University of Ruhuna - Faculty of Technology

Bachelor of Information & Communication Technology Honours Degree
Level 3 (Semester II) Examination – November/December 2023
Academic Year 2021/2022

Time: 2.5 hours

Course Unit: ICT3222 Data Mining and Business Intelligence

Answer all **four (04)** questions

1)

a) Data warehousing is a complete solution for heterogeneous information sources. List down **three (03)** problems associated with heterogeneous information sources.

b) The traditional research approach to integrate various sources is the query driven approach. Briefly explain **two (02)** disadvantages of Query driven approach.

c) What is a Data warehouse? Explain using a suitable example.

d) Assume that, you are working as a data design specialist in a reputed software company. Your company has received a requirement to develop a data warehouse to facilitate Human Resource and Payroll management activities of a client organization. They need to manage employee attendance, payroll transactions, employee trainings, employee performance and employee benefits. In addition to that, client organization consists with various departments in main three locations. Top management expects to have ad-hoc reports by getting the use of Data warehouse for their decision making.

i) Identify **five (05)** facts and **five (05)** dimension tables that can be included in data warehouse according to the given scenario.

ii) What is the suitable conceptual schema to develop above mentioned Data warehouse? Justify your answer by providing valid reasons.

iii) Write **two (02)** hierarchical dimensions included in designed data warehouse.

iv) Kimball proposed a database design methodology with nine steps for data warehouse. Describe the following steps in Kimball methodology.

- Storing pre calculations in the fact table
- Tracking slowly changing dimensions

2) A well-established tea plantation company wants to design a data warehouse to manage their plantation, harvesting and trading records. The company owns number of tea estates around the country and they transport collected harvest daily to tea factories established in specific areas. Received tea leaves are separated based on the size, color and quality. Grading is done based on the quality parameters and market value is determined. Related

to the plantation activities, they need to record rainfall information for their various analytical purposes. In order to handle their sales and distribution activities, they use various channels to distribute their tea products. Online platforms, super markets and other specific stores are among those channels. After extracting, cleaning and transforming, data must be loaded into the warehouse. Typically use batch load utilities in the data loading process.

a) Define the term "data mining" and describe the importance of data mining in plantation company context.

b) State **three (03)** data loading issues that can be occurred in data warehousing.

c) Is the inclusion of data refreshing necessary for the data warehouse of a plantation company? Justify your answer by providing **two (02)** valid reasons.

d) Yearly tea leave harvest of three main estates of the plantation company is represented by below ordered set of data. The identified dimensions of the data cube are tea leaves type, year and estate name. The measure is the weight in metric tons.

(Black tea, 2022, Estate 1, 2389)	(Oolong tea, 2022, Estate 3, 2320)
(White tea, 2021, Estate 2, 5234)	(Black tea, 2020, Estate 1, 4200)
(Oolong tea, 2020, Estate 3, 2345)	(White tea, 2022, Estate 2, 4500)
(White tea, 2020, Estate 3, 5200)	(Black tea, 2021, Estate 3, 3400)
(Oolong tea, 2021, Estate 2, 2200)	(Oolong tea, 2020, Estate 1, 2600)
(White tea, 2021, Estate 1, 1700)	(Black tea, 2022, Estate 2, 1800)

i) Calculate all the necessary aggregated cuboids and clearly show the data cube representation of above information.

ii) Calculate the following aggregate cuboid values.

a. (*, *, Estate 2)

b. (White tea, *, Estate 3)

3)

a) Explain the importance of data preprocessing in data warehousing using an example.

b) State **five (05)** major tasks in data preprocessing and briefly describe **two (02)** of them.

c) What is noisy data? List **three (03)** possible ways that can be used to handle noisy data.

d) The age distribution of employees within a chosen department of an organization is provided below.

