VIT
Vellore Institute of Technology

Final Assessment Test – April 2025

Course: **BCSE206L - Foundations of Data Science**

Class NBR(s): **1985/1997/2002/2013/2023**

Slot: **C1+TC1**

Time: **Three Hours**

Max. Marks: **100**

- **KEEPING MOBILE PHONE/ANY ELECTRONIC GADGETS, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE**
- **DON'T WRITE ANYTHING ON THE QUESTION PAPER**

Answer ALL Questions

(10 X 10 = 100 Marks)

1. A retail company wants to improve its sales forecasting to optimize inventory management and reduce stock shortages. As a Data Scientist, explain how you would approach this problem using the Data Science process.

2. Analyze and solve the following SQL-based problems using the given 'Orders' and 'Customers' tables:

Order ID	Customer ID	Order Date	Total Amount
1	101	2024-01-05	500
2	102	2024-01-10	700
3	101	2024-02-15	300
4	103	2024-02-20	1200
5	104	2024-03-01	900
6	101	2024-03-10	400

Customer ID	Customer Name	City
101	Alice	New York
102	Bob	Chicago
103	Charlie	Boston
104	David	Chicago

a) Write an SQL query to display the customer names of those who placed at least one order in February 2024

b) Compute the total amount spent by each customer in Q1 2024 (January - March). Display the customer name along with the total amount spent.

c) Using a dense rank, assign a rank to customers based on their total spending in Q1 2024, with the highest spender ranked first. Display the customer name, total amount spent, and rank.

d) Find the average order value (total amount/order count) for each customer and compare it with the overall average order value across all orders. Display only those customers whose average order value is greater than the overall average.

e) Write a query to compute the running total of order amounts for each customer, ordered by order date. The output should include customer name, order date, order amount, and running total.

3. A smart healthcare system **is being developed to manage** patient records, medical history, hospital networks, and doctor-patient relationships. The **system needs to choose between** Document Databases, Wide-Column Databases, and Graph Databases **for different aspects of its data management.**

a) **Define** Document Databases, Wide-Column Databases, and Graph Databases

b) **The hospital needs to store** detailed patient records, **including medical history, prescriptions, and test reports, in a** flexible and hierarchical format. **Which NoSQL database should they use and why?**

c) **The hospital wants to track** patient visits and treatment history over time **and quickly generate** reports on disease trends. **Which NoSQL database is best suited for this, and how should the data be structured?**

d) **The system must manage** doctor-patient relationships **and find the** shortest referral paths between specialists **in a hospital network. Which NoSQL database should be used and why?**

e) **The hospital plans to implement** AI-based disease prediction **by analysing patient medical histories and common treatment patterns. Which one of these are more suitable. Justify your answer.**

4. A Production company is set to release the highly anticipated superstar Rajinikanth-starrer "Coolie" in 2025.The production company aims to maximize the movie's box office success by leveraging data analytics to inform their marketing and release strategies.

They have access to historical data from previous Rajinikanth blockbusters, audience demographics, ticket sales, streaming engagement, and social media trends.

Your analytics team is tasked with applying the **Data Analytics Lifecycle** to derive actionable insights and recommend strategies for "Coolie."

5. A media analytics company is evaluating the importance of words in movie reviews for **Rajinikanth's upcoming movie "Coolie"** to identify key themes and audience sentiments. Use **(TF-IDF)** to rank words based on their significance in multiple reviews.

Given the following five movie reviews:

Review 1: "Rajinikanth delivers a power-packed performance in Coolie."

Review 2: "The action sequences in Coolie are mind-blowing, and Rajinikanth shines."

Review 3: "Coolie is an action-packed blockbuster with Rajinikanth at his best."

Review 4: "The direction and screenplay of Coolie elevate Rajinikanth's performance."

Review 5: "Rajinikanth and the cast make Coolie a must-watch action film."

Stopwords: with, and, at, is, in, of, on, the, to, a, an, at.

a) List and describe the key components of **TF-IDF**. What does each component measure?

b) Calculate the **TF value** for the word **"Rajinikanth"** in each review.
Compute the **TF value** for the word **"Coolie"** in each review. Explain the significance of these values.

c) Compute the **IDF value** for the word **"Rajinikanth"** across all five reviews

d) Compute the **TF-IDF score** for "Rajinikanth" in each review

e) Compare the **TF-IDF scores** of the words **"Rajinikanth"** and **"Coolie"** across all reviews. Which word has a **higher TF-IDF score**, and what does it indicate?

