EGCO 425 – Project 1 (Association)

This project can be done by a group of 1-3 students

Total score = 100. Report must be done in THAI

1. This project uses CES data set (https://bit.ly/3b8fqvK). Study data description from the included papers and files.

(10 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, attributes types & meaning, report some interesting statistics

2. From original data set, save records whose income <= 5 to ces_hybrid_lowincome.arff. Save records whose income > 25 to ces_hybrid_highincome.arff. For each new data set, remove attributes that represent city, monthly income, and family size (i.e. the first 23 attributes) & run Apriori on supermarket items

(3 x 2 runs = 6 points) Explain parameter setups for each run. Don't just use default values. Try different setups and choose appropriate ones

((4 x 3) x 2 runs = 24 points) From each run, discuss 3 rules that you find interesting. The discussion should include rule interpretation & measurements (at least support, confidence, lift). The rules from both runs should be related → e.g. identical rules with different measurements, rules with the same LHS but different RHS items, rules that are subsets/supersets of the others -- so you can compare the buying patterns between 2 groups

3. From original data set, save records whose members <= 2 to ces_hybrid_smallfamily.arff. Save records whose members >= 6 to ces_hybrid_bigfamily.arff. For each new data set, remove attributes that represent city, monthly income, and family size (i.e. the first 23 attributes) & run Apriori on supermarket items

(3 x 2 runs = 6 points) Explain parameter setups for each run. Don't just use default values. Try different setups and choose appropriate ones

((4 x 3) x 2 runs = 24 points) From each run, discuss 3 rules that you find interesting. The discussion should include rule interpretation & measurements (at least support, confidence, lift). The rules from both runs should be related → e.g. identical rules with different measurements, rules with the same LHS but different RHS items, rules that are subsets/supersets of the others -- so you can compare the buying patterns between 2 groups

4. *(15 points)* Compare and summarize the buying patterns that you discover from (2) and (3)

5. *(15 points)* Others (writing, format, etc.)

Submission : due Thursday 27 February, 18.00

1. Put the following files in a folder. Name the folder after the ID of one member
   - Report (in .doc or .pdf)
   - Data files in .arff (for 4 subgroups) that you use for association analysis
   - readme.txt containing names & IDs of every one in your group

2. Zip and submit it to rangsipan@gmail.com. Put "EGCO 425 – Project 1" in the email subject. In case of multiple submission, only the earliest version will be marked