DUBLIN CITY UNIVERSITY

AUGUST/RESIT EXAMINATIONS 2016/2017

MODULE: CA4010 - Data Warehousing and Data Mining

PROGRAMME(S):

CASE - BSc in Computer Applications (Sft.Eng.)
ECSAO - Study Abroad (Engineering and Computing)
ECSA - Study Abroad (Engineering and Computing)

YEAR OF STUDY: 4,O,X

EXAMINERS: Mark Roantree (Ph:5636)
Dr. Ian Pitt

TIME ALLOWED: 3 hours

INSTRUCTIONS: Answer 4 questions. All questions carry equal marks.

PLEASE DO NOT TURN OVER THIS PAGE UNTIL 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.

Requirements for this paper (Please mark (X) as appropriate)

☐	*Log Tables*	☐	*Thermodynamic Tables*
☐	*Graph Paper*	☐	*Actuarial Tables*
☐	*Dictionaries*	☐	*MCQ Only - Do not publish*
☐	*Statistical Tables*	☐	*Attached Answer Sheet*

QUESTION 1 ***[Total marks: 25]***

Clustering

1(a) [7 Marks]

What is a *Dendogram*, as used in data clustering?
Describe the basic algorithm for Agglomerative Hierarchical Clustering.

1(b) [8 Marks]

Using a small number of objects (A,B,C,D,E,F,G,H,I,J), show how a Dendogram is constructed and the steps involved in reaching the final cluster. Your answer should contain an illustration of the final Dendogram.

1(c) [10 Marks]

i) Using the *k*-Means algorithm, the 16 points represented by (x,y) in Table 1 have been clustered using the initial centroids C1:(2.2, 5.5), C2:(7.8, 12.2), and C3:(8.2, 12.1). Distances *d1* have been calculated for the first centroid *C1*.
Compute the missing distances for all points from centroids C2 and C3 and make your cluster assignments.

For your answer, copy the table into your exam paper and complete the missing distance values and cluster assignments.

ii) Finally, calculate the new values for all three centroids.

Table 1: Distances d1,d2,d3 from Centroids C1, C2, C3

x	y	d1	d2	d3	cluster
6.8	12.6	8.5			
0.8	9.8	4.5			
1.2	11.6	6.2			
2.8	9.6	4.1			
3.8	9.9	4.7			
4.4	6.5	2.4			
4.8	1.1	5.1			
6.0	19.9	14.9			
6.2	18.5	13.6			
7.6	17.4	13.1			
7.8	12.2	8.7			
5.8	7.7	4.2			
8.2	4.5	6.1			
8.4	6.9	6.4			
9.0	3.4	7.1			
9.6	11.1	9.3			

QUESTION 2 ***[Total marks: 25]***

Classification

age	specRx	astig	tears	C
1	1	1	1	3
1	1	1	2	2
1	1	2	1	3
1	1	2	2	1
1	2	1	1	3
1	2	1	2	2
1	2	2	1	3
1	2	2	2	1
2	1	1	1	3
2	1	1	2	2
2	1	2	1	3
2	1	2	2	1
2	2	1	1	3
2	2	1	2	2
2	2	2	1	3
2	2	2	2	3
3	1	1	1	3
3	1	1	2	3
3	1	2	1	3
3	1	2	2	1
3	2	1	1	3
3	2	1	2	2
3	2	2	1	3
3	2	2	2	3

2(a) [7 Marks]

Why is attribute selection important in the construction of decision trees?
Explain the term *Information Gain* as part of the process for attribute selection.

2(b) [18 Marks]

Use *Information Gain* to rank attribute set (age, specRx, astig, tears) for selection for the *first* branch in the decision tree, from best to worst. In other words, you are required to provide the initial classification only.

[End Question 2]

QUESTION 3 ***[Total marks: 25]***

Association Rule Mining

3(a) [10 Marks]

Describe, as a set of steps, the Apriori-gen algorithm that takes the L_{k-1} itemset and generates the new C_k itemset.
What process generates L_k from C_k?

3(b) [15 Marks]

Suppose that L_3 is the list
{{a, b, c}, {a, b, d}, {a, c, d}, {b, c, d}, {b, c, w}, {b, c, x}, {p, q, r}, {p, q, s}, {p, q, t}, {p, r, s}, {q, r, s}}
i) Which itemsets are placed in C_4 by the *join* step of the Apriori-gen algorithm? (7 marks)
ii) Which are then removed by the *prune* step? (5 marks)
iii) Will there be a L_5? Explain your answer (3 marks).

[End Question 3]

QUESTION 4 ***[Total marks: 25]***

Data Warehousing and OLAP

4(a) [6 Marks]

(i) What is meant by a *Fact Table* in Data Warehousing? In your answer, describe the role of the *measurement*.
(ii) What is the role of *Dimension Tables*? Describe the relationship between Dimension and Fact tables and explain the functionality provided by Dimension tables.

4(b) [13 Marks]

Assume you are a data analyst for a chain of gymnasiums, with fitness area, classes, and swimming pool. Use a diagram to describe the schema with an appropriate Fact Table and at least three Dimension Tables. In your diagram, be sure to highlight those attributes that are keys.
Also, provide a brief discussion on each table.

4(c) [6 Marks]

(i) What is meant by a Roll-up operation?
(ii) Using your diagram in part 4b), provide 2 examples of Roll-up operations using 2 separate dimensions.
Use your fact table, and incorporate sample values to demonstrate the rollup computations in each case.

[End Question 4]

QUESTION 5 ***[Total marks: 25]***

Cube Computation

Assume you are required to create a data cube for analysis for the FlyHigh airline containing Passenger demographics (all the data entered when booking a flight), Flight data, and date with ticket price as the measure. Its purpose is to analyse the price by bookings, destination and season (dates).

Dest	Seating	Month	count()
*	*	*	200
London	*	*	
Paris	*	*	
Munich	*	*	
*	N	*	
*	R	*	
*	P	*	
*	*	J	50
*	*	F	50
*	*	M	50
*	*	A	50

5(a) [6 Marks]

Draw the lattice for the data Cube provide a separate list of all cuboids.

5(b) [11 Marks]

Assume we are looking at a small dataset with 3 city destinations (London, Paris and Munich), 3 passenger seating requirements (None, Reserved seat, Priority boarding) and date as (Jan,Feb,Mar,Apr).
In total we have 200 flights in the sample.
A partially complete sample table is shown in the table above. Use your own sample data to complete the table, taking care to ensure that **all aggregations are correct**.

5(c) [8 Marks]

Define an *ancestor* cell. As part of your answer, define a *descendant* cell in relation to an ancestor cell.
Use the sample table from part (b) of this question to provide examples which clearly illustrate the difference between each type of cell.

[End Question 5]

[END OF EXAM]