General instruction.
1. The following languages are acceptable: Java, C/C++, Matlab, Python and R.
2. You can work in team of up to 3 people. Each team will only need to submit one copy of the source code and report.
3. You need to submit your source code (self contained, well documented and with clear instruction for how to run) and a report through the TEACH site https://secure.engr.oregonstate.edu:8000/teach.php?type=want_auth
   Please clearly indicate your team members’ information.
4. Be sure to answer all the questions in your report. Your report should be typed, submitted in the pdf format. You will be graded based on both your code as well as the report. In particular, the clarity and quality of the report will be worth 10% of the pts. So please write your report in clear and concise manner. Clearly label your figures, legends, and tables.
Kmeans clustering and dimension reduction

For this assignment, you will implement the Kmeans clustering algorithm as well as the PCA dimension reduction method and test them on the digits 4-9 training data set. This dataset has been used previously for training in the logistic regression assignment. For clustering, you will only consider the first 256 features. The last column of the data stores the class label and should not be used during clustering and only be used during evaluation of the clustering results.

(a) (20 pts) Apply your implemented Kmeans algorithm to the usps-4-9-train data to find \( k = 2, 4, 6, 8 \) clusters in the data. For each different \( k \) value, you need to run your Kmeans algorithm with 10 different random initializations and select the clustering solution that achieves the lowest Sum of Squared Errors. This will give you four different clustering solutions, each with a different number of clusters.

(b) (10 pts) For each choice of \( k \in \{2, 4, 6, 8\} \), report the confusion matrix between your clustering result and the class label provided in the last column of the data. A confusion matrix is defined as follows. If the clustering has \( k \) clusters and the ground-truth class labels have \( c \) classes. The confusion matrix \( C \) will be of size \( k \times c \), where \( C(i, j) \) stores the total number of examples assigned to cluster label \( i \) and belong to class \( j \), where \( i \in \{1, \ldots, k\} \) and \( j \in \{1, \ldots, c\} \). For this part, you will need to report four confusion matrices, one for each \( k \) value.

(c) (10 pts) For each choice of \( k \in \{2, 4, 6, 8\} \), in addition to the confusion matrix, please report the class purity measure achieved by your algorithm. Given a clustering result, this measure can be easily computed based on its confusion matrix as follows. For each cluster \( i \in \{1, 2, \ldots, k\} \) (i.e., the \( i \)-th row of the confusion matrix) find the majority class (the largest number in the row, denoted by \( \text{max}_i \)) and assign all examples in the cluster to that majority class. After the assignment of all clusters, we measure class purity as the percentage of correctly assigned examples, i.e., \( \frac{\sum \text{max}_i}{n} \), where \( n \) is the total number of examples in the dataset.

(d) (5 pts) Do you observe any trend in terms of how the purity measure changes as we increase the number of clusters we look for in the data using Kmeans? Provide an explanation for your observation.

(e) (25 pts) Apply Principal Component Analysis (PCA) to the 256 input features to preserve 80% of the total data variance. Repeat the first 3 steps to the reduced data. Compare the new results on the reduced data with the previous results without dimension reduction. What impact do you observe from the dimension reduction both in terms of clustering performance and the run time?