Strathmore UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS

END OF SEMESTER EXAMINATION

DSA 8102: DATA MINING, DATA STORAGE & RETRIEVAL

DATE: 8th September 2021 Time: 2 1/2 Hours

Instructions

1. This examination consists of **FOUR** questions.
2. Answer any **THREE** questions.

Question One

Carrefour was launched in Kenya in 1999 and today it operates five hypermarkets and five supermarkets, employing over 1800 colleagues.

Carrefour operates different store formats, to meet the growing needs of its diversified customer base. In line with the brand's commitment to provide the widest range of quality products and value for money, Carrefour offers an unrivalled choice of more than 30,000 food and non-food products, and an exemplary customer experience to create great moments for everyone every day.

We always strive to provide the best quality and most diverse selection of household goods available in Kenya. Our value packs and combination discount offer mean that we can offer these products at even lower costs, keeping your household essentials at unbeatable prices.

Source: Carrefour. 2021. Accelerator Title. [online] Available at: https://www.carrefour.ke/mafken/en/carrefourandsociety [Accessed 2 July 2021].

As a data miner, you were tasked with finding regularities in the shopping behaviour of customers. Table 1 shows the lists of transactions from seven different customers. Answer the questions that follow:

Table 1: Transactions of items

TID	Items
1	Tofu, Kimchi, Fish, Eggs
2	Yoghurt, Chicken, Lamb
3	Bread, Yoghurt, Lamb
4	Bread, Yoghurt, Chicken, Tofu
5	Chicken, Bread, Lamb, Eggs
6	Fish, Bread, Eggs, Kimchi
7	Bread, Yoghurt, Lamb

1. What is the support and support count of the itemset marked in red?
2. What is the confidence of the itemset marked in red?
3. What is the lift of the itemset marked in red?
4. Given the result of the lift, what is your conclusion?

[Sub total: 20 Marks]

Question Two [20 Marks]

In data warehouse technology, a multi-dimensional view can be implemented by a relational database technique (ROLAP), by a multi-dimensional database technique (MOLAP), and by a hybrid database technique (HOLAP).

1. Briefly describe each implementation technique.
2. For each schema of the multi-dimensional data models, and with a suitable example, explain the following:
   a) Star schema
   b) Snowflake schema
   c) Fact constellation schema

[Sub total: 20 Marks]

Question Three [20 Marks]

a) As a data analyst, discuss the data stream features that make it difficult to process using traditional DBMSs. **(10 Marks)**

b) Using an example, discuss the significance of (preliminary) exploratory data analysis. **(10 Marks)**

Question Four [20 Marks]

a) Explain at least 5 steps of data pre-processing**. (5 Marks)**

b) Using examples, describe at 3 multivariate techniques for feature selection. **(15 Marks)**