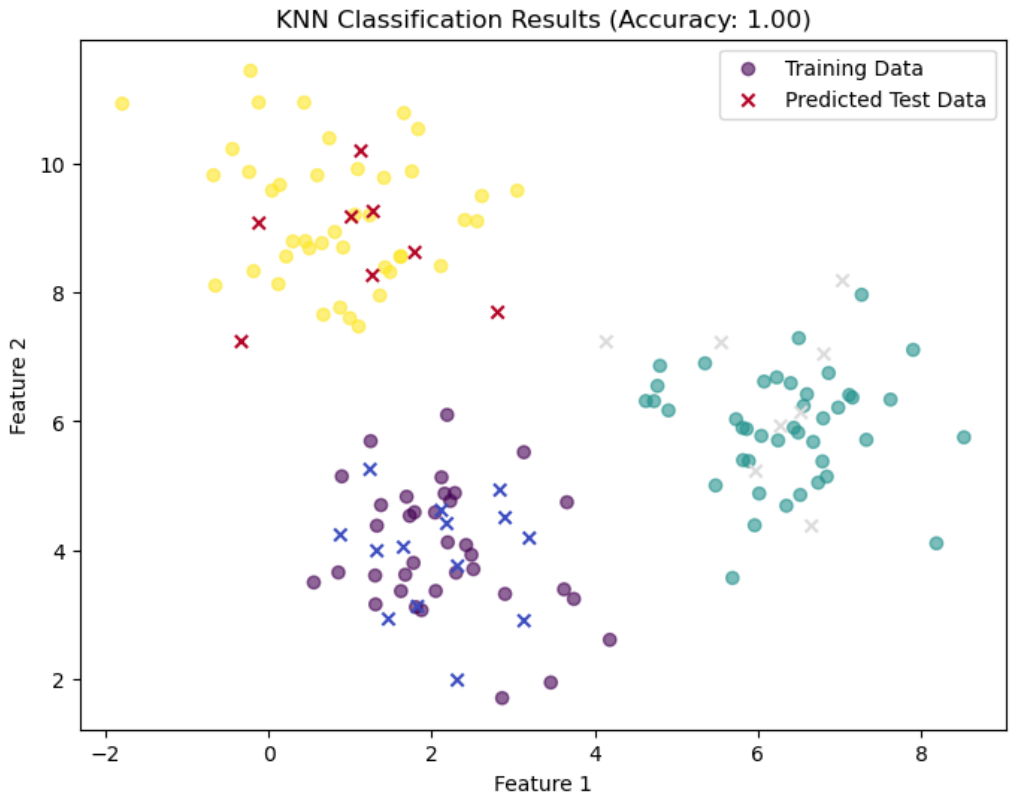
sli61@kent.edu

Assignment 1 - Summary of KNN Analysis



Method

This analysis applies the K-Nearest Neighbors (KNN) classification method to both a real dataset (Iris dataset) and a simulated dataset (generated using make_blobs).

For the Iris dataset, the data was split into 80% training and 20% testing, and a default KNN classifier was trained. The classifier's performance was evaluated using accuracy scores, which demonstrated high accuracy (0.97) in predicting species.

For the simulated dataset, synthetic data with three distinct clusters was generated. The dataset was split similarly into training and testing sets. A KNN classifier with five neighbors (k=5) and Euclidean distance metric was trained. The classifier achieved 100% accuracy on the test set, indicating a perfect separation of the clusters. A scatter plot was created to visually compare the training data and predicted test labels.

Results and analysis

Apparently, the classification on the synthetic dataset was better performed than on the real dataset. The reason behind the classification is that:

- Iris Dataset: The real-world Iris dataset contains three types of flowers, but their features (petal & sepal measurements) overlap to some extent. This makes it harder for KNN to classify every instance correctly.
- Synthetic Dataset: The data was generated with make_blobs, which creates well-separated clusters. This means each class is clearly distinguishable, making classification much easier for KNN.

Overall, this analysis demonstrates the effectiveness of KNN in classifying structured datasets, especially when clusters are well-separated. The results highlight the importance of parameter tuning and dataset characteristics in achieving high classification accuracy.