Republic of the Philippines
POLYTECHNIC UNIVERSITY OF THE PHILIPPINES

COSC-E4
ELECTIVE 4 (DATA MINING)
Home Activity # 4

Name: Cueva, Larry Miguel, R. **Year and Section: BSCS 4-2**

Predictor	**Below Standard (0 pt.)**	**Approaching Standard (2 pts.)**	**At Standard** (4 pts.)	**Above Standard (6 pts)**
Completion	Student turned in assignment but mostly incomplete	Some of the assigned work is complete	Most of the assigned work is complete	All of the assigned work is complete
Accuracy	Little to none of the answers are correct	Some of the answers are correct	Most of the answers are correct	All of the answers are correct
Work Shown	Student did not show any work	Some steps for problem solving are missing	Most work is meticulously shown	All work is meticulously shown
Neatness	Homework is messy, disorganized, and difficult to read or understand	Homework is somewhat neat and organized, but it could benefit from further editing or refinement	Homework is neat, well-organized, and visually appealing. It is easy to read and understand, with no major errors, mistakes, or inconsistencies	Homework is exceptionally neat, well-organized, and visually appealing. It is engaging and easy to read and understand, with no errors, mistakes, or inconsistencies.

Instructions.

Please provide a clear and detailed solution, including all necessary steps and explanations, for each exercise problem. Ensure that your solution is supported by appropriate reasoning, proofs, or calculations, and not solely based on the final answer. Complete solutions with proper justifications will be evaluated.

1. The dataset given below is for the company that produces ***paper tissues***. The company works in biological science field, and they conducted survey to gathered data, asking opinion of people and they want to test the two attributes: the **Acid Durability** and **Strength.** The objective is to test these two attributes to classify whether a special paper tissue is good or not. The company wants to predict how well these products or typically the types of paper tissue they are producing are accepted by their clients.

   Now, the factory produces a new paper tissue that pass laboratory test with **Acid Durability of = 3** and a **Strength** of **7**. Find out how close this particular tissue type, the distant measure. Guess what the classification of this new tissue using **KNN.**

Name	**Acid Durability (x2)**	**Strength (y2)**	**Class**
Tissue_Type-1	7	7	Bad
Tissue_Type-2	2	3	Good
Tissue_Type-3	3	4	Good
Tissue_Type-4	6	7	Bad

Tissue_Type-5	7	4	Bad
Tissue_Type-6	1	4	Good
Tissue_Type-7	2	7	Bad
Tissue_Type-8	3	6	Bad
Tissue_Type-9	3	4	Good

K = 3

Distance to tissue type	Formulae	Calculated Distance
Tissue_Type-1	$\sqrt{(7-3)^2+(7-7)^2}$	4.0
Tissue_Type-2	$\sqrt{(2-3)^2+(3-7)^2}$	4.24
Tissue_Type-3	$\sqrt{(3-3)^2+(4-7)^2}$	3.0
Tissue_Type-4	$\sqrt{(6-3)^2+(7-7)^2}$	3.0
Tissue_Type-5	$\sqrt{(7-3)^2+(4-7)^2}$	4.24
Tissue_Type-6	$\sqrt{(1-3)^2+(4-7)^2}$	3.61
Tissue_Type-7	$\sqrt{(2-3)^2+(7-7)^2}$	1.0
Tissue_Type-8	$\sqrt{(3-3)^2+(6-7)^2}$	1.0
Tissue_Type-9	$\sqrt{(3-3)^2+(4-7)^2}$	3.0

Sorted distances

Distance to tissue type	Calculated Distance
Tissue_Type-7	1.0
Tissue_Type-8	1.0
Tissue_Type-3	3.0
Tissue_Type-4	3.0
Tissue_Type-9	3.0
Tissue_Type-1	3.0
Tissue_Type-6	3.61
Tissue_Type-2	4.24
Tissue_Type-5	4.24

Three nearest neighbors are therefore: Tissue_Type-7, Tissue_Type-8, and Tissue_Type-3, which have classes "Bad", "Bad", and "Good" respectively

Majority of classes of the nearest neighbors are "Bad" therefore new tissue type with Acid Durability = 3 and Strength = 7 is classified as "Bad"

Prepared by:
Mr. Montaigne G. Molejon,
MSIT PUP-CCIS, Instructor I

Republic of the Philippines
POLYTECHNIC UNIVERSITY OF THE PHILIPPINES

COSC-E4
ELECTIVE 4 (DATA MINING)
Home Activity # 4

2. You work as a data analyst for a car rental company. Your company has collected data on customers' ***age*** and ***average monthly mileage*** driven. You want to segment the customers based on these two variables to understand different groups tailor marketing strategies accordingly. You decide to use K-Means clustering with Euclidean distance as the similarity measure.

Dataset:

Customer	Age (years)	Monthly Mileage (miles)
1	35	500
2		800
3	45	300
4	22	200
5		400
6	55	700
7	30	250
8	40	350
9		600
10	50	450
	27	

	48	
	33	

Prepared by:
Mr. Montaigne G. Molejon, MSIT
PUP-CCIS, Instructor I

Republic of the Philippines
POLYTECHNIC UNIVERSITY OF THE PHILIPPINES

COSC-E4
ELECTIVE 4 (DATA MINING)
Home Activity # 4

Assuming you want to perform K-Means clustering with **K=3**, perform K-Means clustering with Euclidean distance using the provided dataset above. Tabulate the assignments during each iteration, and determine the final cluster assignments. Assume the ***initial centroids*** are as follows:

Centroid 1: (35, 500)
Centroid 2: (45, 800)
Centroid 3: (22, 300)

Iteration 0/3:

	Centroid Coordinates	Assigned Points
Cluster Centroid 0	(35, 500)	(35, 500), (30, 400), (48, 600), (33, 450)
Cluster Centroid 1	(45, 800)	(45, 800), (40, 700)

Cluster Centroid 2	(22, 300)	(22, 300), (55, 200), (50, 250), (27, 350)

Iteration 1/3

	Centroid Coordinates	Assigned Points
Cluster Centroid 0	(36.5, 487.5)	(35, 500), (30, 400), (48, 600), (33, 450)
Cluster Centroid 1	(42.5, 750.0)	(45, 800), (40, 700)
Cluster Centroid 2	(38.5, 275.0)	(22, 300), (55, 200), (50, 250), (27, 350)

Iteration 2/3

	Centroid Coordinates	Assigned Points
Cluster Centroid 0	(36.5, 487.5)	(35, 500), (30, 400), (48, 600), (33, 450)
Cluster Centroid 1	(42.5, 750.0)	(45, 800), (40, 700)
Cluster Centroid 2	(38.5, 275.0)	(22, 300), (55, 200), (50, 250), (27, 350)

Iteration 3/3

	Centroid Coordinates	Assigned Points
Cluster Centroid 0	(36.5, 487.5)	(35, 500), (30, 400), (48, 600), (33, 450)
Cluster Centroid 1	(42.5, 750.0)	(45, 800), (40, 700)
Cluster Centroid 2	(38.5, 275.0)	(22, 300), (55, 200), (50, 250), (27, 350)

Prepared by:
Mr. Montaigne G. Molejon,
MSIT PUP-CCIS, Instructor I