Printed Page:-04 **Subject Code**:- **ACSAI0617**

Roll. No:

NOIDA INSTITUTE OF ENGINEERING AND TECHNOLOGY, GREATER NOIDA

(An Autonomous Institute Affiliated to AKTU, Lucknow)

B.Tech

SEM: VI - THEORY EXAMINATION (2024 - 2025)

Subject: Programming for Data Analytics

Time: 3 Hours **Max. Marks: 100**

General Instructions:

IMP: *Verify that you have received the question paper with the correct course, code, branch etc.*

1. *This Question paper comprises of* ***three Sections -A, B, & C.*** *It consists of Multiple Choice Questions (MCQ's) & Subjective type questions.*

2. *Maximum marks for each question are indicated on right -hand side of each question.*

3. *Illustrate your answers with neat sketches wherever necessary.*

4. *Assume suitable data if necessary.*

5. *Preferably, write the answers in sequential order.*

6. *No sheet should be left blank. Any written material after a blank sheet will not be evaluated/checked.*

SECTION-A 20

1. Attempt all parts:-

1-a. What is sensitivity in a confusion matrix?(CO1,K1) 1

(a) The ability of a model to correctly identify negative cases

(b) The ability of a model to correctly identify positive cases

(c) The percentage of true positives out of all actual positives

(d) The percentage of true negatives out of all actual negatives

1-b. Which library is commonly used for data visualization in Python?(CO1,K1) 1

(a) NumPy

(b) Pandas

(c) Matplotlib

(d) Scikit-learn

1-c. Which of the following is lattice command for producing a scatterplot?(CO2,K1) 1

(a) plot()

(b) lm()

(c) xyplot()

(d) anova()

1-d. What is a geom in the ggplot2 system?(CO2,K1) 1

(a) a plotting object like point, line, or other shape

(b) a method for making conditioning plots

(c) a method for mapping data to attributes like color and size

.

(d) a statistical transformation

1-e. The concatenation of the collection name and database name is called a - (CO3,K1) 1

(a) Namespace
(b) MongoDB
(c) Sharding
(d) replica

1-f. A collection and a document in MongoDB is equivalent to..... concepts respectively.(CO3,K1) 1

(a) Table and Column
(b) Table and Row
(c) Column and Row
(d) Database and Table

1-g. Which of the following is a primary use of TensorFlow?(CO4,K1) 1

(a) Creating presentations
(b) Building AI systems
(c) Writing essays
(d) Editing images

1-h. What is the purpose of Word Vectors in natural language processing? (CO4,K1) 1

(a) To represent words as numerical vectors
(b) To delete words from text
(c) To change the font of words
(d) To align words in a document

1-i. Which of the following is not a step in the typical training process of a neural network?(CO5,K1) 1

(a) Forward propagation
(b) Backward propagation
(c) Validation
(d) Regularization

1-j. Which of the following is not a common loss function for regression problems?(CO5,K1) 1

(a) Mean Absolute Error (MAE)
(b) Mean Squared Error (MSE)
(c) Binary Cross-Entropy
(d) Huber Loss

2. Attempt all parts:-

2.a. Explain method used to read a CSV file into a DataFrame in Pandas.(CO1,K2) 2

2.b. Explain the purpose of the mutate() function in the dplyr package. (CO2,K2) 2

.

2.c.	Describe the process of handling errors when executing SQL queries in Python.(CO3,K2)	2
2.d.	Describe the process of visualizing word embeddings using TensorFlow.(CO4,K2)	2
2.e.	Define deep neural networks (DNNs) in one sentence.(CO5,K1)	2

SECTION-B 30

3. Answer any five of the following:-

3-a.	Explain the concept of data preprocessing and its significance in preparing datasets for analysis.(CO1,K2)	6
3-b.	Discuss the challenges associated with handling missing data in datasets and describe techniques for dealing with them.(CO1,K2)	6
3-c.	Describe key functions in dplyr such as filter, mutate, summarise, and arrange, and provide examples of how they are used.(CO2,K2)	6
3-d.	Discuss common functions in stringr such as str_detect, str_replace, and str_split, and demonstrate their usage with examples.(CO2,K2)	6
3.e.	Enlist differences between SQL and NoSQL databases?(CO3,K1)	6
3.f.	Discuss the advanced features and capabilities of Recurrent Neural Networks (RNNs) in TensorFlow for sequential data processing tasks.(CO4,K2)	6
3.g.	Discuss the concept of autoencoders and elaborate on their role in learning efficient data representations.(CO5,K2)	6

SECTION-C 50

4. Answer any one of the following:-

4-a.	Discuss the challenges associated with handling imbalanced datasets extensively, covering techniques such as resampling methods, class weighting, and algorithmic approaches to address class imbalance. (CO1,K2)	10
4-b.	Discuss the role of sensitivity analysis in model evaluation and decision-making, covering techniques for assessing the sensitivity of model outputs to changes in input parameters and assumptions.(CO1,K2)	10

5. Answer any one of the following:-

5-a.	Explain the purpose of descriptive statistics in data analysis and discuss commonly used measures such as mean, median, mode, variance, and standard deviation. Describe how to calculate descriptive statistics for numerical and categorical variables using functions like mean, median, sd, and summary.(CO2,K2)	10
5-b.	Discuss various functions available in dplyr for data manipulation tasks such as filtering, summarizing, mutating, and arranging data. Provide examples of how to use these functions to clean and transform datasets in R. (CO2,K2)	10

6. Answer any one of the following:-

6-a.	Create an orders collection with keys order_id, cust_id, cust_name, phone_no(array field), email_id(optional field), item_name, DtOfOrder, quantity, amount, status(P :pending / D:delivered) (CO3,K4) a) Display all customer names of orders collection with no repetition.	10

.

b) Show results and details of sorting documents based on amount.
c) Show how many orders are placed by each customer.
d) how many orders are pending

6-b. Explain types of NoSQL databases. How is NoSQL different from RDBMS?(CO3,K2) 10

7. Answer any one of the following:-

7-a. Discuss the theoretical underpinnings of CNNs and their practical applications in image recognition, classification, and segmentation tasks.(CO4,K2) 10

7-b. Analyze the principles and techniques behind Distributed TensorFlow for distributed computing and parallel processing of deep learning tasks across multiple devices or machines.(CO4,K4) 10

8. Answer any one of the following:-

8-a. Explain how Improved GANs is better than the traditional GAN architecture, and what are some of the specific techniques and architectures used to stabilize training, improve convergence, and generate higher-quality samples? (CO5,K2) 10

8-b. Disentangled Representation GANs are designed to separate factors of variation in data, but could you provide an in-depth explanation of their architecture, loss functions, and applications in tasks such as style transfer, attribute manipulation, and domain adaptation?(CO5,K2) 10

.