Assignment-1

Venkata sai reddy Peddireddy

Git Hub Link: https://github.com/venkatasai8120/AML_Assignments.git

KNN Classification Analysis on Synthetic Data

Dataset Overview:

For this assignment, the synthetic dataset was generated using the make_blobs function inside the scikit-learn library. This tool is good for generating simple datasets for testing any clustering or classification procedures.

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

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

Number of samples: 150, evenly distributed into three groups

Data Preparation:

The data was split into:

Training set: 120 samples (80%)

Test set: 30 samples (20%)

This data split allowed the model to learn pattern recognition using most of the data while reserving some of it to benchmark the model on unseen points.

Methodology:

Applied A K-Nearest Neighbor (KNN) classifier. The model was run by default: k=5 neighbors, Euclidean distance for distance measure.

There were three sequential operations:

Train the model using the training data.

Predict both training and testing sets.

Evaluate using accuracy score and confusion matrix.

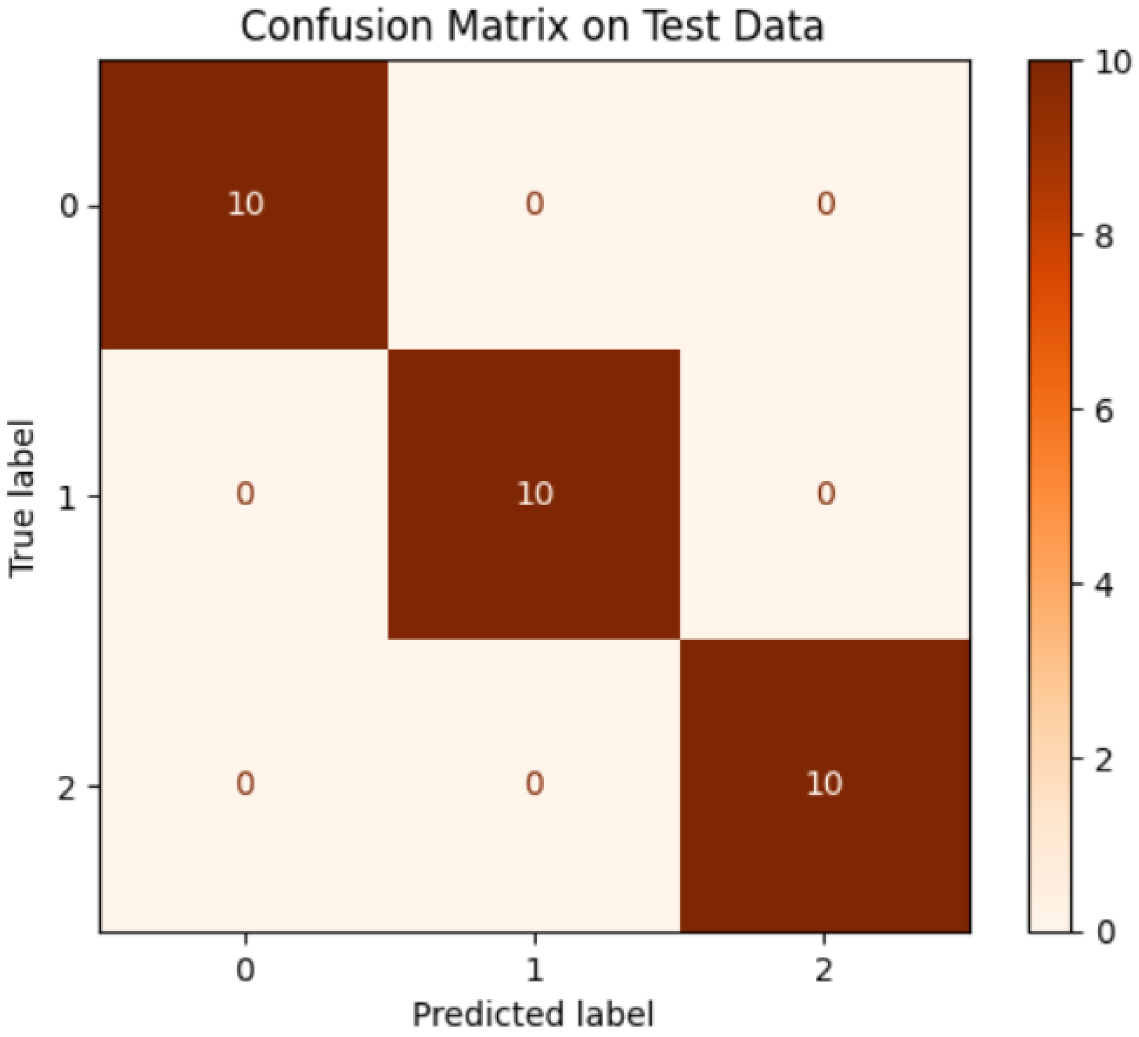
Results:

Training accuracy: 100%

Testing accuracy: 100%

Visualizations:

Confusion Matrix: It compares how the predicted labels match with the actual labels coming from the test set. The matrix shows that each and every sample was correctly identified, presenting a perfect diagonal. This certifies a 100% accuracy of the KNN model.



Breakdown by class from the test set:

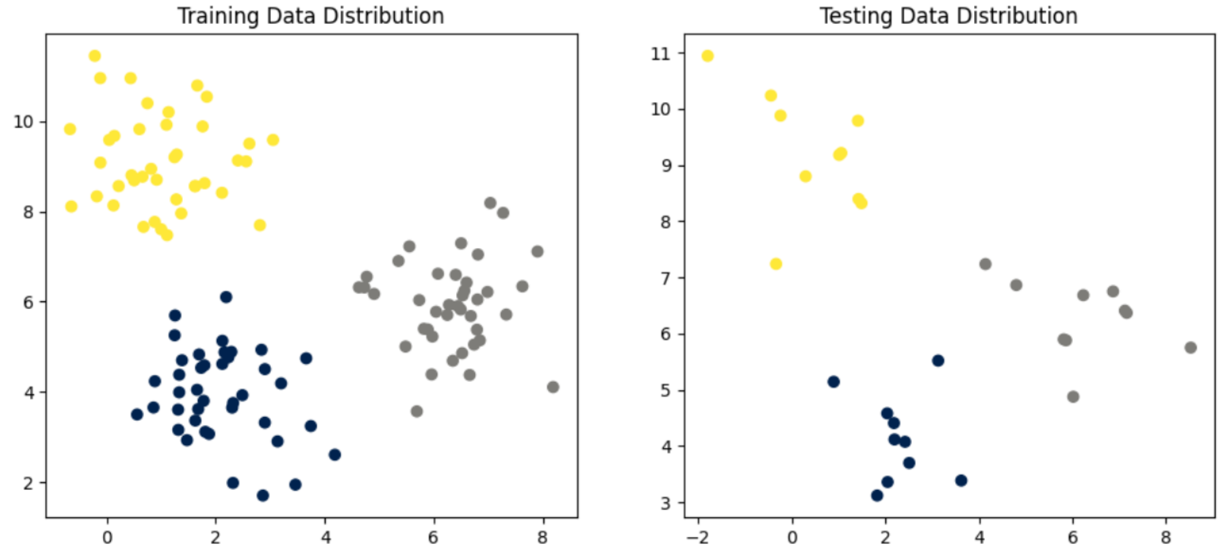
Class 0: All 10 test points were properly marked as Class 0.

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 intentionally generated.