CSCE 4930 - Introduction to Machine Learning
Assignment 1

This is an individual assignment. Rules governing academic integrity are strongly enforced without exception.

1. Assume we are given the location of each object and the type of object (triangle, circle or square) along with an identifying name (e.g., T1, C1, S1). The data is tabulated below. Use the provided dataset to solve the problems below. Make sure to illustrate your steps clearly in each answer.

Shape	Coordinates
Circle	C1:(-3,-2,2), C2:(-1,4,-1), C3:(1,3,0)
Square	S1:(-1,-1,3), S2:(1,2,1), S3:(2,-1,-1)
Triangle	T1:(0,2,-1), T2:(3,-2,2)

   a. A new unidentified object arrives at point **O(2,-1,-2)**. Use KNN to classify the new object **O** as a circle, square or triangle using K-nearest neighbors algorithm using Euclidean distance as your distance / similarity measure.
   Show your prediction result for values of ***k = 1, 4, and 7***. Comment on each result with what you observe. **(10 points)**

   b. Assume we divide the above dataset into a Training set containing {C1, C2, S1, S2, T1} and a validation set containing {C3, S3, T2}. Plot the ROC curve for all possible values of *k.* Identify the best value for *k* based on the plot. Justify your answer **(10 points)**

   c. For the value of *k* you identified in part b, show the following metrics for your model based on the same training and validation sets: Precision, Recall, F-Measure, and Accuracy for each class. **(10 points)**

   d. Show the prediction for object **O** stated in part a. using the value of *k* you identified in part b. Compare your prediction versus the ones you got in part a. illustrating the differences **(10 points)**

2. In this problem, you will be using the MNIST dataset (https://en.wikipedia.org/wiki/MNIST_database) for classifying handwritten digits. You can use the Python libraries we discussed in the lab. The dataset is available in CSV format at (https://github.com/cerndb/dist-keras/blob/master/examples/data/mnist.csv).
   a. Run KNN setting the value of *k* to the square root of the last four digits in your student ID. Use 10-fold cross validation and report the accuracy, precision, recall, F-measure, and AUC for each of the digits classifications **(10 points)**

   b. Modify the parameters of KNN in Python to automatically identify the value of *k* using cross-validation. Report the same measures as in part a. Comment on the results. **(10 points)**

   c. Repeat part b. using Manhattan distance instead of Euclidean distance. Comment on the results **(10 points)**

3. Assume we are given the following data about the users of a movie streaming website as tabulated below. Each user has a set of characteristics (age, gender, country, signup date) and the history of his/her rating on a set of 5 movies. Use 3-NN to answer the questions below. Make sure to illustrate your steps

Age	**Gender**	**Country**	**Signup**	**Lion King**	**Hero**	**StarTrek**	**Wall-E**	**Inception**
28	M	EG	Feb,2018	2.8	3	4.4	4	4.1
22	F	DE	Mar,2016	4.1	4	3.6	4.1	4
24	F	EG	Jul, 2017	3.9	4.5	3.1	3.7	4.6
20	M	EG	Aug,2016	3.2	3.5	4.8	4.3	4.9
32	M	DE	Dec,2017	3.9	3.8	4.1	3.5	4.2

a. Estimate the ratings for all movies for a new user who is a 26 year old male from EG who signed up in December 2016. **(10 points)**

b. Assume we have a new user in the system with missing personal information. We only know about his/her prior rating for Lion King (3.7), Hero (3.9), Inception (3.8), and StarTrek (4.1). Can you use only his/her rating history to predict how this user will rate the movie Wall-E? **(10 points)**

c. Assuming that a rating of **3.7** or more indicates that the user liked the movie, and any rating lower than that means that the user didn't like it. Alice (27 years old female from DE who signed up in May 2017) liked Inception, StarTrek, and Wall-E but didn't like Lion King. Can you predict whether she will like the movie Hero or not? **(10 points)**