Based on the TF-IDF results, how would you use this technique to improve **movie marketing strategies**? Provide two recommendations.

6.a) **Healthcare start-up** is developing an **AI-powered virtual assistant** to assist doctors and patients by automating medical consultations, summarizing patient records, and extracting key insights from medical literature. The assistant will use **Natural Language Processing (NLP)** to understand patient queries, recommend treatments, and generate reports.

Design an NLP-based solution using various **NLP components, stages, and applications.**

OR

6.b) A marketing team wants to segment customers based on their annual income and spending score to create targeted campaigns. They decide to use K-Means Clustering to group customers into clusters.

Customer	Annual Income ($000s)	Spending Score
A	15	39
B	20	81
C	26	6
D	30	77
E	40	20
F	45	50
G	50	90
H	55	10

set K = 2 ,choose centroids (15, 39), (30, 77)

i. Compute the Euclidean Distance of each customer from both centroids, Assign each customer to the nearest centroid.

ii. Compute the new centroids after the first iteration by finding the mean coordinates of the assigned clusters.

iii. Explain how many iterations K-Means need and how you determine when to stop clustering.

7. Given the following 5 points in a 2D space, use the Complete Linkage Algorithm to perform hierarchical clustering. Show the step-by-step process including distance calculations, merging steps, and the final dendrogram using Eucledian distance measure.

Point	X	Y
A	2	3
B	5	4
C	9	6
D	4	7
E	8	1

8.a) An e-commerce website wants to recommend similar products based on customer reviews. The website uses Cosine Similarity to compare text reviews and find similar opinions.

Consider the following two customer reviews for a smartphone:

Review 1: "The phone has an excellent camera and a long-lasting battery."

Review 2: "A great camera phone with amazing battery life."

- Explain why Cosine Similarity is useful for comparing text reviews instead of using simple word matching.
- Compute the dot product of Review 1 and Review 2.
- Compute the Cosine Similarity between the two reviews.
- If the similarity score is above 0.75, should the website recommend both phones as similar to the user? Justify your answer.

Stopwords: the, has, an, and, a, with, is.

OR

8.b) i. What is the difference between a **NumPy array, Pandas** and **DataFrame**? Provide an example of when you would use each. **[2.5]**

ii. Explain how **missing values** are handled in Pandas. Provide an example of how to **fill missing values** in a DataFrame **[2.5]**

iii. Given the NumPy array: [3]

```
import numpy as np
arr = np.array([[10, 20, 30], [40, 50, 60], [70, 80, 90]])
```

Write Python code to:

Convert this **NumPy array to a Pandas Data Frame.**

Rename the columns as **"A", "B", "C".**

Extract only the **second column ("B")** from the Data Frame.

iv. Given the following Pandas data Frame [2]

```
import pandas as pd
data = {'Name': ['Rajini', 'Kamal', 'Saro', 'Ritchie'],
'Age': [25, 30, 35, None],
'Salary': [50000, 60000, 70000, 80000]}
df = pd.Data Frame(data)
```

Write Pandas commands to:

Fill missing values in the **Age** column with the **mean age.**

Add a new column called **"Bonus"**, which is **10% of Salary.**

9. A robotics company is designing an autonomous drone stabilization system. The drone's motion is represented by a state transition matrix, which determines how it responds to external forces.

The engineers need to analyse the eigenvalues and eigenvectors of this matrix to understand the system's stability and behavior over time.

i. The drone's state transition is modelled by the matrix: A = [[4, -2], [1, 1]] Compute the eigenvalues and eigen vectors.

ii. Based on the computed eigenvalues. Explain how they influence the stability of the system?

Calculate through GNU octave code.

10. a) List three **best practices** for designing a Tableau Dashboard. [4]

b) What is the difference between **Dimensions and Measures** in Tableau? [3]
Provide an example.

c) Explain how **Descriptive Statistics** (Mean, Median, Variance) are used in [3]
Tableau for data analysis.

⇔⇔⇔ **Y/D/TY** ⇔⇔⇔