**Birla Institute of Technology & Science, Pilani**

**Work Integrated Learning Programmes Division**

**Second Semester 2023-2024**

**End-Semester Test**

**(EC-3 Regular)**

Course No. : DSECLZG529/AIMLCZG529

Course Title : Data Management for Machine Learning

Nature of Exam : Open Book

Weightage : 40%

No. of Pages = 4

# No. of Questions = 7

Duration : 2 Hours

Date of Exam : ~~06/03/2021 or 19/03/2021 (FN/AN)~~

Note to Students:

1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q1. Suppose that a data warehouse consists of three dimensions: **time, doctor,** and **patient**, and the two measures **count** and **charge**, where charge is the fee that a doctor charges a patient for a visit. **[1 + 2 + 2.5 + 1.5 = 7]**

1. Enumerate two classes of schemas popularly used for modelling data warehouses.
2. Draw a schema diagram for the above data warehouse using one of the schema classes listed in (a).
3. Starting with the base cuboid [**day, doctor, patient**], what specific OLAP operations should be performed to list the total fee collected by each doctor in 2004?
4. To obtain the same list, write an SQL query assuming the data are stored in a relational database with the schema fee (**day, month, year, doctor, hospital, patient, count, charge**).

Q2. You are working on a machine learning project that predicts monthly sales for an e-commerce website. You have trained a sales prediction model using data from the year's first six months (January to June). The model performs well during initial testing.

However, you notice a significant drop in prediction accuracy when you deploy the model in the production environment and start using it to make predictions for the next three months (July to September). You suspect that data drift may be the cause of this decline in performance.

To investigate data drift, you collect the following data for both the training data (January to June) and the production data (July to September):

Training Data (January to June):

* Mean Monthly Sales: $50,000
* Standard Deviation of Monthly Sales: $7,000

Production Data (July to September):

* Mean Monthly Sales: $45,000
* Standard Deviation of Monthly Sales: $8,500

Calculate the following data drift metrics and provide your analysis: **[2 +2 + 2 = 6]**

1. Percentage Change in Mean Monthly Sales between Training and Production Data.
2. Percentage Change in Standard Deviation of Monthly Sales between Training and Production Data.

Q3. You are tasked with designing a data pipeline to process and analyse log data from a website. The logs contain user interactions, including page views, clicks, and demographics. The pipeline consists of three stages: data ingestion, transformation, and analysis.

* Data Ingestion: The raw log data is ingested into the pipeline at an average of 1,000 log entries per second.
* Data Transformation: During the transformation stage, various data processing tasks are performed, including data cleaning, parsing, and feature extraction. The transformation stage processes data at 800 log entries per second.
* Data Analysis: In the analysis stage, machine learning models are applied to the transformed data to predict user behaviour. The analysis stage processes data at an average of 500 log entries per second.

a) Calculate the data throughput for each stage of the data pipeline in log entries per minute. **[3]**

b) Determine the bottleneck stage in the data pipeline based on the calculated throughputs. **[1]**

Q4. Imagine you are a data scientist working on a machine learning project for a healthcare organisation. Your task is to build a predictive model to identify patients at high risk of developing a specific medical condition based on their health records.

In your machine learning project, you trained three models: Model A, Model B, and Model C, each with various hyperparameters and feature engineering techniques. As part of your model metadata, you have recorded the model's architecture, hyperparameters, training data, evaluation metrics, and the date of each model's creation. Additionally, you have documented any noteworthy observations or lessons learned during the modelling process.

You are reviewing the model metadata for Models A, B, and C, and you notice that Model C consistently outperforms the other models in terms of accuracy and recall on the validation dataset. However, Model C is also significantly larger regarding memory usage than Models A and B. Given this information:

a) Why is it important to keep track of model metadata in your machine learning project? **[2]**

b) What are the potential advantages of using Model C despite its higher memory usage? **[2]**

c) How can you ensure that the large memory usage of Model C is smooth in a production environment? **[1]**

Q5. You are tasked with designing a data architecture for a large e-commerce platform. The platform handles many daily customer transactions, product updates, and user interactions. The architecture must efficiently support real-time analytics, reporting, and data storage. You decide to use a data warehouse for this purpose.

Here are some critical metrics for the platform:

* 10,000,000 customer transactions per day
* 1,000,000 product updates per day
* 5,000,000 user interactions per day

Your data architecture needs to handle and process this data efficiently. Design a data architecture that includes the following components and provide an estimate of the required storage capacity: **[4]**

* Data Ingestion Layer
* Data Storage Layer
* Data Processing Layer
* Data Analytics and Reporting Layer

Q6. Consider a dataset containing information about customer orders for an e-commerce website. Here are 20 records with various data quality issues. Identify at least 5 data quality issues and compute data quality metrics to help illustrate the importance of data quality assessment. **[10]**



Q7. A healthcare organisation is sharing medical research data with a research partner while ensuring privacy protection using a privacy-preserving technique called "k-anonymity." The dataset contains information about patients' medical conditions and their ages. The organisation wants to disclose aggregate statistics about patient ages while protecting individual privacy.

The organisation chooses to achieve 3-anonymity, meaning that there are at least three patients with the same age and medical condition for any combination of age and medical condition. The dataset contains the following information:

* Patient A: Age 45, Medical Condition X
* Patient B: Age 30, Medical Condition Y
* Patient C: Age 45, Medical Condition Z
* Patient D: Age 35, Medical Condition X
* Patient E: Age 50, Medical Condition Y
* Patient F: Age 45, Medical Condition X
* Patient G: Age 30, Medical Condition Z
* Patient H: Age 30, Medical Condition X

a) Calculate the transformed dataset that satisfies the 3-anonymity requirement. **[3]**

b) Explain how the transformed dataset ensures privacy protection for individual patients. **[1]**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*