REG.NO.:

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

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

SLOT: C2+TC2

Programme Name & Branch : **BTECh, CSE**
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 : **29/01/2025**
Exam Duration : **90 minutes** **Maximum Marks: 50**

General instruction(s):

- Answer All Questions
- M - Max mark; CO – Course Outcome; BL – Blooms Taxonomy Level (1 – Remember, 2 – Understand, 3 – Apply, 4 – Analyse, 5 – Evaluate, 6 – Create)
- Course Outcomes (Type the CO statements covered in this question paper. Use the CO number as per the syllabus copy)
    a. Ability to obtain fundamental knowledge on data science.
    b. Demonstrate proficiency in data analytics.

Q. No	Question	M	CO	BL
1.	Consider that you have been hired by a company that is working with weather data. You have been asked to collect data about cyclones for the past 100 years and develop a model. As a data scientist describe in detail the process that you will follow in developing a model for the weather forecasting. **1. Understanding the Problem** **Objective:** • Predict the **formation, intensity, and trajectory** of cyclones using historical weather data. • Improve **early warning systems** to minimize damage and loss of life. • Provide insights for **climate change studies** and disaster management. **Key Challenges:** • **Data availability & quality** (cyclone data from 100 years may have missing values). • **Climate change impact** (patterns may change over time). • **Feature selection** (choosing relevant meteorological parameters).	10	1	2

Sources of Data:

Data Type	Source	Description
Historical Cyclone Data	NOAA, IMD, JTWC, NASA	Tracks, wind speeds, pressure levels, impact data
Satellite Images	GOES, Meteosat, Himawari	Cloud patterns, sea surface temperatures
Weather Sensor Data	Buoy stations, AWS	Real-time wind speeds, air pressure, temperature
Topographical Data	Google Earth, GIS	Elevation maps, land-sea interaction
Oceanographic Data	Copernicus, ARGO	Sea Surface Temperature (SST), ocean currents

Data Preprocessing & Cleaning

Data Cleaning Steps:

✅ Handle **missing values** (e.g., interpolate time series gaps).
✅ Standardize **units** (e.g., wind speed in km/h vs. mph).
✅ Convert **satellite images** into structured features using **Computer Vision (CNNs)**.
✅ Remove **duplicate records** from different sources.

4. Exploratory Data Analysis (EDA)

EDA helps to understand patterns in cyclone formation and intensity.

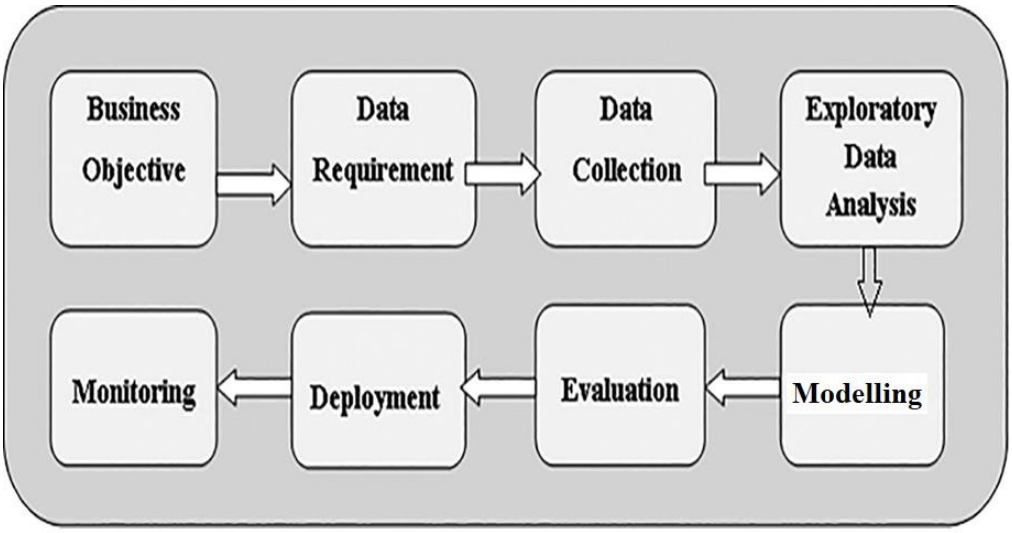
Visualization Techniques:

Time Series Analysis → How cyclone patterns have changed over the last 100 years.
Heatmaps → Geographic distribution of cyclones.
Correlation Matrix → Identifying relationships between pressure, wind speed, and cyclone strength.
Boxplots → Outlier detection in extreme cyclone intensities.

