REG.NO.:

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
WINTER SEMESTER 2024-2025

SLOT: C2+TC2

Programme Name & Branch : **BTech. CSE, BDS**
Course Code and Course Name : **BCSE206L, Foundations of Data Science**
Faculty Name(s) : **Common for all sections**
Class Number(s) : **Common for all sections**
Date of Examination : **18/03/2025**
Exam Duration : **90 minutes** **Maximum Marks: 50**

General instruction(s):

- Answer All Questions
- **Note: Step by step solution is expected.**

CO3: Apply advanced tools to work on dimensionality reduction and mathematical operations.

CO4: Handle various types of data and visualize them using through programming for knowledge representation.

CO5: Demonstrate numerous open source data science tools to solve real-world problems through industrial case studies.

Q. No	Question	M	CO	BL

1.	A hospital wants to reduce emergency room (ER) wait times by analysing patient arrival patterns, staff availability, and treatment durations. How can the Data Analytics Lifecycle be applied to optimize ER wait times? Describe each stage and its significance in this context.
	
	KEY:
	
	Phase 1: Discovery
	
	Objective: Define the problem and project goals.
	
	• **Key Challenge:** Long ER wait times leading to patient dissatisfaction and overcrowding.
	• **Goals:**
	  ○ Identify peak hours of patient influx.
	  ○ Optimize staff allocation and resource utilization.
	  ○ Reduce average wait time by X%.
	• **Stakeholders:** Hospital administrators, doctors, nurses, and IT teams.
	
	Significance: Ensures alignment with hospital objectives and sets measurable success criteria.
	
	Phase 2: Data Preparation
		10	3	2

Objective: Collect, clean, and integrate data.

Data Sources:

- **Patient Records:** Arrival time, triage level, treatment time.
- **Staff Schedules:** Number of doctors/nurses on shift.
- **Historical Data:** Peak admission times, average treatment duration.

Tasks:

- Handle missing or inconsistent data.
- Format timestamps for proper time-series analysis.
- Merge different data sources into a unified dataset.

Significance: Ensures high-quality, structured data for accurate analysis.

Data Exploration (EDA - Exploratory Data Analysis)

Objective: Identify trends, patterns, and outliers.

- **Analyze patient arrival patterns** (e.g., peak hours, seasonal trends).
- **Visualize resource utilization** (e.g., number of available beds vs. patient inflow).
- **Identify bottlenecks** in the treatment process.

Significance: Helps uncover critical insights to guide model selection.

Phase 3: Model Planning

Objective: defines the approach and techniques used to analyze ER wait times and predict improvements.

- **Defining Data Features**
- **Selecting Analytical Techniques**
- **Developing a Data Sampling Strategy**
- **Establishing Success Metrics**

Significance: By forecasting peak hours, hospitals can dynamically adjust staff allocation to meet patient demand. Understanding treatment durations and arrival patterns enables efficient resource utilization, ensuring that medical staff and equipment are available when needed.

Phase 4: Model Building

	Objective: Develop predictive models for ER optimization. Potential **Machine Learning Models:** • Forecast patient inflow at different times- ARIMA, LSTM • Predict patient wait times based on staff availability- Linear Regression, Random Forest Regression • Predict High-Risk Patients- Logistic Regression, Random Forest **Significance:** Enables data-driven decision-making for efficient ER management. **Phase 5: Deployment & Monitoring** **Objective: Implement and integrate the solution.** • **Real-time Dashboard:** Displays predicted peak hours & staff recommendations. • **Automated Alerts:** Notify hospital administrators about expected overcrowding. • **Continuous Monitoring:** Track model performance and adjust as needed. **Significance:** Transforms insights into **real-world improvements** in ER efficiency.			
2.	Suppose we have the following documents: • **Doc1:** "machine learning is powerful." • **Doc2:** "deep learning is a subset of machine learning." • **Doc3:** "artificial intelligence includes machine learning and deep learning." A user searches for **"machine learning applications."** Given a query and a set of three documents, remove stop words and apply the Vector Space Model using Term Frequency - Inverted Document Frequency method to order the documents based on cosine similarity. **KEY:** We remove stopwords and tokenize: • **Doc1:** ["machine", "learning", "powerful"] • **Doc2:** ["deep", "learning", "subset", "machine", "learning"] • **Doc3:** ["artificial", "intelligence", "includes", "machine", "learning", "deep", "learning"]	10	4	3

- **Query:** ["machine", "learning", "applications"]

Compute TF

Term	Query (3 terms)	Doc1 (3 terms)	Doc2 (5 terms)	Doc3 (7 terms)
machine	1/3 = 0.3	1/3 = 0.3	1/5 = 0.2	1/7 = 0.14
learning	1/3 = 0.3	1/3 = 0.3	2/5 = 0.4	2/7 = 0.28
applications	1/3 = 0.3	0	0	0
powerful	0	1/3 = 0.3	0	0
deep	0	0	1/5 = 0.2	1/7 = 0.14
subset	0	0	1/5 = 0.2	0
artificial	0	0	0	1/7 = 0.14
intelligence	0	0	0	1/7 = 0.14
includes	0	0	0	1/7 = 0.14

Compute TF-IDF Weights

Now, **TF-IDF = TF × IDF** for each term.

