EGCO 425 – Project 2 (Clustering)

This project can be done by a group of 3-4 students

Total score = 100. Report must be in THAI

1. This project uses bank marketing data set (https://bit.ly/3f5zNyj). Study data description from the website (and files bank-names.txt + targeted_marketing.pdf). You may search for more background information by yourself. Use the small data set (bank.csv with 4521 records) for clustering. All analysis and reports are based on this data set

(15 points) Write down a report. It should include at least: the source of this data set, what it is about, the number of records & attributes, etc. Report interesting statistics that will back further analysis in (2) & (3). Don't just copy statistic reports or tables from the papers

2. Clean and transform this data set until it is ready for clustering. Save it in ARFF

(10 points) Explain the reason you apply each transformation and how. The file you submit must be the final version that is ready for clustering. For example, if you need to normalize any attribute, do it and submit the file after normalization. I'll run only clustering on your data and won't do any cleansing or transformation. If your data set doesn't require any transformation, explicitly say so

3. Run 2 clustering algorithms on the data in (2), excluding the last attribute

(5 x 2 runs = 10 points) Parameter setup for each run. Don't just explain the meaning of a parameter, but explain why you set it to a certain value.

((10 x 2) x 2 run = 40 points) From each run, discuss 2 cluster profiles. These profiles should be different enough that you can tell how 2 clusters differ. Include at least 4-5 attributes in your profile discussion. Use some background/additional information (e.g. from (1)) to make sense of the outputs. You may also check whether the cluster assignment agree with the excluded attribute

4. *(15 points)* Compare and summarize the results from 2 algorithms in (3). Do they agree with each other ? Or which one is better ? Don't repeat what you have said in (3). Generalize your findings as the patterns of customers of this bank

5. *(10 points)* Others (writing, format, etc.)

Submission : Thursday 15 April, 18.00

1. Put the following files in a folder. Name the folder after the ID of your group representative
   - Report in only 1 PDF file
   - Data file in ARFF → the one that is ready for clustering
   - File readme.txt containing names & IDs of every one in your group

2. The group representative submits the whole project to Google Classroom. The other group members submit only readme.txt to Google Classroom