Assignment 7

Q1) KNN – Is a supervised learning technique that considers the k(number) nearest neighbour. Consider the following training and validation set for a movie dataset. Your objective is to identify the movie class/category given in the test data set based on the number of comedy and action scenes. You are advised to use the concept of array to accomplish this task.

Variable used the code should be a combination of alphabet of your name. (Example if name is Rishav Singh then the variable names are RSG, SIN, RIS_SIN etc.)

Training Data:

Sl No.	No of Comedy Scene	No Of Action Scene	Actual Classes
1	100	0	Comedy
2	0	100	Action
3	15	90	Action
4	85	20	Comedy

Validation Data:

Sl No.	No of Comedy Scene	No Of Action Scene	Actual Classes
1	10	95	Action
2	85	15	Comedy

Note: Validation data is used to check the accuracy of your algorithm.

Test Data:

Sl No.	No of Comedy Scene	No Of Action Scene	Classes
1	6	70	?
2	93	23	?
3	50	50	?

Hint: For simplicity you can change the datatype of class to Boolean (Action – 0, Comedy - 1)
Step1: Compute similarity using Euclidean Distance for each validation data row to each train data row:

$$D1 = \sqrt{((10-100)^2 + (95-0)^2)}$$

Consider "No. Of Comedy Scene" in X – axis and "No. Of Action scene" in y-axis
You will be getting 8 distance values (D1 to D8) as shown below for validation data.

Sl No	Validation Data	Train Data	Predicted Class	Euclidean Distance
D1	1	1		
D2	1	2		
D3	1	3		
D4	1	4		
D5	2	1		
D6	2	2		
D7	2	3		

Assignment 7

D8	2	4		

Step2: Sort the above matrix based on Euclidean Distance in ascending order.

Sl No	Sorted Distance	Train Data	Validation Data

Step 3: Compute the accuracy for different values of ‘k’

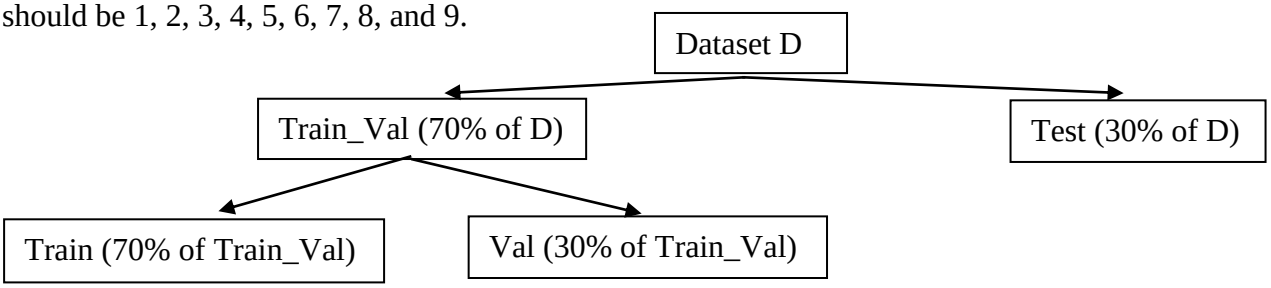
Sl No	Val Data	Accuracy for K=1	Accuracy for K=3

Accuracy = (No. Of Correct Predictions/ Total number of samples) * 100

Step 4: Use the value of ‘k’ that achieved the highest accuracy on the validation set when evaluating the test dataset.

Sl No.	No of Comedy Scene	No Of Action Scene	Predicted Class	Value of ‘k’
1	6	70		
2	93	23		
3	50	50		

Q2) Apply the concept used in Q1 on the Iris dataset (D). Find the highest accuracy obtained on the test dataset and the corresponding value of k used during the test phase. The values of k considered should be 1, 2, 3, 4, 5, 6, 7, 8, and 9.



Note:

1. **Convert the Jupyter file to PDF and upload the PDF file. No other file format shall be acceptable.**
2. **PDF Name should be A07.pdf**