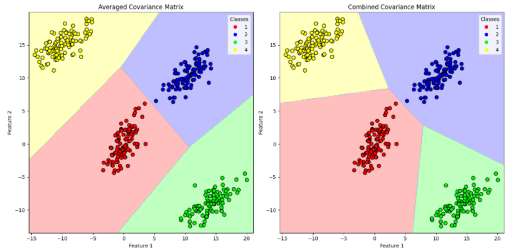
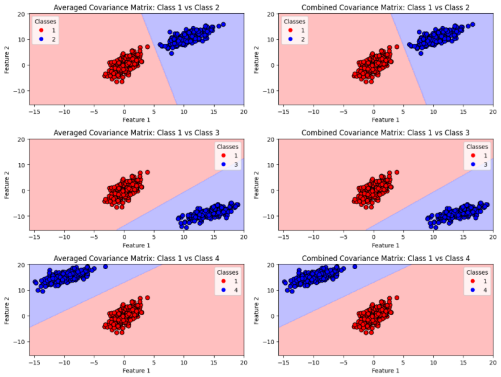


Figure 73: Covariance Matrix Testing Dataset

9.2.6 Decision Boundary Training Dataset



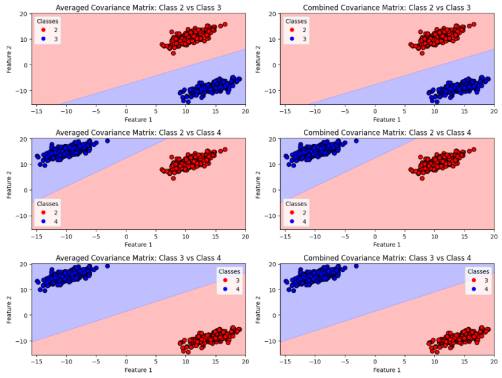


Figure 74: Decision region plot for every pair of classes

9.3 Linearly Separable Dataset Case 3

9.3.1 Metric Analysis

```
Classification Accuracy: 1.0
Class 1 - Precision: 1.0, Recall: 1.0, F1-score: 1.0
Class 2 - Precision: 1.0, Recall: 1.0, F1-score: 1.0
Class 3 - Precision: 1.0, Recall: 1.0, F1-score: 1.0
Class 4 - Precision: 1.0, Recall: 1.0, F1-score: 1.0
Mean Precision: 1.0
Mean Recall: 1.0
Mean F1-score: 1.0
```

Figure 75: Precision, Recall, F1 Score, Accuracy

9.3.2 Confusion Matrix

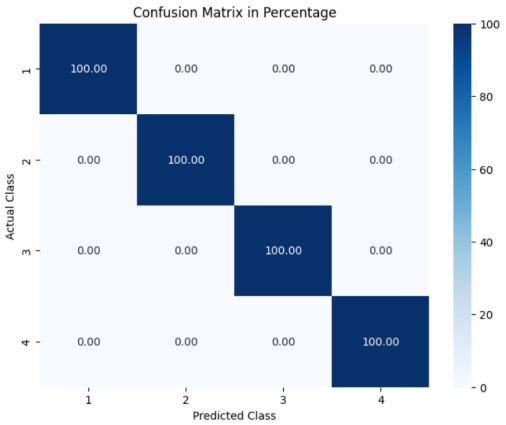


Figure 76: Confusion Matrix

9.3.3 Decision boundary for every pair of class.

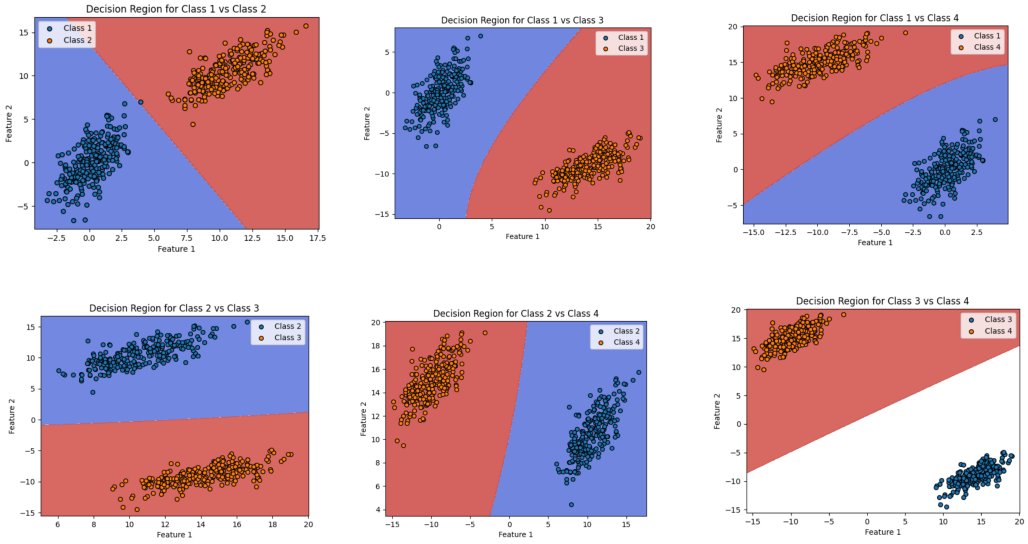


Figure 77: Pairwise Plot

9.3.4 Decision Boundary of Training Data superimposed

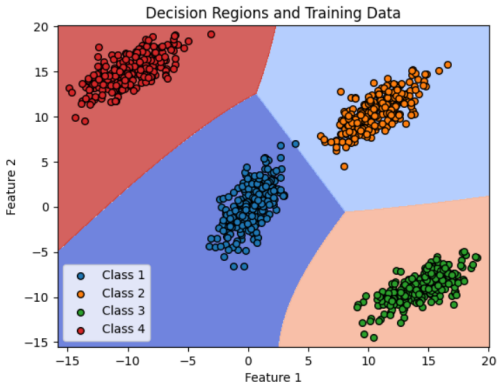


Figure 78: Decision Boundary for Training Dataset

9.3.5 Decision Boundary of Testing Data superimposed

Figure 79: Decision Boundary for Testing Dataset

9.4 Linearly Separable Dataset Case 4

9.4.1 Metric Analysis

Classification Metrics for Each Class:

	Class	Precision	Recall	F-Measure
0	1	1.0	1.0	1.0
1	2	1.0	1.0	1.0
2	3	1.0	1.0	1.0
3	4	1.0	1.0	1.0
4	Mean	1.0	1.0	1.0

Accuracy: 1.0000

Figure 80: Precision, Recall, F1 Score, Accuracy

9.4.2 Confusion Matrix

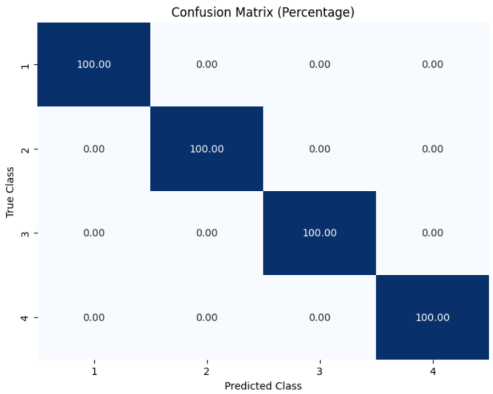


Figure 81: Confusion Matrix

9.4.3 Decision boundary for every pair of class.

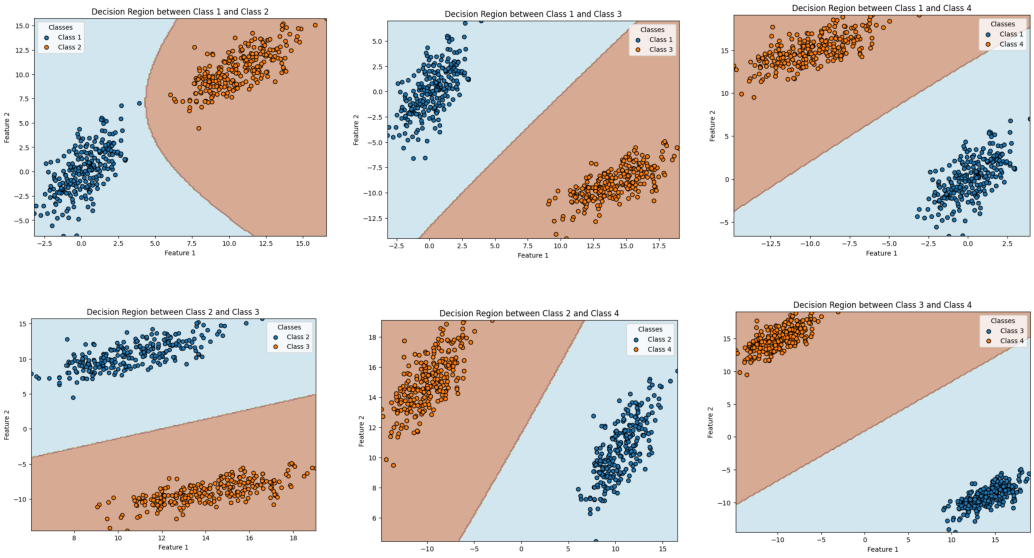


Figure 82: Pairwise Plot

9.4.4 Decision Boundary of Training Data superimposed

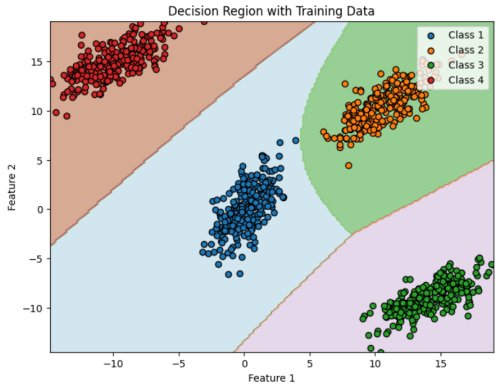


Figure 83: Decision Boundary for Training Dataset

9.4.5 Decision Boundary of Testing Data superimposed

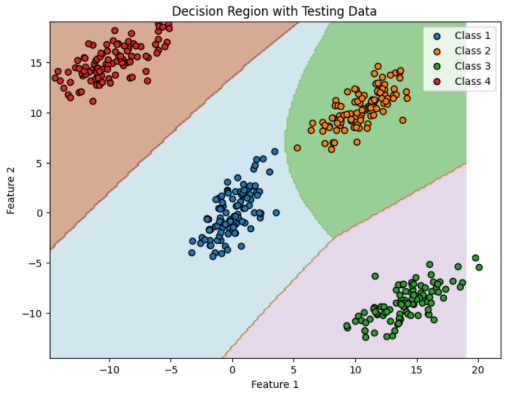


Figure 84: Decision Boundary for Testing Dataset

9.5 Non-Linearly Separable Dataset Case 1

Validation Accuracy: 50.67Test Accuracy: 52.67

9.5.1 Metric Analysis

	Class	Precision	Recall	F-Measure
0	Accuracy	0.526667	NaN	NaN
1	1.0	0.215278	0.516667	0.303922
2	2.0	0.814103	0.529167	0.641414
3	Mean	0.514690	0.522917	0.472668