6. Model Evaluation & Validation

Metrics for Model Performance:

📌 **Mean Absolute Error (MAE):** Measures deviation in cyclone strength prediction.
📌 **Root Mean Squared Error (RMSE):** Measures cyclone intensity forecast accuracy.

	📌 **Precision & Recall:** For classifying cyclones into different categories. 📌 **F1-Score:** Balances precision and recall for cyclone impact classification. **7. Model Deployment & Real-Time Forecasting** Once the model is validated, it should be **integrated into a real-world forecasting system**. **Deployment Steps:** ✅ **API Development** (using FastAPI, Flask) → Allow meteorological agencies to access predictions. ✅ **Cloud Deployment** (AWS, Google Cloud, Azure) → Scale processing for real-time forecasting. ✅ **Automated Data Ingestion** (Kafka, Apache Spark) → Real-time weather updates. 			
2.	Netflix is one of the largest streaming services today. To enhance the user experience and to expand its user base, should Netflix use business intelligence tools or would data science methods be more appropriate? Provide your views and recommend tools for the task. **KEY** Compare data science and business intelligence on: Datasource, Methods Skills, Focus,complexity etc. **1. Enhancing User Experience (Data Science Focus)**	10	1	3

Netflix's user experience is primarily driven by **personalization and engagement**. Data science is crucial for:

1. **Personalized Recommendations**
   - Using collaborative filtering and deep learning models to suggest movies and shows based on a user's preferences and behavior.
   - Real-time updates on the recommendations when a user watches or interacts with content.
2. **Content Customization**
   - **A/B Testing** to determine the best UI design or thumbnail images that attract more clicks.
   - **Predictive Models** to recommend different versions of the same content (e.g., trailers, subtitles) based on the user's region or preferences.
3. **Sentiment Analysis**
   - Analyzing user feedback, social media comments, and reviews to gauge the audience's mood and expectations.
4. **Reducing Buffering Time and Improving Streaming Quality**
   - Using real-time data and machine learning models to optimize content delivery based on a user's device and network conditions.

2. Expanding the User Base (Data Science Focus)

To attract new users, Netflix can leverage predictive analytics and marketing insights:

1. **Churn Prediction**
   - Build models to predict which users are likely to cancel subscriptions and intervene with targeted retention strategies (e.g., personalized discounts or reminders about new releases).
2. **Targeted Marketing Campaigns**
   - Use clustering algorithms to identify potential subscriber segments based on demographics, viewing habits, and social behavior.
   - Optimize advertising spend by predicting the effectiveness of marketing campaigns in different regions.
3. **Content Development and Acquisition**
   - Use machine learning to predict what content will be successful in specific markets or among different age groups.
   - Analyze competitors and market trends to find new opportunities for growth.

for AI-driven personalization.

User Experience Goal	Best Approach	Tools
Streaming Quality Monitoring	BI for real-time reports	Tableau, Power BI
Personalized Recommendations	DS for AI-based content suggestions	TensorFlow, PyTorch
UI/UX Optimization	DS for A/B testing	Optimizely, Google Optimize
Sentiment Analysis	DS for user feedback insights	NLP, Apache Spark
Churn Prediction	DS to detect at-risk users	Scikit-learn, BigQuery

3.

	EMPLOYEE_ID	FIRST_NAME	LAST_NAME	EMAIL	PHONE_NUMBER	HIRE_DATE	JOB_ID	SALARY
2	198	Donald	OConnell	DOCONNEL	650.507.9833	21-JUN-07	SH_CLERK	2600
3	199	Douglas	Grant	DGRANT	650.507.9844	13-JAN-08	SH_CLERK	2600
4	200	Jennifer	Whalen	JWHALEN	515.123.4444	17-SEP-03	AD_ASST	4400
5	201	Michael	Hartstein	MHARTSTE	515.123.5555	17-FEB-04	MK_MAN	13000
6	202	Pat	Fay	PFAY	603.123.6666	17-AUG-05	MK_REP	6000
7	203	Susan	Mavris	SMAVRIS	515.123.7777	07-JUN-02	HR_REP	6500
8	204	Hermann	Baer	HBAER	515.123.8888	07-JUN-02	PR_REP	10000
9	205	Shelley	Higgins	SHIGGINS	515.123.8080	07-JUN-02	AC_MGR	12008
10	206	William	Gietz	WGIETZ	515.123.8181	07-JUN-02	AC_ACCOUNT	8300
11	100	Steven	King	SKING	515.123.4567	17-JUN-03	AD_PRES	24000
12	101	Neena	Kochhar	NKOCHHAR	515.123.4568	21-SEP-05	AD_VP	17000

