Reg. No															

B.Tech DEGREE EXAMINATION, DECEMBER 2023

Second Semester

18AIC101J - FOUNDATION OF DATA ANALYSIS

(For the candidates admitted during the academic year 2020 - 2021 & 2021 - 2022)

Note:

i. **Part - A** should be answered in OMR sheet within first 40 minutes and OMR sheet should be handed over to hall invigilator at the end of 40th minute.

ii. **Part - B** and **Part - C** should be answered in answer booklet.

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

PART - A (20 × 1 = 20 Marks)

Answer **all** Questions

		Marks	BL	CO
1.	Which of the following are the knowledge domains of the data analyst? (A) Artificial Intelligence, Mathematics (B) Mathematics, Statistics, Computer Science and Professional fields (C) Only Computer Science and Maths (D) Python, C and C++	1	1	1
2.	What is the operation that is carried out in the data visualization process? (A) Animation (B) Visualizing Charts and Graphs (C) Both a and b (D) Data modeling	1	1	1
3.	Determine the nature of the following data. Shoe sizes (4, 5, 6, 6.5) of different age groups. (A) Continuous (B) Discrete (C) Integer (D) String	1	2	1
4.	Interpretation without sacrificing important information is _____ (A) Visualization (B) Extraction (C) Normalization (D) Summarization	1	2	1
5.	If b = np.array([[1.3, 2.4],[0.3, 4.1]]). What is b.size and b.shape? (A) 4 and (2, 2) (B) (2, 2) and 4 (C) 4, 4 (D) 2, 2	1	2	2
6.	To create a two-dimensional array 3*3, full of zeros, the syntax is (A) np.zero((3,3)) (B) np.zeros((3,3)) (C) numpy.zeros((3,3)) (D) nzeros((3,3))	1	2	2
7.	np.dot(A,B) is same as (A) A.dot(B) (B) np.(A,B) (C) np.dot(B,A) (D) np.(B,A)	1	2	2
8.	What is the output of np.arange (0,6,0.6)? (A) array([0.0, 0.6, 1.2, 1.8, 2.4, 3.0 , 3.6, 4.2, 4.8, 5.4]) (B) array([0.0, 6, 12 , 18, 24, 30,36,42,48,54]) (C) array([0.6,1.2, 1.8, 2.4, 3.0 , 3.6, 4.2, 4.8, 5.4]) (D) array([6, 12 , 18, 24, 30,36,42,48,54])	1	2	2
9.	The serialization and de-serialization of the data object of the cPickle module can be done through (A) dumps() function only (B) loads() function only (C) dumps() and loads() function (D) loads() and dumps() function	1	2	3
10.	A Series by default have numeric data labels starting from _____ (A) 3 (B) 2 (C) 1 (D) 0	1	2	3

11.	The two operations in pivoting with hierarchical indexing are _____ (A) Pivot() and Unpivot() (B) Stack() and Unstack() (C) Serialize() and Deserialize() (D) Queue() and Dequeue()	1	1	3
12.	The -------- object supports the operation of an iteration to generate a sequence of two tuples containing the name of the group together with the data portion. (A) concat() (B) join() (C) groupby() (D) merge()	1	1	3
13.	Which Matplotlib function is used to create a scatter plot? (A) plt.line() (B) plt.scatter() (C) plt.bar() (D) plt.plot()	1	2	4
14.	How can you set the color of a line in a Matplotlib plot? (A) plt.color("red") (B) plt.line_color("red") (C) plt.set_color("red") (D) plt.plot(color="red")	1	1	4
15.	In Matplotlib, _____ function is used to save a plot as an image file? (A) plt.save() (B) plt.save_plot() (C) plt.savefig() (D) plt.export()	1	1	4
16.	Which one of the following is a non-parametric algorithm? (A) Decision tree (B) K-Nearest Neighbours (C) Linear Regression (D) Support Vector Machine	1	1	4
17.	The effectiveness of an SVM depends upon _____. (A) Selection of Kernel (B) Kernel Parameters (C) Soft Margin Parameter (D) All of the above	1	1	5
18.	Naïve Bayes algorithm is based on _____ and used for solving classification problems. (A) Bayes Theorem (B) Candidate elimination algorithm (C) EM algorithm (D) None of the above	1	1	5
19.	An optometrist wishes to use people's characteristics to predict their contact lens type for a data mining system. In this case, the contact lens type would be an example of a (A) Class (B) Table (C) Feature (D) None of the above	1	2	5
20.	How can you add a title to a Matplotlib plot? (A) plt.title("My Title") (B) plt.add_title("My Title") (C) plt.label("My Title") (D) plt.set_title("My Title")	1	2	5

PART - B (5 × 4 = 20 Marks)

Answer **any 5** Questions

		Marks	BL	CO
21.	Explain the different types of data in statistics?	4	2	1
22.	Describe briefly the Qualitative and Quantitative Analysis.	4	2	1
23.	Given a NumPy array arr = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]]), Write python code to extract the second column of the array.	4	3	2
24.	What is a DataFrame in pandas? How do you read a CSV file into a data frame?	4	2	3
25.	Write a Python program using Matplotlib to create a figure with two subplots. In the first subplot, plot a sine wave, and in the second subplot, plot a cosine wave. Customize the plot by adding labels, titles, and adjusting the appearance.	4	3	4
26.	What are the support vectors in the context of SVM?	4	2	5
27.	Explain the concept of cross-validation in machine learning. Describe two common types of cross-validation techniques and their advantages?	4	2	5

		Marks	BL	CO
	PART - C (5 × 12 = 60 Marks) Answer **all** Questions			
28.	(a) List the various steps of the data analysis process. Apply the data analysis steps for designing the student information management system in college. (Use academic data and attendance data as fields).	12	3	1
	(OR)			
	(b) Explain the importance of programming skills for a Data Analyst. Describe a situation where you used Python or another programming language in a data analysis project.			
29.	(a) (i) Explain universal function and aggregate functions in NumPy? (6 marks) (ii) Give an example of using an aggregate function to find the mean of an array. (6 marks)	12	3	2
	(OR)			
	(b) Create a nested dictionary for '5' employees using the below fields: Employee1(ID, Name, Age, Sex, Department, Salary) .... Employee5(ID, Name, Age, Sex, Department, Salary). i. How to add another new dictionary Employee6 to the already existing dictionary. ii. How to delete the 3rd and 4th employee record? iii. How to iterate through a nested dictionary?			
30.	(a) Discuss the potential issues associated with duplicate entries in a dataset. Write a Python code snippet using Pandas to identify and remove duplicate rows from a DataFrame.	12	3	3
	(OR)			
	(b) Explain the concept of 'GroupBy' in Pandas. Provide a Python code snippet that demonstrates how to use GroupBy to group data based on a specific column and calculate the mean for each group.			
31.	(a) How would you create a i) basic line plot ii) scatter plot iii) multiple subplots using Matplotlib?	12	3	4
	(OR)			
	(b) How would you create a simple line chart using Matplotlib in Python? Discuss a scenario where a line chart would be more appropriate than a bar chart.			
32.	(a) Write a Python program to determine the line of best fit using linear regression.	12	2	5
	(OR)			
	(b) Explain the concept of a hyperplane in SVM. How does the choice of the kernel affect the performance of an SVM?			

* * * * *