Term	Doc1 TF-IDF	Doc2 TF-IDF	Doc3 TF-IDF	Query TF-IDF
machine	0.3×0=0	0.2×0=0	0.14×0=0	0.3×0=0
learning	0.3×0=0	0.4×0=0	0.28×0=0	0.3×0=0
applications	0	0	0	Ignored
powerful	0.3×0.477=0.158	0	0	0
deep	0	0.2×0.176=0.034	0.14×0.176=0.02	0
subset	0	0.2×0.477=0.09	0	0
artificial	0	0	0.14×0.477=0.067	0
intelligence	0	0	0.14×0.477=0.067	0
includes	0	0	0.14×0.477=0.067	0

	Cosine similarity score Since all **TF-IDF values for query terms are 0**, its magnitude is **0**, meaning the cosine similarity will be **0** for all documents.			
3.	A university wants to develop an AI-powered chatbot to assist students with course registration. The chatbot must understand student queries like: "**I want to register for Data Science. What are the prerequisites? and how can I check the available slots?**" How can Natural Language Processing (NLP) components such as Tokenization, Stopword removal, PoS tagging, Stemming and Lemmatization be applied to improve the chatbot's ability to process and respond to student queries effectively? Provide explanations and significance of each component. **KEY:** **Tokenization** Splits the text into smaller meaningful units (tokens), like words or sentences, for easier analysis. ["I", "want", "to", "register", "for", "Data", "Science", ".", "What", "are", "the", "prerequisites", "?", "And", "how", "can", "I", "check", "the", "available", "slots", "?"] **Stopword Removal** Removes commonly used words (e.g., "the," "is," "and," "to") that do not contribute much meaning. ["register", "Data", "Science", "prerequisites", "check", "available", "slots"] **Impact:** ✓ Reduces noise, allowing focus on important words. ✓ Speeds up processing by eliminating unnecessary terms. **Part-of-Speech (POS) Tagging** Assigns grammatical categories (noun, verb, adjective, etc.) to each word, helping in intent recognition. [("register", Verb), ("Data Science", Noun), ("prerequisites", Noun), ("check", Verb), ("available", Adjective), ("slots", Noun)] **Impact:**	10	4	2

	✓ Distinguishes actions (verbs) from subjects (nouns). ✓ Helps recognize intent (e.g., "register" indicates course registration, "check" indicates query for availability). **Stemming** Reduces words to their root form by removing suffixes (e.g., "-ing", "-ed"). ["regist", "data", "scienc", "prerequisit", "check", "avail", "slot"] **Impact:** ✓ Helps in pattern matching, but may produce incorrect roots (e.g., "scienc" instead of "science"). ✓ Works well for simple keyword-based systems. **Lemmatization** Converts words into their base dictionary form while preserving meaning. ["register", "Data", "Science", "prerequisite", "check", "available", "slot"] **Impact:** ✓ More accurate than stemming because it produces correct words. ✓ Enhances query understanding and response accuracy.			
4.	A university wants to cluster students into different performance groups based on their exam scores out of 10 in Engineering and Science course. We have 5 students with their scores: S1=(5.3,7.2), S2=(8.5,9.7), S3=(6.1,8.3), S4=(4.4,6.5), S5=(9.9,8.2) Construct a distance matrix using the manhattan proximity measure and perform complete linkage hierarchal clustering for the above mentioned two-dimensional data. Show cluster results by drawing a dendrogram. **KEY:** **Distance Matrix:**	10	5	3

	S1	S2	S3	S4	S5
S1	0.0	5.7	1.9	1.6	5.6
S2	5.7	0.0	3.8	7.3	2.9
S3	1.9	3.8	0.0	3.5	3.9

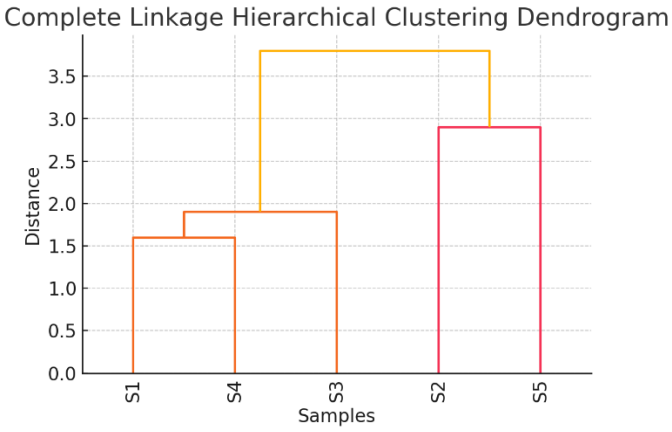
S4	1.6	7.3	3.5	0.0	7.2
S5	5.6	2.9	3.9	7.2	0.0

The **smallest** distance is **1.6**, between **S1 and S4 is called C1**

The next **smallest** distance is **1.9**, between **C1 and S3** called **C2**

The next **smallest** distance is **2.9**, between **S2 and S5** called **C3**.

Now only **C2 (C1(S1, S4), S3)** and **C3 (S2, S5)** remain



5.	Consider following the Stock Level (Units) and Sales Volume (Units Sold) of four products in a supermarket: P1=(100,80),P2=(50,30),P3=(200,190),P4=(75,50) Compute the Eigen vectors and Eigen values for the given data using the Principal Component Analysis (PCA) Algorithm. **KEY:** Cov=[4322.92 4687.50 4687.50 5091.67] **eigenvalues:** λ1= 9410.52 λ2= 4.06	10	5	3

	λ1 eigenvector is: v1=(0.68,-0.73) **λ2 eigenvector is:** v2=(−0.73,-0.68)			

****