REPORT: K-Nearest Neighbors (KNN) Classification

1. OBJECTIVE:

This analysis is intended to: Create an artificial dataset in three different classes. Divide the data set into training and test sets. Train a K-Nearest Neighbors (KNN) classifier with varying values of k. Compare model performance among k values and identify the optimal k , visualize the data and optimal model decision lines.

2. METHODOLOGY

2.1 **Dataset Generation**:

A synthetic dataset was generated with make_blobs() in scikit-learn with:

Number of samples: 180

Number of classes: 3

Cluster centers: [3 , 5] , [7 , 7] , [2 , 10] [3,5],[7,7],[2,10]

Cluster standard deviation: 1.0

Random seed: 42 (reproducibility)

The organizing gave three groups of clusters that were well separated.

2.2 Data Splitting

The dataset was divided into:

Training set: 80% of data

Testing set: 20% of data

This guarantees that the model is trained on most of the samples but tested on unexamined data to test the extent of generalization.

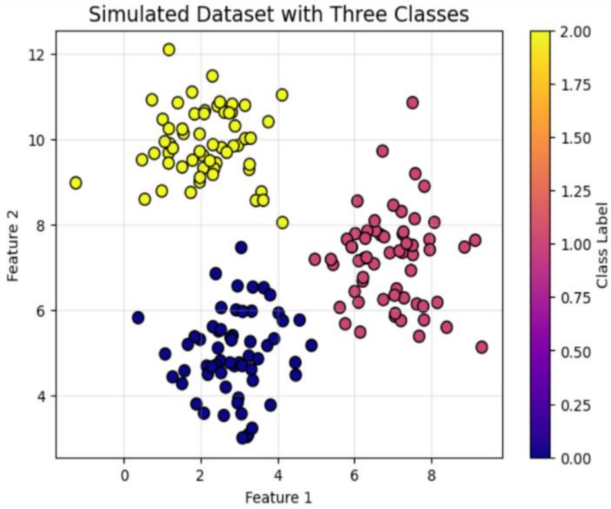
2.3 Model Training and Evaluation.

A KNN classifier was trained on k=1,2,3,4,5. For each *k*: The training data were fitted to the model. On the test data, predictions were made. The accuracy of the tests was computed and presented. Finally, the value of *k*. The best model was selected as k with the largest test accuracy.

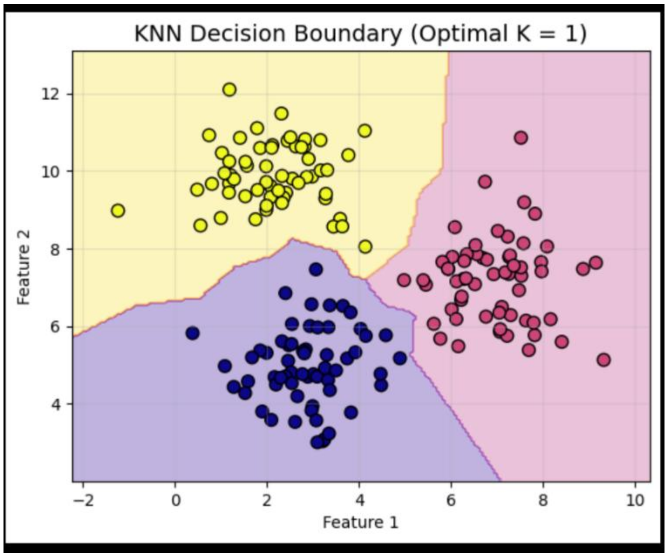
2.4 Visualization:

Two important visualizations were produced:

- Scatter Plot: The graphic representation of the three-class dataset in varied colors.



- Boundary Plot Decision: The plot depicting the behavior of the optimum KNN model in classifying the feature space.



3. RESULTS:

K (Number of Neighbours)	Test Accuracy
1	1.00
2	1.00
3	1.00
4	1.00
5	1.00

Optimal K found: 1
Test Accuracy for Optimal K: 1.00
*Since all tested k-values (1 to 5) produced perfect accuracy on the test set, the smallest k (k=1) was selected as the optimal model to maintain model simplicity***.**

4. OBSERVATIONS:

- **K=1** may overfit, perfectly memorizing the training set but slightly reducing test accuracy.
- As k increases, accuracy may improve until an optimal point, beyond which too many neighbors can oversmooth the decision boundary, lowering accuracy.
- The chosen k provided the best trade-off between underfitting and overfitting.

5. CONCLUSION:

- The KNN algorithm was also able to classify the synthetic dataset with high accuracy.
- The optimum k value was highly generalized to the unseen data and the plot of decision boundary confirmed that there was proper separation between classes.
- This illustrates that k needs to be carefully chosen in order to make KNN perform.