University of Ruhuna – Faculty of Technology

Bachelor of Information & Communication Technology Honours Degree
Level 3 (Semester II) Examination – November/December 2025
Academic Year: 2023/2024

Course Unit: ICT3233 – Data Science and Analytics (Written)

Answer **all four (04)** questions **Duration: 2.5 hours**

1)

a) A data warehouse (DW) provides a permanent solution for number of issues associated with data analysis. What is the main problem addressed by a data warehouse as a solution for data analysis?

[10 Marks]

b) "A DW is a subject-oriented, integrated, time-varying, non-volatile collection of data that is used primarily in organizational decision making." Explain what is meant by non-volatile collection by using a suitable example.

[20 Marks]

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

[20 Marks]

d) A large hospital chain operates multiple branches across the country. Each branch uses its own operational system to record patient admissions, diagnoses, lab tests, treatments, surgeries, billing details, and doctor availability. These systems work well for day-to-day activities but do not provide a unified view for long-term analysis.
To improve strategic decision-making, the hospital builds a Data Warehouse that integrates data from all branches.
Once implemented, the DW allows hospital administrators to:

- Analyze patient admission trends (e.g., peak months for dengue or heart diseases).
- Identify high-demand departments (e.g., emergency, cardiology, pediatrics).
- Monitor performance metrics such as average patient waiting time, surgery success rates, and bed occupancy rates.
- Evaluate doctors' workload across branches to support staffing decisions.
- Predict medicine and equipment demand based on historical usage patterns.

i) List three (03) Online analytical processing (OLAP) rules that can be considered in designing DW according to the business scenario of the hospital chain.

[15 Marks]

ii) The development team has decided to follow OLAP approach to design the data warehouse. Do you agree with the suggested approach? Give two (02) valid reasons to support your answer.

[15 Marks]

iii) The development team wants to select a reporting architecture to view information stored in the data warehouse. Describe specific characteristics when selecting reporting architecture.

[20 Marks]

2) Serendib Insurance Corporation (SIC) is one of the largest insurance providers in Sri Lanka, offering a wide range of products such as life insurance, health insurance, motor insurance, property insurance, and corporate insurance packages. The company operates through multiple branches across the island, each using different operational systems including policy management systems, claims processing systems, finance and billing applications, customer relationship management (CRM), fraud detection systems, and agent performance tracking tools.
However, the company faces major challenge due to fragmented data stored in sperate systems. As a result, top management and analysts struggle to answer critical strategic and operational questions. To resolve these issues, SIC initiates a large-scale **Data Warehouse (DW) project** aimed at integrating data from all operational systems into a single analytical platform.

a) List three (03) basic architectures used for constructing a data warehouse.

[15 Marks]

b) What is the most suitable architecture for developing the data warehouse of Serendib Insurance Corporation (SIC). Give valid reasons to support your answer.

[20 Marks]

c) You are part of a team designing a data warehouse for the insurance firm aiming to analyze sales trends, customer behaviors and claim history.

i) What is the most suitable schema for developing this data warehouse? Explain your answer.

[15 Marks]

ii) List five (05) dimension tables which can be included in the data warehouse design.

[15 Marks]

iii) Suggest a data visualization tool that the company can use after implementing the data warehouse. Briefly explain two (02) reasons for your recommendation

[15 Marks]

iv) Discuss the importance of data refreshing in the Serendib Insurance Corporation (SIC) context.

[20 Marks]

3)

a) What is Data Mining? Explain in detail using a real-world example.

[15 Marks]

b) Data monitoring is an essential step in Data warehouse development process. This process detects changes to an information source that are of interest to the warehouse. List three (03) monitoring techniques used in the data warehouse.

[15 Marks]

c) Data cleaning is an essential step involved in the data warehouse development process. Discuss the importance of data cleaning in data warehouse development.

[20 Marks]

d) The following figures show the daily harvest of a tea plantation company in Kg s within 21 working days. Partition the harvested values given below into 3 bins using:

56,34,61,63,39,43,52,65,73,45,40,39,72,56,57,56,34,43,39,40,52

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

[20 Marks]

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

[15 Marks]

e) The sales revenue attribute of tea leaves has minimum and maximum values specified as Rs. 10,000 and Rs. 212,000 respectively. Calculate the transformed value for sales revenue of Rs. 88,000 using minmax normalization within the range [0:0, 0:1]

[15 Marks]

4)

a) Consider the following transactions of a database presented in the following Table Q4a. Assuming a minimum level of support min_sup = 3 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.

[25 Marks]

T_ID	Item List
T01	K, A, D, B
T02	D, A, C, E, B
T03	C, A, B, E
T04	B, A, D

Table Q4a

b) You are assigned to assess the feasibility of issuing credit card for a chosen subset of customers in a private bank. The data collected from credit card handling center is presented in Table Q4b.

[50 Marks]

Income (Rs)	Age (Years)	Have taken any Loans or Not	Job Field	Issuing a credit card or not **(Class Label)**
Below 50K	Above 35	No	Non-IT	Yes
50K-100K	Below 30	Yes	IT	No
50K-100K	30 – 35	Yes	Non-IT	Yes
Above 100K	Below 30	No	IT	Yes
Below 50K	30 – 35	No	IT	No
Above 100K	30 – 35	Yes	Non-IT	Yes
Above 100K	Below 30	No	IT	Yes
50K-100K	Above 35	Yes	IT	Yes
Below 50K	Above 35	Yes	Non-IT	Yes
50K-100K	30 – 35	No	Non-IT	No
50K-100K	Above 35	No	Non-IT	Yes
Above 100K	Below 30	Yes	Non-IT	Yes
Above 100K	30 – 35	No	IT	Yes
50K-100K	Below 30	No	Non-IT	No
Above 100K	30 – 35	Yes	IT	Yes

Table Q4b

Note: Use the log table given at the end of the paper for your calculations.

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) Use the Naïve Bayes Classifier to predict the class for the following dataset given in Table Q4c.

[25 Marks]

Income (Rs)	Age (Years)	Have taken any Loans or Not	Job Field	Issuing a credit card or not **(Class Label)**
Below 50K	Above 35	No	IT	?
Above 100K	30 – 35	Yes	Non-IT	?

Table Q4c

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...../.....