Age: 45, 56, 62, 43, 39, 28, 31, 44, 56, 61, 53, 32, 28, 26, 39, 41, 53, 60, 40, 52, 34

Partition the age values given above into 3 bins using:

i) Equal depth and smooth by bin mean and bin boundaries.

ii) Equal width and smooth by bin mean and bin boundaries.

e) Min-max normalization and Z-score normalization are commonly used data normalization strategies in data transformation.

i) The income attribute has minimum and maximum values specified as Rs. 12,000 and Rs. 92,000 respectively. Calculate the transformed value for an income of Rs. 84,000 using minmax normalization within the range [0:0, 0:1]

ii) The income attribute has Rs. 42,000 and Rs. 54,000 as mean and standard deviation respectively. Calculate the transformed value for an income of Rs. 92,000 using z-score normalization.

4)

a) Consider the following transactions of a computer equipment shop presented in the following Table Q4a.

T_ID	Item List
T01	Keyboard, Mouse, Speaker, Hard drive, RAM, Scanner
T02	Mouse, RAM
T03	Mouse, Keyboard, Speaker, Hard drive
T04	RAM, Keyboard, Speaker, Mouse
T05	Mouse, Scanner
T06	Mouse, Keyboard, Speaker, Projector
T07	Speaker, Scanner
T08	Speaker, Keyboard, RAM, Hard drive, Mouse
T09	Mouse, RAM, Keyboard
T10	Mouse, RAM, Keyboard, Scanner

Table Q4a

i) Association rule was given as "*Mouse, RAM → Speaker*" State the association rule in your own words.

ii) Find the *support* and *confidence* for the associations given below.

(a) Mouse → RAM

(b) Mouse, RAM, Speaker → Keyboard

iii) Assuming a minimum level of support min_sup = 4 and a minimum level of confidence min_conf = 80%. Find the most frequent itemset using the Apriori algorithm. For each iteration show the candidate and acceptable frequent itemset.

b) You are tasked with assessing the feasibility of issuing credit cards to a randomly chosen subset of customers in a bank. The data collected from individual customers is presented in Table Q4b.

Income	Credit Rating	Marital Status	Government or Private Sector	Issue a Credit card or not (Class Label)
High	Good	Single	Government	Yes
Medium	Poor	Married	Private	No
Low	Excellent	Married	Private	Yes
Medium	Poor	Single	Government	No
Medium	Good	Single	Private	Yes
High	Poor	Married	Government	Yes
Low	Good	Married	Government	No
High	Good	Single	Private	Yes
High	Poor	Married	Government	Yes
Medium	Excellent	Single	Government	Yes
Low	Excellent	Single	Private	Yes
Medium	Good	Married	Private	No
Medium	Excellent	Married	Private	Yes
High	Poor	Single	Private	Yes
High	Good	Married	Government	Yes

Table Q4b

i) Calculate the overall Entropy before splitting.

ii) Calculate the overall Entropy after splitting each attribute.

iii) At which attribute should the decision tree split first? Explain the reason for your selection.

c) Briefly describe **two (02)** strengths and **two (02)** weaknesses of Decision Trees.

Log Table

$\log_2(x)$										
x	**0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0	inf.	-6.64	-5.64	-5.06	-4.64	-4.32	-4.06	-3.84	-3.64	-3.47
0.1	-3.32	-3.18	-3.06	-2.94	-2.84	-2.74	-2.64	-2.56	-2.47	-2.4
0.2	-2.32	-2.25	-2.18	-2.12	-2.06	-2	-1.94	-1.89	-1.84	-1.79
0.3	-1.74	-1.69	-1.64	-1.6	-1.56	-1.51	-1.47	-1.43	-1.4	-1.36
0.4	-1.32	-1.29	-1.25	-1.22	-1.18	-1.15	-1.12	-1.09	-1.06	-1.03
0.5	-1	-0.97	-0.94	-0.92	-0.89	-0.86	-0.84	-0.81	-0.79	-0.76
0.6	-0.74	-0.71	-0.69	-0.67	-0.64	-0.62	-0.6	-0.58	-0.56	-0.54
0.7	-0.51	-0.49	-0.47	-0.45	-0.43	-0.42	-0.4	-0.38	-0.36	-0.34
0.8	-0.32	-0.3	-0.29	-0.27	-0.25	-0.23	-0.22	-0.2	-0.18	-0.17
0.9	-0.15	-0.14	-0.12	-0.1	-0.09	-0.07	-0.06	-0.04	-0.03	-0.01

.... **End of the Paper**............