2 3

Considering the data is stored in a database, find the following

i) Find the range of the salary of the employees
Select min(salary) as minRange, max(salary) as maxRange from employee;

ii) Find the employee with the most experience in the company
select first_name,last_name from employee order by hire_date limit 1;

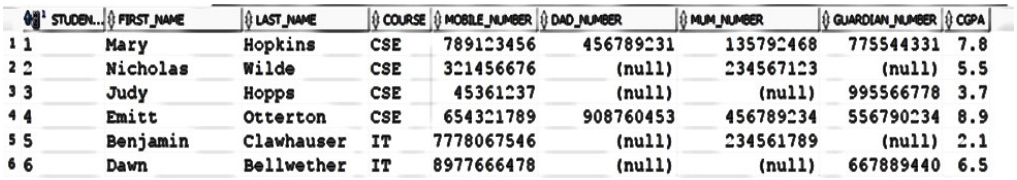
or

If there's a tie (multiple employees hired on the same earliest date), the second query will return all of them.

SELECT * FROM employee WHERE hire_date = (SELECT MIN(hire_date) FROM employee);

iii) List all the high-ranking employees with salary greater than 10000
select first_name,last_name from employee where salary>10000;

iv) List the representatives of various departments.
select first_name,last_name from employee where job_id like

	'%REP'; v) Find the average salary of a clerk. **Select avg(salary) from employee where job_id like '%CLERK';**				
4.	 Consider the above dataset: i) Display 'low' for students with cgpa less than 4, 'medium' for students with cgpa between 4 and 7.5 and 'high' for other students. **select first_name,last_name,** **case** **when cgpa<4 then 'low'** **when cgpa<7.5 then 'medium'** **else 'high'**	10	2	3

STUDEN...	FIRST_NAME	LAST_NAME	COURSE	MOBILE_NUMBER	DAD_NUMBER	MUM_NUMBER	GUARDIAN_NUMBER	CGPA
1	Mary	Hopkins	CSE	789123456	456789231	135792468	775544331	7.8
2	Nicholas	Wilde	CSE	321456676	(null)	234567123	(null)	5.5
3	Judy	Hopps	CSE	45361237	(null)	(null)	995566778	3.7
4	Emitt	Otterton	CSE	654321789	908760453	456789234	556790234	8.9
5	Benjamin	Clawhauser	IT	7778067546	(null)	234561789	(null)	2.1
6	Dawn	Bellwether	IT	8977666478	(null)	(null)	667889440	6.5

	end as cgpaRange **from studentContact;** ii) Display the first available emergency contact for the students from dad, mom or guardian numbers. **select first_name, last_name,** **coalesce(dad_number,mum_number,guardian_number) as emergencyNumber from studentContact;** iii) List the different courses available for the students **select distinct course from studentContact;** iv) Display 'Yes' along with the student name if the student belongs to CSE 'No otherwise. **select first_name,last_name,** **case course** **when 'CSE' then 'Yes'** **else 'No'** **end as belongsToCSE** **rom studentContact;**			

5.	You are a data analyst for a movie streaming platform. The platform has the following tables: Viewership(view_id, movie_id, user_id, view_date, view_duration) Movies(movie_id, movie_title, genre, release_year) Users(user_id, user_name,subscription_type) (subscription type can be: Basic, Standard, Premium). Write SQL queries to answer the following using window functions: i) List the top watched movies by genre (5 Marks): Find the total viewing duration for each movie and rank them within their genre based on viewing duration (highest to lowest). Include movie_title, genre, total_duration, and rank. **Select movie_title, genre, sum(view_duration) over(partition by movie_id) as total_duration, rank() over(partition by genre order by total_duration desc) as movieRank from movies m join viewership v on m.movies_id = v.movies_id;** ii) Find the cumulative viewing time for users (5 Marks): Calculate the cumulative viewing time for each user over time, ordered by view_date. Include user_name, view_date, view_duration, and cumulative_duration. **Select user_name, view_date, view_duration, sum(view_duration) over(partition by u.user_id order by view_date rows between unbounded preceding and current row) as cumulative_duration from users u join viewership v on u.user__id=v.user_id;**	10	2	3

****