Reg. No															

B.Tech DEGREE EXAMINATION, DECEMBER 2023

Fifth & Seventh Semester

18CSE396T - DATA SCIENCE

(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 is an example of semi structured data? (A) C (B) XML (C) RDBMS (D) Python	1	1	1
2.	Which of the following characteristic of big data is more related to Twitter data? (A) Velocity (B) Variety (C) Volume (D) veracity	1	1	1
3.	Which of the following is an example of NOSQL (A) RDBMS (B) Cassandra (C) Excel (D) Ms Access	1	2	1
4.	_____ is a temporary storage area between the data sources and a data warehouse (A) A data staging area (B) Buffer (C) Secondary memory (D) Data wrangler	1	1	1
5.	Which of the following is the process of cleaning data, normalizing datasets, and performing transformations on the data. often viewed as a preprocessing step? (A) Data wrapping (B) Data normalization (C) Data ordering (D) Data conditioning	1	1	2
6.	Developing an Initial Hypotheses process is done on _____ phase of data science life cycle (A) Model building (B) Model planning (C) Data preparation (D) Data discovery	1	1	2
7.	Which of the following is not a key role of a successful analytics project ? (A) Project manager (B) Data base administrator (C) Data catalogue manager (D) Data Engineer	1	1	2
8.	In which phase of data science lifecycle sets up a pilot project to deploy the work in a controlled way before broadening the work to a full enterprise or ecosystem of users? (A) Data preparation (B) Model building (C) Operationalize (D) Model planning	1	2	2
9.	What is the output of the following code? > A<-"Welcome" > class(A) (A) Integer (B) Real (C) Char (D) Imaginary	1	1	3

10.	Pick up the statement in R used to skip the iteration of loop without terminating it (A) Break (B) Continue (C) Apply (D) Next	1	2	3
11.	The function divides numeric vectors into different ranges (A) quantile() (B) range() (C) cut() (D) quartile()	1	2	4
12.	In R Language the following are all atomic data types EXCEPT (A) integer (B) logical (C) data frame (D) character	1	1	4
13.	How to calculate sensitivity? (A) TP/(TP+FN) (B) TP/(FP+FN) (C) TP/(TP+FP) (D) TP/(TP-FP)	1	1	5
14.	which algorithm is used for clustering the data? (A) Decision tree (B) SVM (C) Naïve bayes (D) K means	1	1	5
15.	_____ is the ratio of correct positive predictions to the total positive predictions (A) recall (B) precision (C) sensitivity (D) specificity	1	2	5
16.	How many different types of Logistic Regression available? (A) 3 (B) 4 (C) 2 (D) 1	1	2	5
17.	Data visualization is also an element of the broader _____. (A) deliver presentation architecture (B) data presentation architecture (C) dataset presentation architecture (D) data process architecture	1	1	6
18.	What is the functionality of KNITR package? (A) Assembles the contents of the original document with code and results (B) Presents the illustration alone (C) Executes the code and stores the result (D) Creates the new document with extraction of comments	1	2	6
19.	Which of the following global options are available for figures in knitr? (A) fig.height (B) fig.size (C) fig.value (D) fig.column	1	1	6
20.	Choose a statement which is not included in documentation (A) Presents the step by step presentation of the deployment (B) Installation of software package (C) share the work (D) Comments of the results	1	2	6

PART - B (5 × 4 = 20 Marks)

Answer **any 5** Questions

		Marks	BL	CO
21.	Build 2 Data Frames with student data of your choice then join using rbind() command and print it.	4	3	1
22.	Brief about big data characteristics	4	3	1
23.	Explain about 4 main key deliverables of an analytics project	4	4	2

24.	Compare Null and alternate hypothesis with suitable examples	4	3	3
25.	Consider a problem statement to analyse the brain tumour detection with the True positive rate 380, false positive 20, True negative 400, False negative 40. Find out the precision, recall and F1 score	4	3	4
26.	What is dirty data? Give an example with bar plot.	4	4	5
27.	Mention the use of hexbin plot. Write the R code for hexbin plot with random values 2000 & bin count will be 60.	4	4	6

PART - C (5 × 12 = 60 Marks)

Answer **all** Questions

		Marks	BL	CO
28.	(a) Explain in detail about structured, unstructured and Quasi structured data with suitable examples.	12	4	1
	(OR)			
	(b) Describe the skillsets of a data scientist in general and technical aspects.			
29.	(a) It has been decided to start a project with the main objective to study ways to reduce the frequency of costly unplanned emergency repair work and downtime in the satellite station. By predicting future satellite failures or necessary maintenance work in good time to allow for organized maintenance to take place. From the above scenario, answer the following question. Write the data preparation key points with the roles and tools involved. Also explain the methods, to create the appropriate model planning with roles and tools.	12	4	2
	(OR)			
	(b) Explain the data discovery phase and communicate the results to the users' phase of data science life cycle			
30.	(a) Write short note on: i. Attribute categories in R, (6 marks) ii) Descriptive statistics methods (6 marks)	12	3	4
	(OR)			
	(b) Explain in detail about student's T -Test and Welch T-Test with suitable examples			
31.	(a) Explain in detail about the Linear and Logistic regression with R code.	12	4	5
	(OR)			
	(b) Elaborate discuss about the evaluation of clustering techniques with an example			
32.	(a) Demonstrate the visualization models used for data exploration and presentation of multiple variables.	12	3	6
	(OR)			
	(b) Discuss about Statistical Properties of Anscombe's Quartet and also discuss KNITR package for documentation in detail.			

* * * * *