Figure 85: Precision, Recall, F1 Score, Accuracy

9.5.2 Confusion Matrix

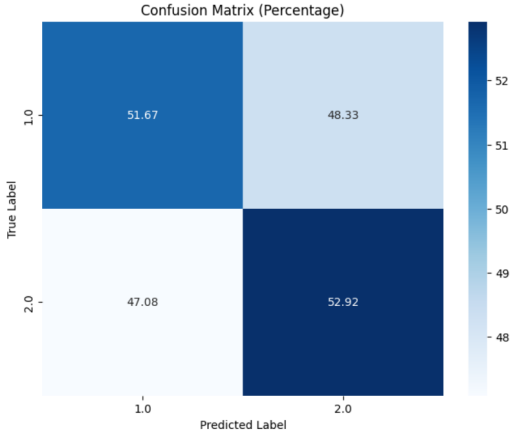


Figure 86: Confusion Matrix

9.5.3 Decision Boundary

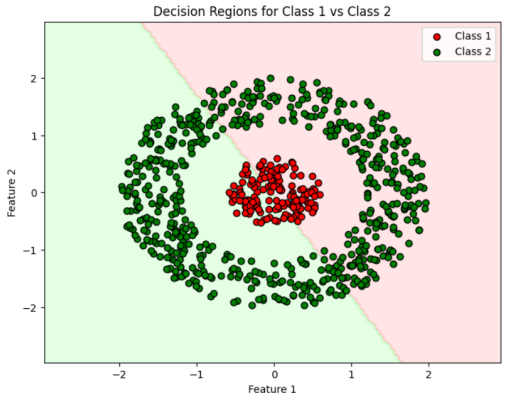


Figure 87: Decision Boundary

9.5.4 Decision Boundary of Training Dataset

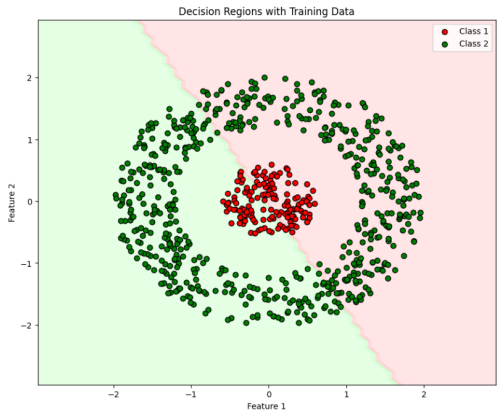


Figure 88: Decision Boundary of Training Dataset

9.5.5 Decision Boundary of Testing Dataset

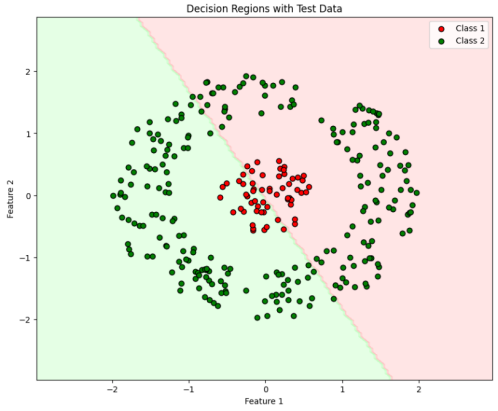


Figure 89: Decision Boundary of Testing Dataset

9.6 Non linearly Separable Dataset Case 2

Full covariance matrix for all the classes and is same for all the classes 1. Same covariance matrix for all the classes may be obtained by taking average of covariance matrices of all the classes 2. Same covariance matrix for all the classes by computing the covariance matrix of training data of all the classes combined.

1st approach - Covariance matrix calculated by taking average of covariance matrices of all classes

Class 1 has 150 samples. Class 2 has 600 samples. Test Accuracy: 51.67Validation Accuracy: 50.22

9.6.1 Metric Analysis

```
Classification Accuracy: 0.5167
Class 1:
  Precision: 0.2109
  Recall: 0.5167
  F1-Score: 0.2995
Class 2:
  Precision: 0.8105
  Recall: 0.5167
  F1-Score: 0.6310
Mean Precision: 0.5107
Mean Recall: 0.5167
Mean F1-Score: 0.4653
```

Figure 90: Precision, Recall, F1 Score, Accuracy

9.6.2 Confusion Matrix

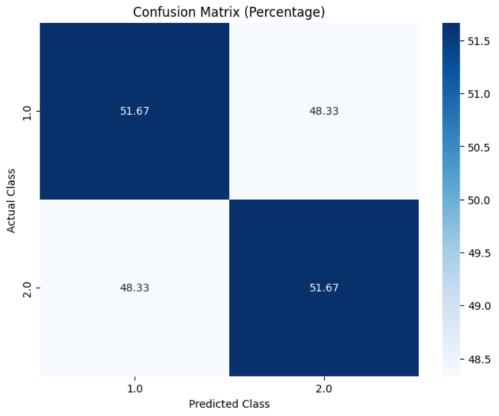


Figure 91: Confusion Matrix

9.6.3 Decision Region of Training Dataset

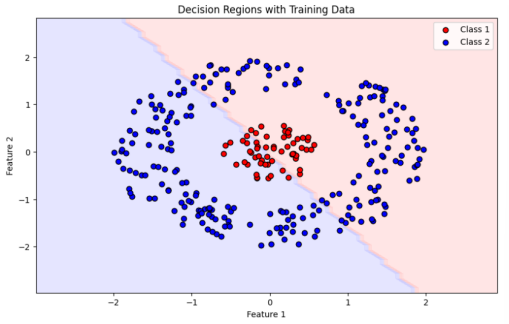


Figure 92: Decision Region of Training Dataset

9.6.4 Decision Region of Testing Dataset

Figure 93: Decision Region of Testing Dataset

2nd Approach covariance matrix for all the classes by computing the covariance matrix of training data of all the classes combined.

