The assignment uses K-Nearest Neighbours (KNN) classification approach to generate the produced dataset and validate its effectiveness. Analyzing the results allows us to assess the classifier's ability to effectively recognize and categorize data points based on feature similarity.

Dataset description

The dataset was built using scikit-learn's make_blobs method. This function is frequently used to generate synthetic data for clustering and classification purposes.

centers = [[2, 4], [6, 6], [1, 9]]

The dataset consists of three separate classes, each representing a unique cluster of points in a two-dimensional space:

Class 0 (Cluster Center: 2 and 4):
Represents the first set of data points concentrated around the center at coordinates (2, 4).

Class 1 (Cluster Center: 6 and 6):
Represents the second group, with points grouped around the center at (6, 6).

Class 2 (Cluster Center: 1 and 9):
Represents the third group, located around coordinates (1, 9).

A balanced dataset was created by generating 150 samples evenly distributed across three classes.

Data Split:

The training set includes 80% of the data (120 samples).
The testing set contains 20% of the data (30 samples).

Graph Plots:

Training Dataset Plot:

Displays three unique clusters, each representing one of the classes.

Testing Dataset Plot:

This shows a comparable distribution, indicating that the testing data has the same structure as the training data.

Methodology:

Data Splitting:

An 80-20 split was used to divide the dataset into training and testing sets. This ensures that the model is trained on the bulk of the data while leaving a fraction for evaluation.

Model Training:

- A (KNN) classifier method has been used.
- Distance metric: Euclidean distance.

Predictions are produced for both training and testing datasets to assess model performance.

Evaluation metrics:
Accuracy Score: Calculated for both training and testing sets to determine the percentage of correctly categorized cases.
Confusion Matrix: Used to evaluate the distribution of correct and wrong classifications.

Results:
Training accuracy: 100 percent.
Testing accuracy: 100 percent.
Confusion Matrix Analysis: The model correctly categorized all test samples, with no misclassifications.
Each of the three classes (0, 1, and 2) was correctly predicted, indicating the model's ability to distinguish between the created clusters.

Class 0 - 14 data instances are correctly classified

Class 1 - 8 data instances are correctly classified

Class 2 - 8 data instances are correctly classified

Conclusion:

- The KNN model achieved 100% classification accuracy on both training and testing datasets.
- The clear separation of classes, as demonstrated by the visualization and confusion matrix, supports the model's efficacy on this dataset.
- As the result shows 100% accuracy, it is important to observe that the dataset is artificially generated and it may not accurately reflect the intricacies of real-world data. Further testing on different datasets is recommended to evaluate the model's resilience.