GOA COLLEGE OF ENGINEERING

“Bhausaheb Bandodkar Technical Education Complex”

Experiment No: 3 **Date:**

Lab Session 3: K Nearest Neighbour Classifier

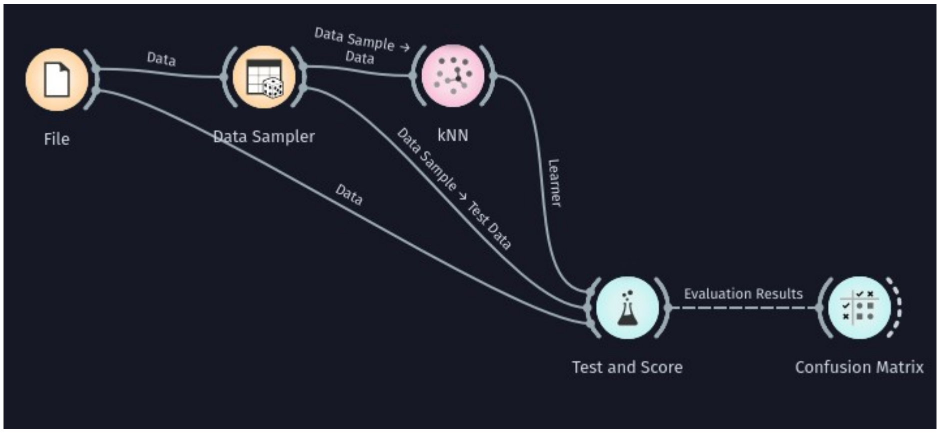
Aim: Implement KNN classifier on the following datasets:
a. Iris Dataset
b. Titanic Dataset
c. Placement Dataset

Problem Description: Implement the KNN Classifier and note the following observations for each of the datasets:
a. The need for Data Sampler
b. The impact of k on the performance of the classifier
c. The impact of the splitting ratio on the performance

Widgets Used:

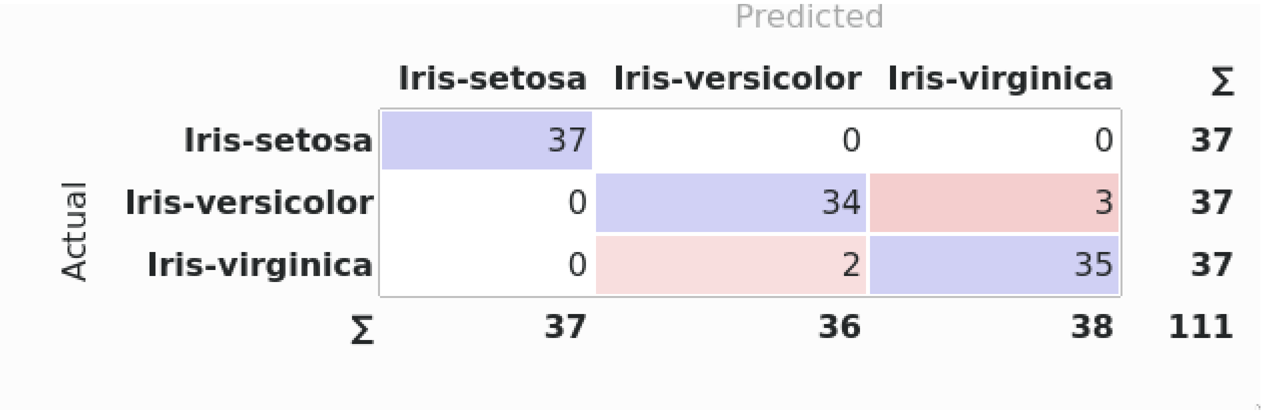
- Data Sampler (Refer Error: Reference source not found)
- Test and Score
- KNN
- Confusion Matrix
- CSV File Import

Data Workflow:



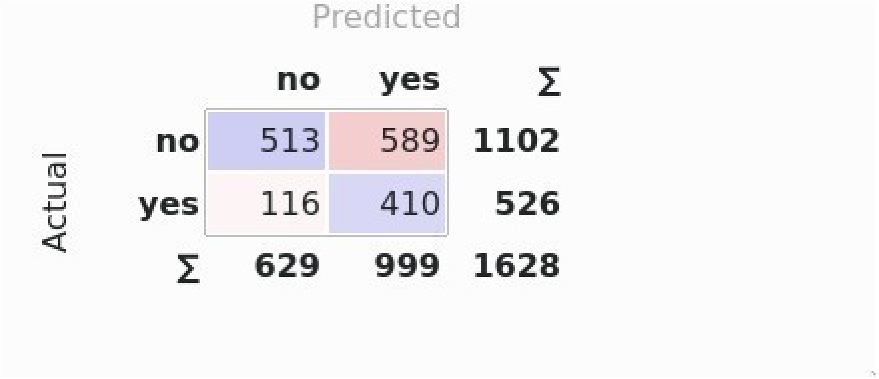
Output:

Iris Dataset: 1) k=3



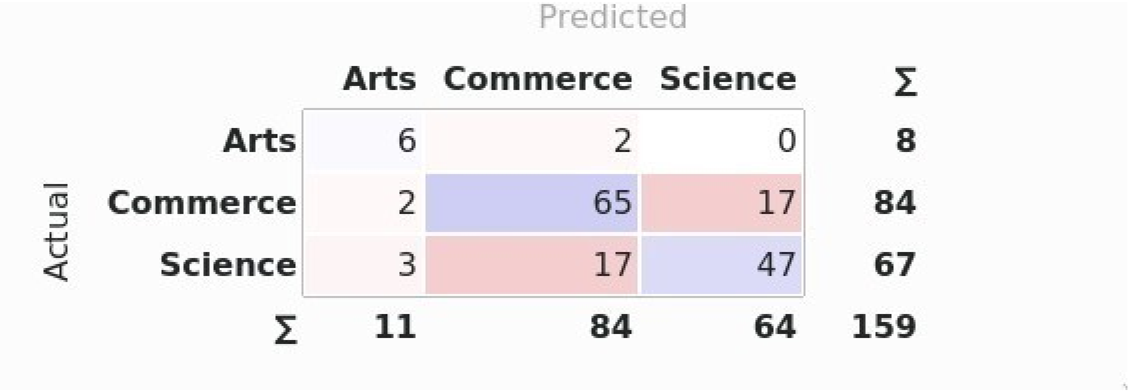
		Predicted			
		Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Actual	Iris-setosa	37	0	0	37
	Iris-versicolor	0	34	3	37
	Iris-virginica	0	2	35	37
	Σ	37	36	38	111

Titanic Dataset: 1) k=3



		Predicted		
		no	yes	Σ
Actual	no	513	589	1102
	yes	116	410	526
	Σ	629	999	1628

Placement Dataset: 1) k=3



		Predicted			
		Arts	Commerce	Science	Σ
Actual	Arts	6	2	0	8
	Commerce	2	65	17	84
	Science	3	17	47	67
	Σ	11	84	64	159

Observations and Conclusion:

- The impact of k is seen as follows on the evaluation metrics:
  - Iris Dataset

	Accuracy	Precision	Recall	F score
K=3	0.955	0.955	0.955	0.955
K=5	0.964	0.965	0.964	0.964
K=7	0.973	0.975	0.973	0.973
K=11	0.991	0.991	0.991	0.991

  - Titanic Dataset

	Accuracy	Precision	Recall	F score
K=3	0.567	0.685	0.567	0.575
K=5	0.504	0.656	0.504	0.501
K=7	0.353	0.566	0.353	0.255
K=11	0.353	0.566	0.353	0.255

  - Placement Dataset

	Accuracy	Precision	Recall	F score
K=3	0.742	0.746	0.742	0.743
K=5	0.66	0.654	0.66	0.654
K=7	0.616	0.605	0.616	0.607
K=11	0.56	0.533	0.56	0.531

- The impact of split ratio for a particular value of k=11
  - Iris Dataset

Split Ratio	Accuracy	Precision	Recall	F score
50-50	0.987	0.987	0.987	0.987
60-40	0.983	0.984	0.983	0.983
70-30	0.978	0.979	0.978	0.978
90-10	1.0	1.0	1.0	1.0

  - Titanic Dataset

Split Ratio	Accuracy	Precision	Recall	F score
50-50	0.351	0.561	0.351	0.251
60-40	0.340	0.522	0.340	0.233
70-30	0.344	0.537	0.344	0.240
90-10	0.359	0.595	0.359	0.261

  - Placement Dataset

Split Ratio	Accuracy	Precision	Recall	F score
50-50	0.561	0.533	0.561	0.538
60-40	0.593	0.563	0.593	0.565
70-30	0.5	0.465	0.5	0.469
90-10	0.476	0.433	0.476	0.433

Conclusion:

- For the iris dataset with 3 target classes, increasing the k value seems to increase the accuracy of the classifier
- For the titanic and placement dataset with 2 and 3 classes respectively, increasing the k value seems to decrease the accuracy of the classifier
- Increasing the sampling ratio seems to decrease the accuracy
- The 90-10 sampling ratio appears to cause an increase in the accuracy as compared to previous ratios possibly due to lack of testing data
- For the placement dataset, increase split ratio seems to increase accuracy intially but for larger splits the accuracy decreases.