Assignment 1
(Statistical Machine Learning)

Instructions:

1. All your implementations should be from scratch, and preferably use a utils.py file to implement the algorithms as classes.
2. You can (preferably should) import them into a notebook and run them.
3. Make sure the notebooks are self-explanatory, easy to read, and all plots are well labeled.
4. You can use existing libraries to load the images.
5. You can use numpy library for array/matrix processing

Q1. In this question, you're going to use PCA to select the required subset of features from MNIST and then perform classification on it.

a) Download the dataset from here: https://www.kaggle.com/datasets/scolianni/mnistasjpg
b) Load the image dataset in your environment and convert it into a suitable format for creating an ML model.
c) Implement PCA from scratch i.e. define your own PCA function or class for it. You are not allowed to use ANY pre-built implementation from any library. Note: Your implementation of PCA should be general, i.e., it can work for any dataset.
d) Use kNN to train ML models on the training set of both original features and the transformed features (number of PCs = 5, 25, 125) obtained from PCA. Run the trained models on the test set and report the classification accuracies.
e) Plot explained-variance (ratio of eigenvalue and sum of all eigenvalues) v/s PCs. How many PCs do you need to select to cover at least 80% of the variance? Do all this through programming.

Q2.Implement k-means clustering and silhouette analysis algorithms to cluster the given data and find the optimal 'k' value. Also, implement fuzzy c means and report the J (objective function) value when c= optimal k.Assume m=2 and beta=0.3.

Data:
https://drive.google.com/file/d/1-0zx-cXze6ja777SN_NkMYVCzidU2lXw/view?usp=sharing

Q3.
https://drive.google.com/file/d/15-6l7_51OZ3wIw37d8a6SX2dfWxUQGl4/view?usp=sharing
Implement the mean shift algorithm and perform image segmentation for the above image. Use a suitable bandwidth such that the vegetables look as separated as possible.

Q4. Implement ICA from scratch to separate mixed signals

- Generate a sinusoidal wave and a ramp wave. Plot them.
- Use the following mixing matrix
  [[0.5, 1],
  [1, 0.5]]
  to mix them. Plot the mixed signals.
- Now use ICA to recover the original signals and plot them.

Q5.

AGE	LOAN(Million Dollars)	HPI	BHK
25	40	135	2
35	60	256	3
45	80	231	3
20	20	267	4
35	120	139	4
52	18	150	2
23	95	127	2
40	62	216	4
60	100	139	2
48	220	250	3
33	150	264	4

On the above data, use kNN to find HPI (continuous) and BHK (discrete) for a test instance having Age = 37 and LOAN = 142 as its feature values. Do it for k=1, 2, 3.