Reg. No															

B.Tech DEGREE EXAMINATION, NOVEMBER 2023

Fifth Semester

18ECE339T - DATA ANALYSIS AND VISUALIZATION

(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. In a confidence interval, the range of values above and below the sample statistics is called _____ 1 1 1

(A) standard error of mean (B) margin of error

(C) Confidence Level (D) interval estimate

2. _____ rule is applied to anticipate probable outcomes in a normal distribution. 1 1 1

(A) Empirical (B) Probability

(C) Position adjustment (D) Bonferroni

3. Which is a measure to indicate the extent to which two random variables change in tandem. 1 2 1

(A) Variance (B) Standard Deviation

(C) Correlation (D) Covariance

4. ______ distributions are used to model time to failure/product lifetime and are common in engineering to study product reliability. 1 1 1

(A) Weibull (B) Sampling

(C) Income (D) Central Limit

5. In _____ method of estimating conditional density function, a specific functional form for density model is assumed 1 2 2

(A) parametric (B) non-parametric

(C) semi-parametric (D) normal distribution

6. Which is a special type of regression analysis that is applied to survival or "time to event "data? 1 2 2

(A) simple linear (B) non-linear

(C) cox (D) logistic

7. It works great when it comes to taking decisions on data by creating branches from a root, which are essentially the conditions present in the data, and providing an output known as a leaf. The algorithm is ? 1 2 2

(A) Decision tree (B) SVM

(C) Logistic regression (D) Bayesian theorem

8. ____ is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. 1 1 2

(A) Coefficient of Determination (B) Residual Standard Error

(C) Total sum of squares (D) Residual Sum of Squares

9. ___ imputation replaces missing values with the exact prediction of the regression model. 1 1 3

(A) Stochastic regression (B) Deterministic regression

(C) Hot deck (D) Substitution regression

10. ____ is used for storing and managing data in Relational Database Management System. 1 1 3

(A) XML
(B) JSON
(C) SQL
(D) Data repository

11. In XML, what does the acronym "DTD" stand for? 1 1 3

(A) Data Type Definition
(B) Document Type Definition
(C) Dynamic Text Descriptor
(D) Document Text Data

12. Which data format can be used for sending images for web pages, such as logos or photographs. This format maintains its size and quality throughout multiple saves and changes. 1 2 3

(A) PNG
(B) PDF
(C) MP4
(D) ODP

13. A dataset is ______ if the number of data points available for each class is not similar. 1 1 4

(A) balanced
(B) unbalanced
(C) missing
(D) relavent

14. ____ is a process in which a database designer creates a data model that supports his application. Its purpose is to represent how database objects interact and how they solve business problems. 1 2 4

(A) Infograph
(B) Data visualization
(C) Data designing
(D) Data modeling

15. Which charts can be used for complex datasets with intersecting points. They can be used to show a variety of variables that may not translate as well into two dimensions. 1 2 4

(A) multidimensional
(B) geospatial
(C) hierarchial
(D) temporal

16. In predictive analysis using data visualization, what is the primary purpose of creating predictive models and visualizations? 1 1 4

(A) To summarize historical data trends
(B) To forecast future trends and make data-driven predictions
(C) To showcase the aesthetic appeal of data visualizations
(D) To validate existing theories with visual representations

17. Which is a hierarchy variant where the most relevant information or the most important entities are at one end and the least important are at the opposite end 1 2 5

(A) Categorical
(B) Dependence
(C) Causality
(D) Importance

18. What is the primary purpose of a pie chart in data visualization? 1 1 5

(A) To show the distribution of data over time
(B) To display the relationship between two variables
(C) To represent parts of a whole and their proportions
(D) To compare data points in a scatter plot

19. Relative proximity measures how closely data points are related in terms of their _____ or meaning 1 1 5

(A) semantic
(B) proximity
(C) distance
(D) layout

20. In a heat map, what do color variations typically represent? 1 2 5

(A) The temperature in the environment where the data was collected
(B) The presence or absence of outliers in the dataset
(C) The distribution of data values within a two-dimensional grid
(D) The number of data points in a scatter plot

PART - B (5 × 4 = 20 Marks)

Answer **any 5** Questions

		Marks	BL	CO
21.	What is the graphical summary that is used for bivariate data?	4	2	1
22.	Present the pros and cons of kitchen sink model.	4	2	2
23.	Infer about data leakage in machine learning.	4	1	3
24.	Appraise on information - centric data visualization with example.	4	2	4
25.	List the possible questions while exploring the data.	4	3	5
26.	Deduce the tips for creating effective data visualizations.	4	1	4
27.	Organize the cultural conventions of leverage common color associates.	4	3	5

PART - C (5 × 12 = 60 Marks)

Answer **all** Questions

		Marks	BL	CO
28.	(a) Summarize univariate, bivariate and multivariate data with examples and graphical analysis **(OR)** (b) For some computers, the time period between charges of the battery is normally distributed with a mean of 50 hours and a standard deviation of 15 hours. Rohan has one of these computers and needs to know the probability that the time period will be between 50 and 70 hours.	12	3	1
29.	(a) Examine the types of linear regression. Give the advantages, disadvantages and use cases of linear regression. **(OR)** (b) How would you decide which type of linear regression (simple or multiple) is more appropriate for a given research or prediction task, and what considerations and statistical techniques would you use to support your choice?	12	2	2
30.	(a) What is multiple imputation. Give the methods and models to impute missing data. **(OR)** (b) (i) Find the outlier in the given data using interquartile range. Data: 22, 24, 25, 28, 29, 31, 35, 37, 41, 53, 64 [6 Marks] (ii) Articulate the SQL operators with examples. [6 Marks]	12	2	3
31.	(a) How does the use of data visualization contribute to the persuasive power of a presentation or report, and why is it considered an effective communication tool for conveying complex information? **(OR)** (b) Comment on the benefits of data visualization. What are the tools that may be used for data visualization?	12	2	4
32.	(a) Present the quantitative, comparative, relational and spatial formats of common visualization layouts and axis styles. **(OR)** (b) Infer keys vs direct labeling of data points and explain the pitfalls.	12	2	5

* * * * *