Training Accuracy: 52.93Test Accuracy: 51.67

9.6.5 Metric Analysis

```
Classification Report on Test Data
              precision    recall  f1-score     support
1.0            0.210884  0.516667  0.299517   60.000000
2.0            0.810458  0.516667  0.631043  240.000000
accuracy       0.516667  0.516667  0.516667    0.516667
macro avg      0.510671  0.516667  0.465280  300.000000
weighted avg   0.690543  0.516667  0.564738  300.000000

Summary Metrics
           Metric     Score
0        Accuracy  0.516667
1  Mean Precision  0.510671
2     Mean Recall  0.516667
3   Mean F1-Score  0.465280
```

Figure 94: Precision, Recall, F1 Score, Accuracy

9.6.6 Confusion Matrix

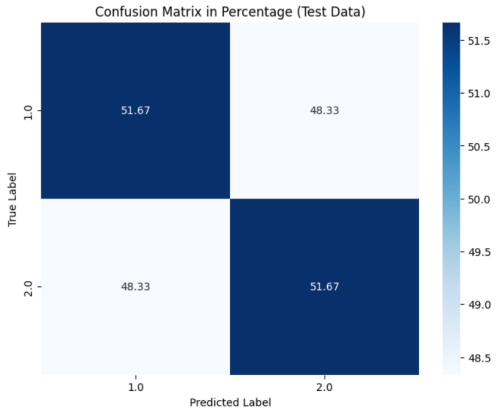


Figure 95: Confusion Matrix

9.6.7 Decision Region of Training Dataset

Figure 96: Decision Region of Training Dataset

9.6.8 Decision Region of Testing Dataset

Figure 97: Decision Region of Testing Dataset

9.7 Non-Linearly Separable Dataset Case 3

9.7.1 Metric Analysis

```
Classification Accuracy: 94.33%
Class 1 - Precision: 1.00, Recall: 0.72, F-measure: 0.83
Class 2 - Precision: 0.93, Recall: 1.00, F-measure: 0.97
Mean Precision: 0.97
Mean Recall: 0.86
Mean F-measure: 0.90
```

Figure 98: Precision, Recall, F1 Score, Accuracy

9.7.2 Confusion Matrix

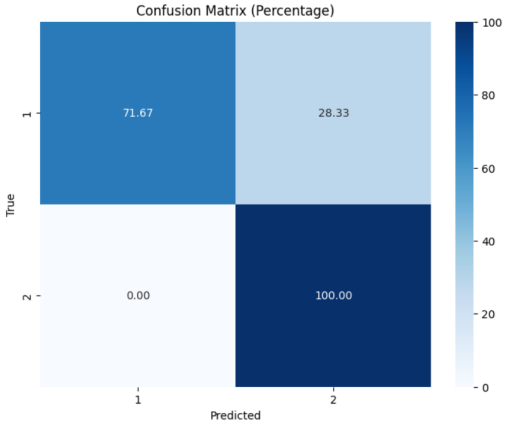


Figure 99: Confusion Matrix

9.7.3 Decision Region of Training Dataset

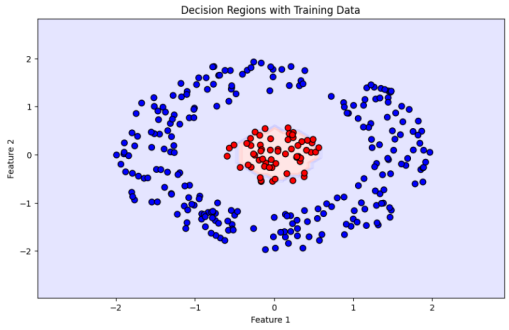


Figure 100: Decision Region of Training Dataset

9.7.4 Decision Region of Testing Dataset

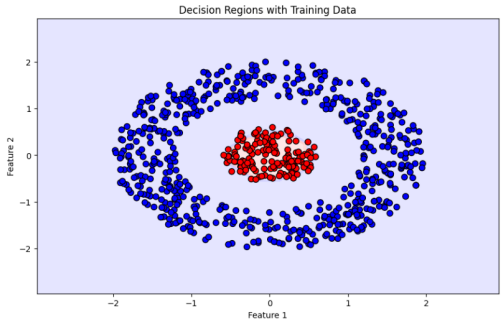


Figure 101: Decision Region of Testing Dataset

9.8 Non-Linearly Separable Dataset Case 4

9.8.1 Metric Analysis

```
Classification Accuracy: 0.9266666666666666
Class 1:
  Precision: 1.0
  Recall: 0.6333333333333333
  F-measure: 0.7755102040816326
Class 2:
  Precision: 0.916030534351145
  Recall: 1.0
  F-measure: 0.9561752988047809

Mean Precision: 0.9580152671755725
Mean Recall: 0.8166666666666667
Mean F-measure: 0.8658427514432068
```

