Assignment 2 – KNN Digits Classifier

**Description:**

KNN is a classification algorithm that makes predictions based on the distance between a testing sample and the samples in the training set. Though KNN is a simple algorithm, it may work surprisingly well if the test data and train data are from the same data distribution. For this assignment, you need to build a KNN classifier for digits classification using the *scikit-learn* *digits* dataset.

**Purpose:**

* Get familiar with Python programming language and the scikit-learn library.
* Develop a KNN algorithm for a given task.

**Directions:**

For this assignment, you need to build a KNN classifier from scratch. Below is a detailed instruction of what you may need to do.

* Dataset Preparation
  + You need to load the dataset using *sklearn.datasets.load\_digits*.
    - More information about the function can be found at: <https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html#sklearn.datasets.load_digits>
  + After loading the dataset, randomly shuffle the dataset to split the dataset to train/dev/test sets.
    - Use the 70% of data for the train set, 15% for the dev set, and 15% for the test set
    - You need to make sure that the labels and images are still matching after shuffling the data.
    - You may want to use the random shuffle function provided by Numpy.
* KNN Development
  + You need to write your own distance comparison function
  + Use the train set as the training data, and use the dev set to determine the best K and best distance metric.
    - You may need to test multiple K values and distance metrics to select the optimal ones.
* Test the Model
  + After the optimal K value and distance metric are selected, test the model using the test set.
* Submission
  + You need to submit a written report for this assignment.
  + For this report, you need to:
    - Explain what you have done
      * E.g., what distance metrics you have tested, what K values you have tested, etc.
    - Report the best performance on the test set (in terms of accuracy)
      * You also need to indicate the K value and distance matric for achieving this result
    - Visualize the prediction result
      * Randomly select 10 data samples from the test set and specify the ground truth label and the predicted label for each of the samples.
    - Include your code as an appendix
      * You could save your Colab code as a PDF file and attach it to your report, or you could copy and paste your code into the report.
        + If you want to copy/paste your code, make sure to maintain the appropriate indentation and make the code readable.

**Evaluation Criteria:**

This assignment is worth 100 points.

* (40 pts) The code must be runnable and works as expected
* (60 pts) The report must contain the four components