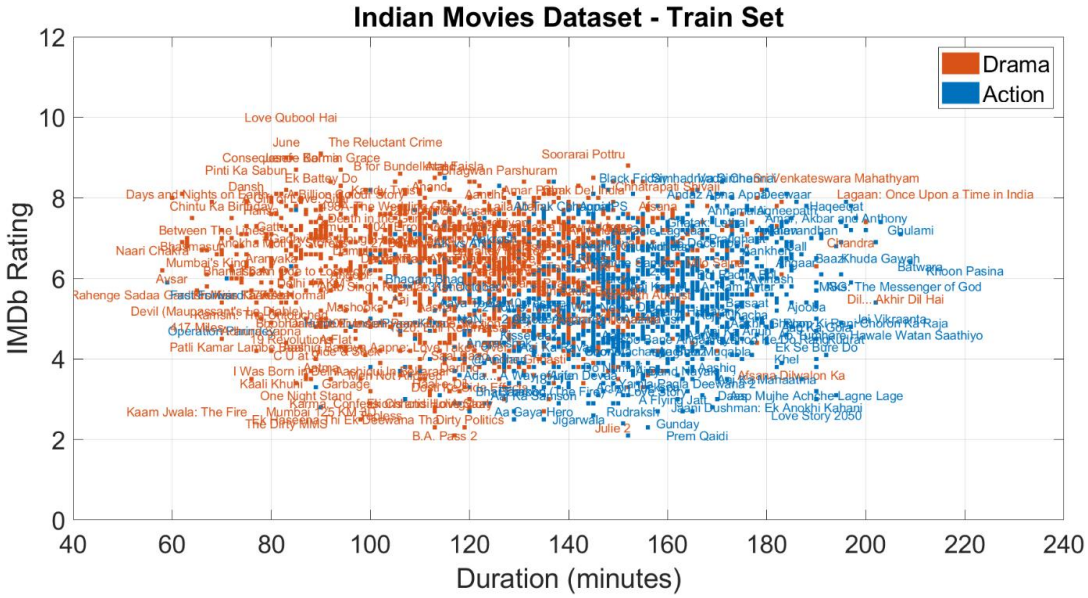
MACHINE LEARNING AND PATTERN RECOGNITION

LAB WEEK 6

1) Import the following libraries:
    - *Pandas*
    - *Matplotlib.pyplot*
    - *Numpy*

2) Import the training and test datasets using *pandas*.

3) Plot the training and test data sets having 'Rating' on the y-axis and 'Duration' on the x-axis. Also, label the data into the two classes 'Action' and 'Drama' and annotate each data point with the name of the movie:



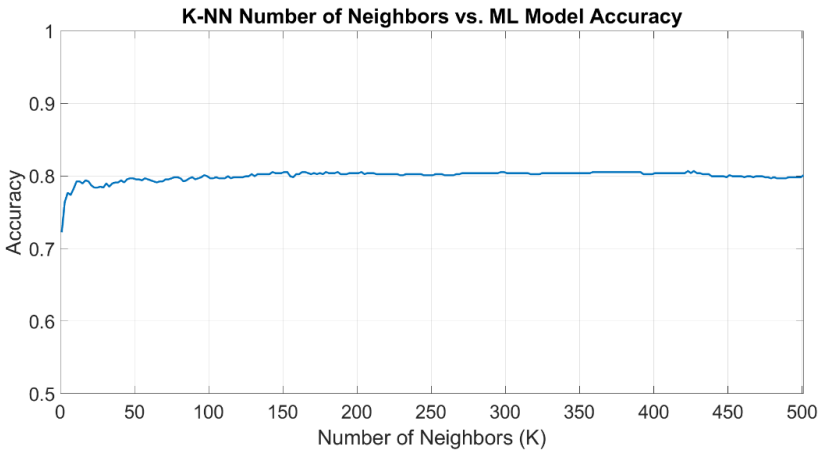
4) Model Development and Evaluation:
    - Import *KNeighborsClassifier* from *sklearn.neighbors*
    - Import *confusion_matrix, ConfusionMatrixDisplay* from *sklearn.metrics*
    - Define a list: *knn_neighbors* = [1, 3, 5, ……., 499]
    - for each *k* in *knn_neighbors*:
        1. apply the k-NN algorithm on the train set

(For Reference: https://www.geeksforgeeks.org/k-nearest-neighbor-algorithm-in-python/)

2. use the test dataset to the find the predictions set
3. use the test dataset and the predictions set to get the confusion matrix for each *k*
4. use the Confusion Matrix to calculate the accuracy for each *k*:

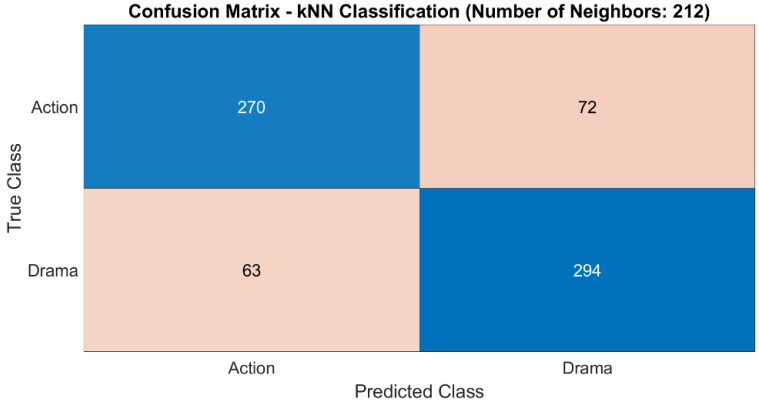
   Accuracy = (True Negatives + True Positives) / Total no. of data points

5) Plot Accuracy vs No. of Neighbors (*k*):



6) Find the value of *k* for which the accuracy is the maximum. Use the *max()* and *list.index()* to find that value of *k* and the corresponding maximum accuracy.

7) For that value of *k*, display the confusion matrix and calculate the performance metrics i.e. precision, recall, overall_precision, overall_recall and F_score.



Report: Answer the following questions within your report:

1.What does the 'k' value in KNN determine?

2. How is the optimal 'k' value in KNN typically chosen?

3. How does the value of 'k' affect variance and bias in KNN?

4. Is KNN sensitive to outliers, and why?

5. How does the scale of features affect the performance of KNN?

Submission Instructions:

- **Items to be uploaded on LMS:** Main code file along with the outputs and the report answers in one file in **PDF format**. Ensure that your file contains the following plots:
    - Train dataset plot annotated with movie names
    - Test dataset plot annotated with movie names
    - Accuracy vs No. of Neighbors plot
    - Final confusion matrix and performance metrics (screenshot)

- **Please note:** Write your **comments along with the code** for each step as required. All plots must be **correctly labelled**.

- Your file should be as follows. Yourname_lab6.pdf

- Due time and date are given on LMS. Submit it before the deadline.