Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

SEMESTER 1 EXAMINATIONS 2021/2022

MODULE: CA270 - Data Warehousing and OLAP

PROGRAMME(S):
DS BSc in Data Science

YEAR OF STUDY: 2

EXAMINER(S):
Mark Roantree (Internal) (Ext:5636)

TIME ALLOWED: 2 Hours

INSTRUCTIONS: **Answer all 3 questions.**

PLEASE DO NOT TURN OVER THIS PAGE UNTIL YOU ARE INSTRUCTED TO DO SO.
The use of programmable or text storing calculators is expressly forbidden.
Please note that where a candidate answers more than the required number of questions, the examiner will mark all questions attempted and then select the highest scoring ones.

There are no additional requirements for this paper.

EXAM INSTRUCTIONS

This exam comes with a different set of appendices for each student. The Appendix is not attached to this exam paper but will be available on Loop at the SAME TIME at which this exam starts.

Use the **last 4 digits** of your **student ID** to download the correct Appendix.

WARNING: For open ended questions like Q3, no 2 students should have the same answer!

HINT: Q3 may take longer than Q1 & Q2. Take this into consideration when working on your solutions.

QUESTION 1 (Classification) **[TOTAL MARKS: 30]**

This question uses the dataset in Appendix 1 which is the result of 21 covid-19 tests. The attributes captured are Gender, Age, Setting (where they believed the exposure took place), TripsMade (in the last 5 days), and Home (their living arrangements).

Your task is to determine the result of Test_22 which has the values:
{F, 16, Sports, 3, Family}.

Q 1(a) *[6 Marks]*
Why is it necessary to transform some of the columns? Be precise as to why and the extract transformation carried out.
Provide the new dataset in your answer book and describe the transformation(s) on which you decided.

Q 1(b) *[5 Marks]*
Calculate the prior probability for Test_22.

Q 1(c) *[15 Marks]*
Build a matrix with the full set of conditional and prior probabilities.

Q 1(d) *[4 Marks]*
What is the prediction for Test_22?

[End of Question 1]

QUESTION 2 (Cluster Analysis) **[TOTAL MARKS: 30]**

This question uses the dataset in Appendix 2.
Students are asked to comment Yes or No if they have any of 5 options available to them to travel to college. Each response Yes or No is considered equal.
You are required to cluster students using agglomerative hierarchical clustering.

Q 2(a) *[10 Marks]*
Begin by constructing a dissimilarity matrix.
Explain the distance function you used and how it calculates *one* of your distance values.

Q 2(b) *[10 Marks]*
Show the formation of the first cluster and the subsequent re-computation of the matrix. Explain your decision for the modified values in the matrix.

Q 2(c) *[10 Marks]*
Complete the clustering process, showing each matrix re-computation, until you construct the final dendrogram.

[End of Question 2]

QUESTION 3 (Association Rule Mining) **[TOTAL MARKS: 40]**

This question uses Appendix 3, which contains a list of airport codes. For example, Dublin airport is DUB and De Gaulle airport is CDG. The list represents the minimum support for numbers of evening flights landing at these airports.

Apply the Apriori algorithm to determine the itemsets that exceed a threshold confidence. The table in Appendix 3 is regarded as the set C_1.

Begin by selecting 9 airports (you choose the 9 airports!) which exceed the minimum support and form the set L_1.
Write the set L_1 (as selected by you) in your answer book.

This is important. You are making the decisions as to which itemsets are supported. Choose your itemsets so that when the time comes, C_4 has at least 1 item!

Q 3(a) *[4 Marks]*
Generate and write the set C_2 into your answer book.

Q 3(b) *[4 Marks]*
Now select 8 itemsets from C_2 (you choose this set!) which exceed the minimum support and form the set L_2.
Write the set L_2 in your answer book.

Q 3(c) *[16 Marks]*
Generate and write the set C_3 into your answer book. Clearly explain the optimisation process used in the construction on C_3.

Q 3(d) *[16 Marks]*
From now on, assume that all itemsets of size 3 and above are supported.

Run the Apriori algorithm to completion and clearly show what is happening at each step.
Explain how the algorithm terminates.

[End of Question 3]

[END OF EXAM]