AML Project1 – Exploring Labeled Data Using kNN

The objective of this project is to understand the workflow of supervised learning. You will utilize the simple and popular classifier kNN (k-nearest neighbors) to explore a labeled dataset of your choice. You are required to pick an interesting dataset to demonstrate your understanding of kNN and how to evaluate a machine learning model.

This project is to be done in a group of **3** students. The responsibilities of the students in a group must be clearly delimited and stated in a statement (attached to your project) so that each of you can be graded fairly and separately.

Requirement:

1. Source code in Python (done in a group).
2. **Group project presentation** in class (12 min), Q&A (3 min)
3. **Individual project report** (3/4 pages, no longer than 5 pages, fonts>=12)

Specific requirements:

- Your data must be an appropriate dataset to illustrate the workflow of kNN and it should have at least 4 attributes and 40 instances (data points).

- Your Python script must answer the following questions:
    1. How many features/attributes does the dataset have?
    2. What is the class distribution?
        a. How many instances in class1, how many in class2, …
        b. Visual display of the instances (data points) with different classes colored differently
            i. Data points can be plotted in reduced dimensions (several 2 or 3)
    3. Dataset partition (training, testing); how many percent of the data is used for training and how many percent for testing
    4. What distance metric is used?
    5. Testing result; what's the estimated accuracy of your model on future data
        a. Present your accuracy in a confusion matrix.

- Your individual report reports and explains your exploration of kNN and the dataset you picked. It must be well organized (nice and neat) and should include:
    1. The description of the problem you want to solve using kNN
    2. Background of your data
    3. Statistical summary of your data
        - what are: max, min, mean, median, mode, and standard deviation of each class.
    4. Short description of your understanding of kNN

5. Summary of your classification results, include accuracy and confusion matrix/matrices for a few distance metrics, k values, and different partitions of training and testing datasets
6. Discussion of your results
    - What distance metric your best kNN uses; why you think there is a good reason the metric works the best
    - The k that gives the best performance; your intuitive explanation on why
    - Explain how the accuracy depends on the partitions of your dataset; What are the pros and cons of having a large training subset
    - User your dataset to illustrate the data leakage problem and explain the implementation of your model does not have data leakage issue
7. Conclusion of your exploration. Wrap up what you've learned in the project, the pros and cons of kNN as a supervised learning model for your dataset.
8. (**Graduate students**) A summary of commonly used nearest neighbor finding algorithms (additional 1 or 2 pages)

Your presentation must show:

1. Your understanding of supervised learning and kNN
2. A good story about your dataset and
3. What your exploration tells us
4. Demo of your code

Submission instructions:

1. Due at the beginning of the due day
2. An electronic copy of your Python scripts, drop it to the drop box at Brightspace (yes, need only one copy per group)
3. An electronic copy of your **individual project report** (words or pdf), drop it to the drop box at Brightspace
4. ppt presentation slides, drop it to the drop box at Brightspace (yes, need only one copy per group)
5. An individual statement of the responsibilities of each member in your group

Here a couple of example datasets:

1. Iris dataset: The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor), 150 instances, 4 attributes, 3 classes. The dataset in Excel can be downloaded **here**. The task: determine the species of a given unknown iris plant.
   The description of the dataset: https://en.wikipedia.org/wiki/Iris_flower_data_set. Make sure you understand what the data set contains.
2. Wisconsin breast cancer dataset: 569 instances, 32 attributes, 2 classes (malignant/benign). Dataset can be founded at: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)