Figure 102: Precision, Recall, F1 Score, Accuracy

9.8.2 Confusion Matrix

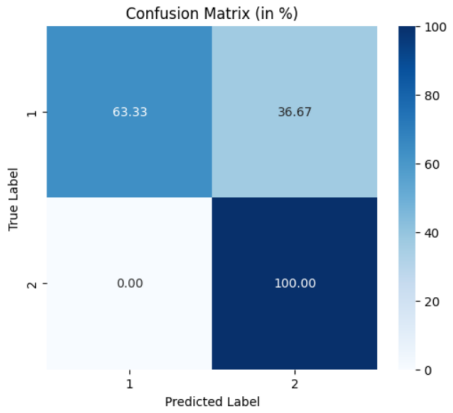


Figure 103: Confusion Matrix

9.8.3 Decision Region of Training Dataset

Figure 104: Decision Region of Training Dataset

9.8.4 Decision Region of Testing Dataset

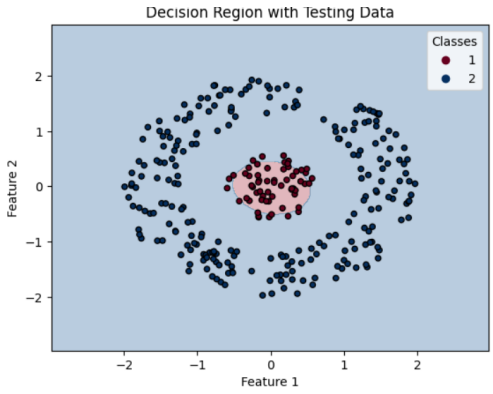


Figure 105: Decision Region of Testing Dataset

9.9 Overlapping Dataset Case 1

9.9.1 Metric Analysis

	Class	Precision	Recall	F-Measure
0	Accuracy	0.825000	NaN	NaN
1	1.0	0.738739	0.820	0.777251
2	2.0	0.858586	0.850	0.854271
3	3.0	0.853659	0.700	0.769231
4	4.0	0.861111	0.930	0.894231
5	Mean	0.828024	0.825	0.823746

Figure 106: Precision, Recall, F1 Score, Accuracy

9.9.2 Confusion Matrix

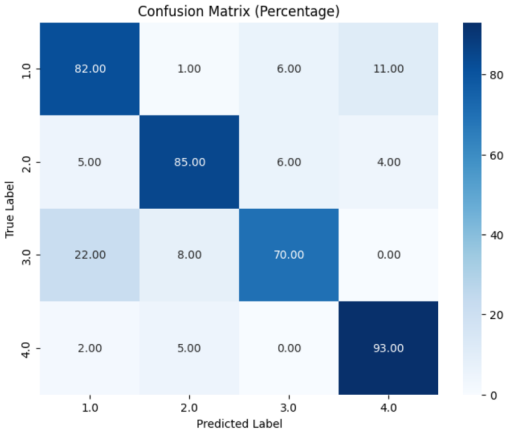


Figure 107: Confusion Matrix

9.9.3 Decision boundary for every pair of class.

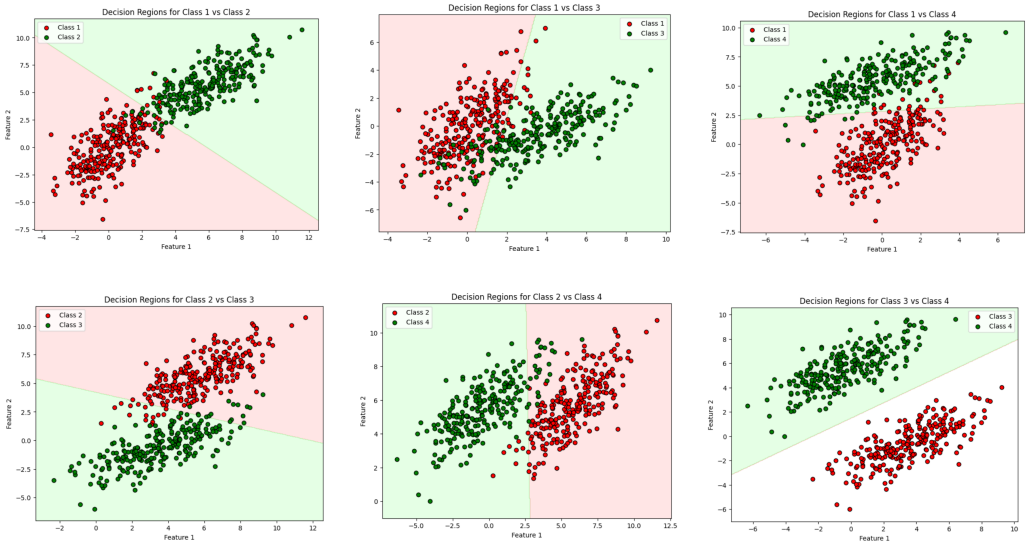


Figure 108: Pairwise Plot

9.9.4 Decision Boundary of Training Data superimposed

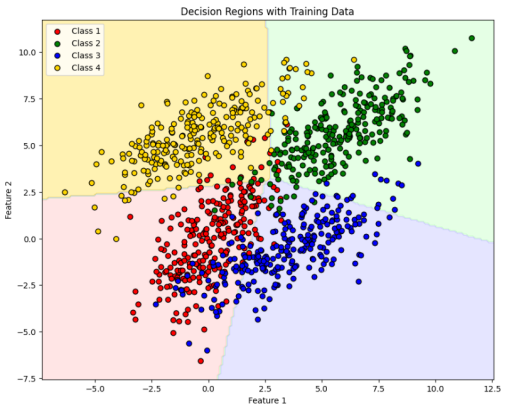


Figure 109: Decision Boundary for Training Dataset

9.9.5 Decision Boundary of Testing Data superimposed

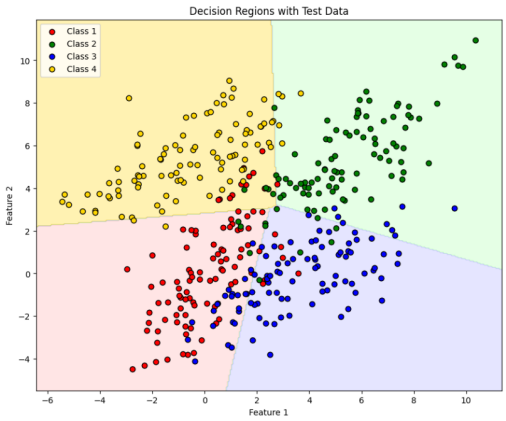


Figure 110: Decision Boundary for Testing Dataset

9.10 Overlapping Dataset Case 2

Full covariance matrix for all the classes and is same for all the classes 1. The same covariance matrix for all the classes may be obtained by taking the average of covariance matrices of all the classes 1. The same covariance matrix for all the classes by computing the covariance matrix of training data of all the classes combined.

