KNN Classification Analysis on Synthetic Data

Dataset Overview

In this assignment a synthetic dataset was created using scikit-learn’s make_blobs function. This tool is handy for generating simple datasets to test out clustering or classification techniques.

Cluster centers: [2, 4], [6, 6], [1, 9]

Number of classes: 3 (one for each cluster center)

Total samples: 150, spread evenly across the three groups

Data Preparation

The data was split into:

Training set: 120 samples (80%)

Testing set: 30 samples (20%)

This split allowed the model to learn patterns from most of the data while leaving a smaller portion aside to check how well the model performs on unseen points.

Methodology

A K-Nearest Neighbors (KNN) classifier was applied. The model used the default settings: k=5 neighbors and Euclidean distance to measure similarity.

The process followed three steps:

Train the model on the training data.

Generate predictions for both the training and testing sets.

Evaluate performance using accuracy scores and a confusion matrix.

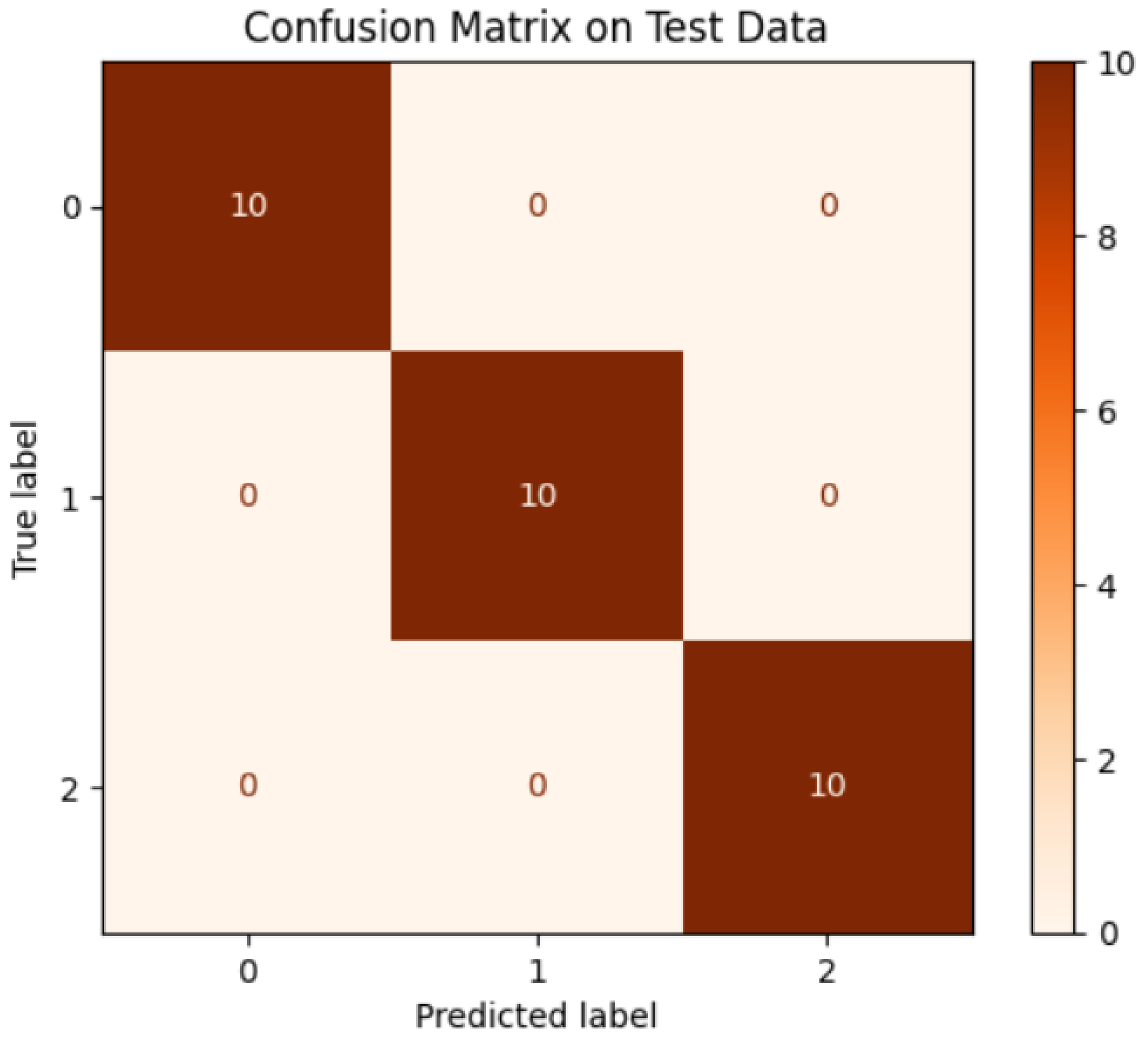
Results:

Training accuracy: 100%

Testing accuracy: 100%

Visualizations:

Confusion Matrix: The confusion matrix compares the predicted labels with the actual labels from the test set. The matrix displays that every sample was correctly classified, creating a perfect diagonal. This confirms the KNN model's 100% accuracy.



Breakdown by class from the test set:

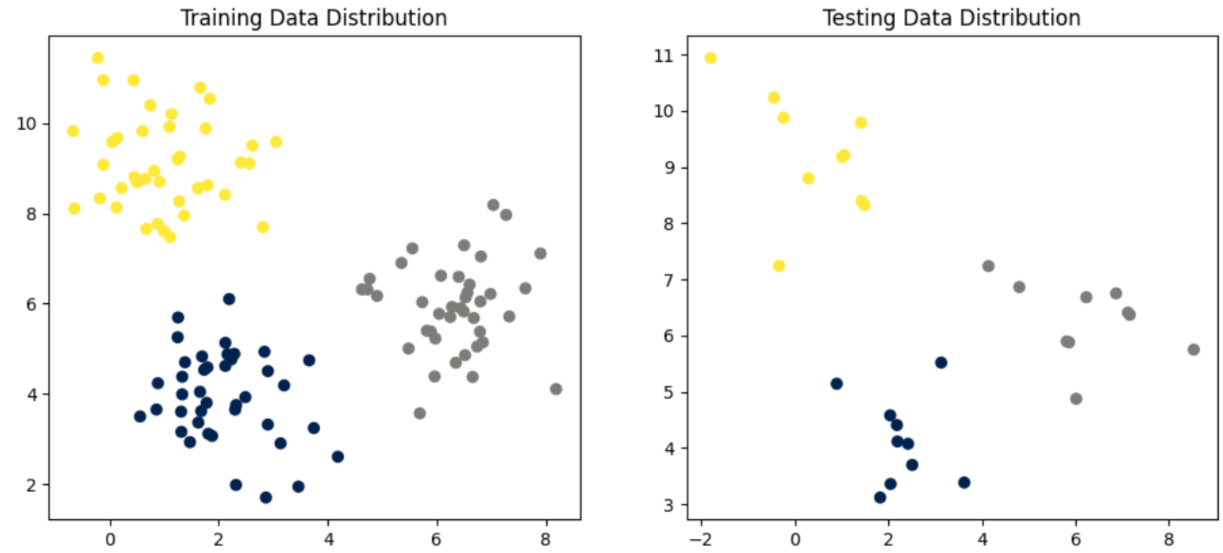
Class 0: All 10 test points were properly marked as Class 0.

Class 1: The model correctly placed all 10 test points in Class 1.

Class 2: All 10 test points were placed in Class 2 without an error.

The confusion matrix shows that all test samples were classified correctly for each class, with no mistakes across the three groups.

Scatter Plots: The scatter plot code displays the training and testing data next to each other. It positions each point based on its features and colors them by class. The plots clearly show that the clusters are well separated. This separation explains why the KNN model achieved perfect accuracy.



Conclusion

On this dataset, the KNN model performed perfectly. Given how distinct and simple it is to distinguish the clusters, it is not surprising that both the training and testing accuracy approached 100%. However, this result is not necessarily like the complicated real-world data because the sample was relatively simple and intentionally generated.