Homework 4

Due date: by 11:59**pm** on February 21, 2022

For this homework, you will try multiple supervised learning algorithms on the given dataset. The dataset can be found in the same folder where this file is.

Things to do:

1. **[10 points]** Notice that the labels are in the first column (at 0^{th} column in a Pandas dataframe). The rest of the columns are representing features. Separate your X and y accordingly.

2. **[10 points]** Randomly choose 20% of the samples as your test set and the remaining samples as your training set.

3. **[10 points]** Now pick any sample from your training set. Reshape it into a (28, 28) array. Research "imshow" for Python. Then visualize the sample using imshow. It should look like one of the 10 digits.

4. **[20 * 3 = 60 points]** Run the following algorithms on the dataset and report your *test* accuracy.
    a. Logistic Regression
    b. KNN
    c. One algorithm of your choice (*suggestions*: Random Forest, Support Vector Machine etc.). Any algorithm that can perform classification tasks will be accepted.

5. **[10 points]** Show the performance of the three algorithms using a visualization method of your choice (e.g. column charts).

6. **[bonus(optional): 10 points]** Write a report (minimum 2 pages). Make sure the report has the following sections:

    a. Introduction
    b. Description of the dataset
    c. Challenges faced
    d. Algorithms used
    e. Results

What to submit:

1. Put the dataset, your Jupyter Notebook (and the optional report) in a folder. Your code should refer to this folder when loading the data. Zip the folder, and name it: yourLastName_yourFirstName.zip.

2. Upload the zipped folder on Canvas.