9.10.1 Metric Analysis

Evaluation Results with Averaged Covariance Matrix:

	Class	Precision	Recall	F-measure
0	1	0.790000	0.7900	0.790000
1	2	0.957447	0.9000	0.927835
2	3	0.854167	0.8200	0.836735
3	4	0.909091	1.0000	0.952381
4	Mean	0.877676	0.8775	0.876738

Evaluation Results with Combined Covariance Matrix:

	Class	Precision	Recall	F-measure
0	1	0.790476	0.83	0.809756
1	2	0.907216	0.88	0.893401
2	3	0.897727	0.79	0.840426
3	4	0.890909	0.98	0.933333
4	Mean	0.871582	0.87	0.869229

Figure 111: Precision, Recall, F1 Score, Accuracy

9.10.2 Avg Confusion Matrix

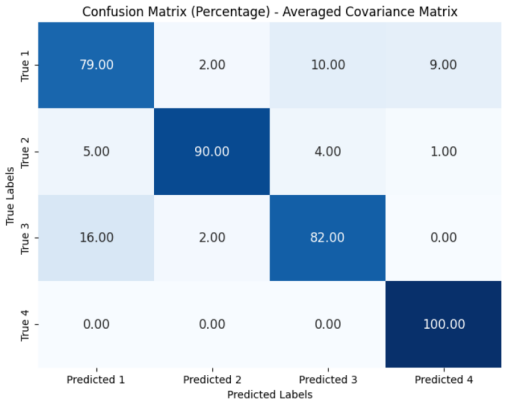


Figure 112: Avg Confusion Matrix

9.10.3 Combined Confusion Matrix

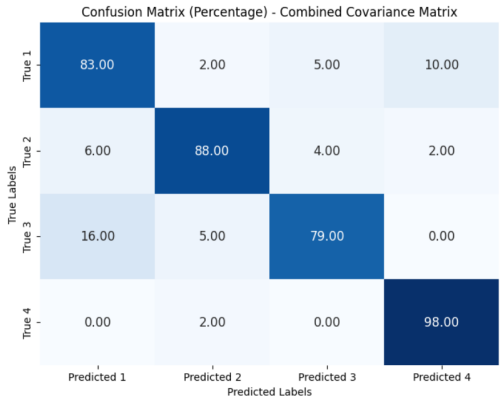
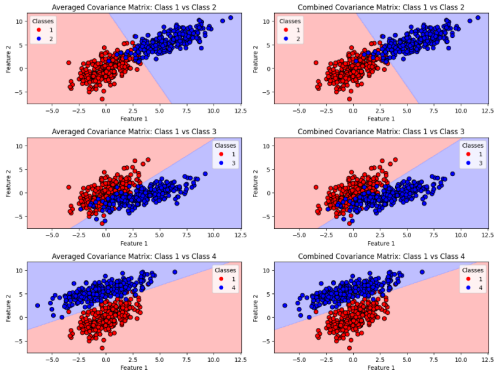


Figure 113: Combined Confusion Matrix

9.10.4 Decision region plot for every pair of classes



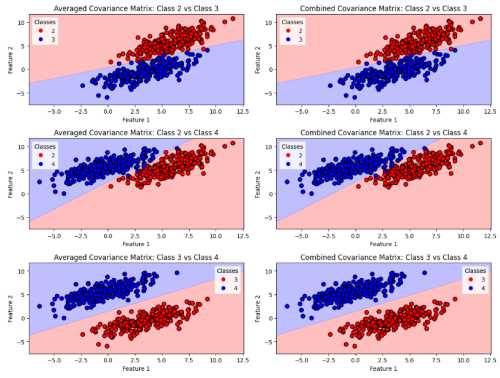


Figure 114: Pair-wise Plots

9.10.5 Decision region plot for all the classes together with the training data superposed

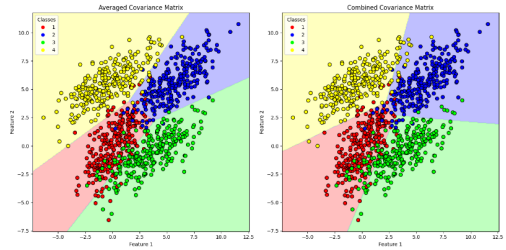


Figure 115: Decision Region Plot for Training Data

9.10.6 Decision region plot for all the classes together with the testing data superimposed

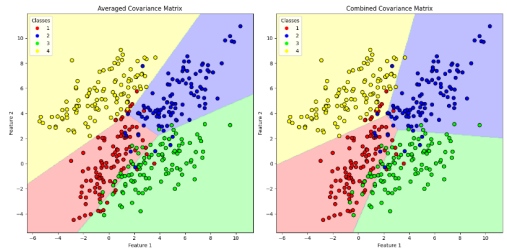


Figure 116: Decision Region Plot for Testing Data

9.11 Overlapping Dataset Case 3

9.11.1 Metric Analysis

```
Classification Accuracy: 0.825
Class 1 - Precision: 0.7619047619047619, Recall: 0.8, F1-score: 0.7804878048780488
Class 2 - Precision: 0.8645833333333334, Recall: 0.83, F1-score: 0.8469387755102041
Class 3 - Precision: 0.8588235294117647, Recall: 0.73, F1-score: 0.7891891891891892
Class 4 - Precision: 0.8245614035087719, Recall: 0.94, F1-score: 0.8785046728971962
Mean Precision: 0.827468257039658
Mean Recall: 0.825
Mean F1-score: 0.8237801106186595
```

