EET 4501 – KNN ASSIGNMENT

OBJECTIVE: KNN in Scikit-Learn

Scikit-learn is one of the commonly-used Python libraries for machine learning which is built on Numpy, SciPy, and matplotlib. Scikit-learn helps in data preprocessing, feature selection, classification, regression, clustering, and model selection.

HELPFUL RESOURCES:
Reviewing the following resources and viewing the recorded lab tutorial video before starting the lab exercise will be helpful:

- Scikit-learn interactive documentation
- KNN in Scikit-learn
- Distance metrics in Scikit-learn
- Feature scaling with Scikit-learn
- Randomly split data into test and train subsets in Scikit-learn

Write Python code and document it to do the following:
First, download the sample data (**'knnData.csv'**) from Canvas. This dataset contains twodimensional samples that belong to one of the two classes: 0 or 1. The class labels (0 or 1) for each sample are located in the last column. The first and second columns contain the coordinates of the 2D points representing the samples.

1. Exploring of Data **(4 marks)**

- **a.** Load data into a dataFrame named (**df**) using Pandas. **(1 mark)**
- **b.** How many samples per class do we have? **(1 mark)**
- **c.** Visualize the dataset samples using a scatter plot in a 2D space, using different colors and point shapes for each class. **(2 marks)**

2. Training Phase **(4 marks)**

- **a.** Split (**df**) into two data frames: features and labels **(2 marks)**
- **b.** Split data into two random subsets: train (70%) and test (30%) **(2 marks)**

3. Learn a *k*-Nearest Neighbor (*k*-NN) model **(8 marks)**

a. Select **two** different distance metrics and set the number of neighbors to **1, 5, 10, 25.** Then, estimate the accuracy of the *k*-NN classifier for different distance metrics and different k values and display results for the following 8 different settings: **(4 marks)**

	K = 1	K = 5	K = 10	K = 25
Distance Function 1	Accuracy			
Distance Function 2				

For example:

Distance Function 1 = "Euclidean"

K = 1: Accuracy is

K = 3: Accuracy is

…

b. Discuss the results obtained in part (c). What are the best hyperparameters for this data (the number of neighbors, distance metric)? Which hyperparameter has more effect on accuracy (distance function or the number of neighbors)? How do you know? **(1 marks)**

c. Plot the confusion matrix for the *k*-NN algorithm with the best hyperparameters obtained in part (d) **(1 mark)**

d. Report other evaluation metrics (including Precision, Recall, and F-Score) per each class and on average for the k-NN with the best hyperparameters. **(2 marks)**

4. Data Scaling **(4 marks)**

a. Scale the data using "MinMaxScaler" scaler. **(1 marks)**

b. Compare the performance of the *k*-NN classifier **with** and **without** scaling data, using different numbers of neighbors (**1, 5, 10, 25**). **(2 marks)**

c. Which one has better accuracy? Does scaling improve the performance of the *k*-NN classifier for this dataset? **(1 mark)**

	K = 1	K = 5	K = 10	K = 25
Scaled Data	Accuracy			
Non-scaled Data				

WHAT TO SUBMIT ON Canvas:

1. 'assignment4.ipynb' file, including the documentation and written python code
2. 'FirstnameLastname_ assignment4.pdf' (for example, AmnaMazen_ assignment4.pdf), including the code and outputs of each part of the lab exercise

Note: Proper documentation should be provided for your code using markdown.
EVALUATION: You will be evaluated based on your solutions for the problems based on the following scheme:

1. Does the code run and meet specifications?
   - Is input adequate and input data type properly validated?
   - Is processing adequate?
   - Is output correct and adequate?
   - Is the code compliable?
   - Is the code run properly?
2. Is the code properly commented?
   - Are the program title, programmer's first and last name, and the date posted at the top in a multi-line comment?
   - Is each significant step of the program properly commented?
   - Are comments added to clarify details?
   - Are comments clear, accurate, and neatly formatted?

IMPORTANT: ASK QUESTIONS IF YOU GET STUCK, BUT DO YOUR OWN CODE. ANY CODE SUSPECTED TO BE SIMILAR TO ANOTHER SUBMISSION WILL CAUSE BOTH SUBMISSIONS TO RECEIVE A ZERO MARK ON ALL LABS AND BE REPORTED FOR PLAGIARISM.