Figure 117: Precision, Recall, F1 Score, Accuracy

9.11.2 Confusion Matrix

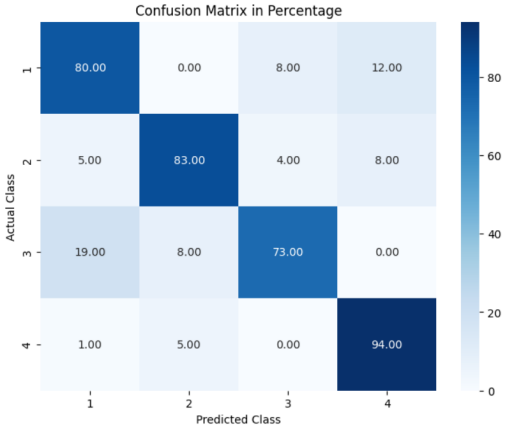


Figure 118: Confusion Matrix

9.11.3 Decision boundary for every pair of class.

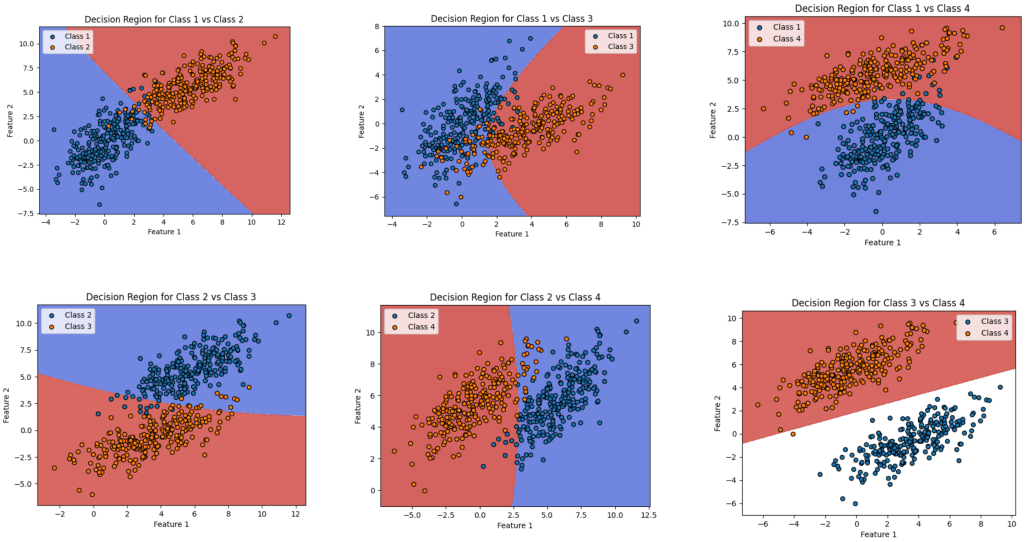


Figure 119: Pairwise Plot

9.11.4 Decision Boundary of Training Data superimposed

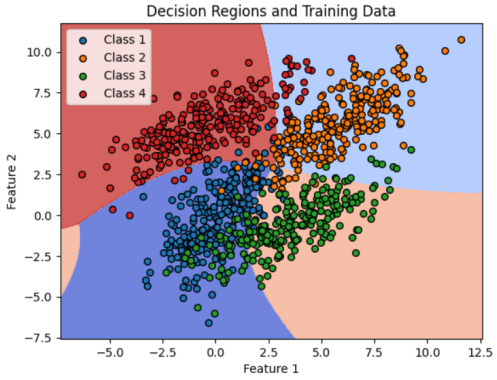


Figure 120: Decision Boundary for Training Dataset

9.11.5 Decision Boundary of Testing Data superimposed

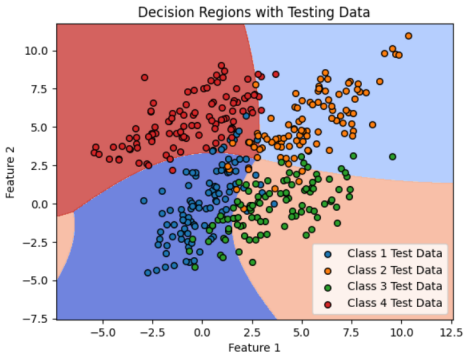


Figure 121: Decision Boundary for Testing Dataset

9.12 Overlapping Dataset Case 4

9.12.1 Metric Analysis

Classification Metrics for Each Class:

	Class	Precision	Recall	F-Measure
0	1	0.796460	0.9000	0.845070
1	2	0.956522	0.8800	0.916667
2	3	0.882979	0.8300	0.855670
3	4	0.970297	0.9800	0.975124
4	Mean	0.901564	0.8975	0.898133

Accuracy: 0.8975

Figure 122: Precision, Recall, F1 Score, Accuracy

9.12.2 Confusion Matrix

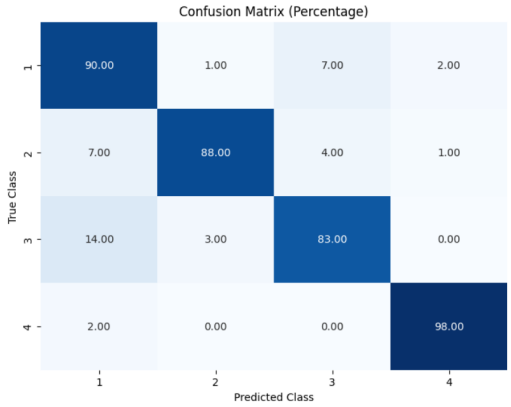


Figure 123: Confusion Matrix

9.12.3 Decision boundary for every pair of class.

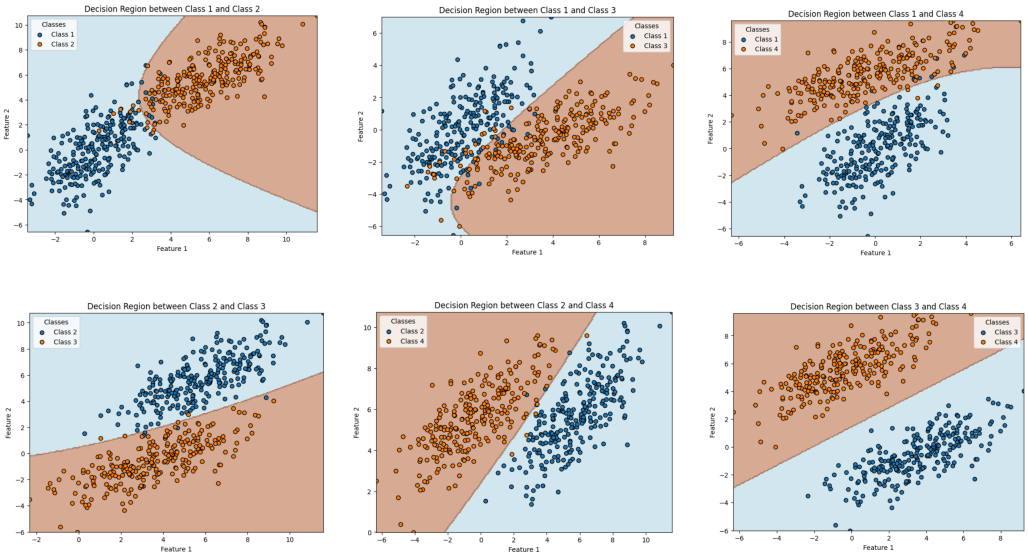


Figure 124: Pairwise Plot

9.12.4 Decision Boundary of Training Data superimposed

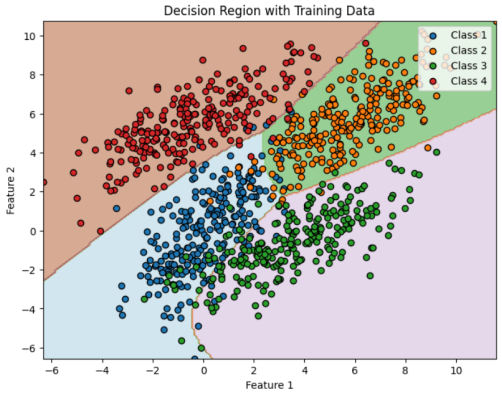


Figure 125: Decision Boundary for Training Dataset

9.12.5 Decision Boundary of Testing Data superimposed

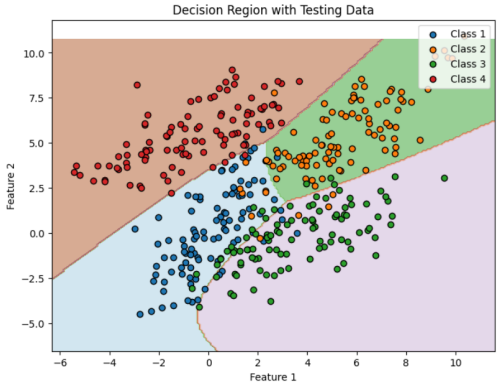


Figure 126: Decision Boundary for Testing Dataset

10 Observation

- **NN classifier:** It had 100% accuracy on linearly and non-linearly separable dataset. Although it had 86.25% accuracy on overlapping dataset.

- **KNN classifier:** The most optimal value of K from elbow method for linearly and non-linearly separable data turned out to be 1, which indicates a simplified nn classifier. Although for an overlapping dataset, optimal value for K turns out to be 9. It gave an accuracy of 88% on overlapping dataset.

- **Reference Template Classifier:**
  (a) Mean vector as reference template: This classifier worked poorly on non linear and overlapping dataset. It handled the linearly separable dataset with 100% accuracy but had a dropped accuracy of 52% and 82.25% accuracy on non-linearly separable dataset and overlapping dataset respectively.
  (b) Mean vector and covariance matrix as reference template: Taking covariance matrix also as reference template has improved the performance of reference template classifier on non linearly separable data. Though the classification accuracy is 80%, this method failed to provide the decision plot due to its extreme non-linearly separable nature, However, the classifier had an improved performance on the overlapping dataset with a classification accuracy of 89.75% when compared to the previous case.

- **Bayesian Classifier:** Bayesian Classifier had a slight improvement with accuracy reaching 90% on overlapping dataset. The performance of this classifier did have minor difference on changing the way covariance matrices were assumed. On contrary to this, there was a drop in accuracy for the non-linearly separable dataset enforcing the point that Bayes classifier is not the best classifier for the given non-linearly separable data and hence allowing explore other methods of classification.

11 Conclusion

All the classifiers performed with 100% accuracy on linearly separable dataset.

• On using KNN classifier for overlapping dataset we obtained a improved accuracy from 86.25% to 88% for K = 9.

• Accuracy of reference template classifier improves on overlapping data when covariance matrix is also used as reference template along with mean vector but failed to provide a decision boundary for non-linearly separable data despite achieving an accuracy of 80%.

• Bayesian Classifier performed with highest accuracy on overlapping datasets. However, this classifier did not work the best for the non-linearly separable dataset provided due to its extreme non-linearity. This concluded that, for the given non-linearly separable dataset, methods like SVM, Kernel methods and others works best and will be experimented in the future.