Reg. No

B.Tech DEGREE EXAMINATION, DECEMBER 2023

Fifth & Seventh Semester

18CSO106T - DATA ANALYSIS USING OPEN SOURCE TOOL

(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 use a specific name or number to index either rows or columns? (A) Datasets (B) Functions (C) Data frames (D) Data	1	1	1
2.	Which is used to produce a sequential vector:c((1,2,3,4,5,6,7,8))? (A) Seq(8) (B) Seq(10) (C) Seq(12) (D) Seq(15)	1	1	1
3.	Infer the wrong statement from the following: (A) break will execute a loop while a condition is true (B) for will execute a loop a fixed number of times (C) if and else tests a condition and acting on it (D) break is used to break the execution of a loop	1	1	1
4.	Outline the number of statements that provide explicit looping in R (A) 2 (B) 4 (C) 5 (D) 3	1	1	1
5.	A residual is defined as (A) The square root of the slope (B) The difference between the actual Y values and the predicted Y values (C) The difference between actual Y values and the mean of Y (D) The predicted value of Y for the average X value	1	1	2
6.	In R, Which function used for linear regression ? (A) lr(formula, data) (B) lrm(formula, data) (C) lm(formula, data) (D) regression.linear (formula, data)	1	1	2
7.	__________ is defined as the sum of squares of the difference between the observations and the line in the horizontal direction in the scatter diagram that can be minimized to obtain the estimates (A) formal regression (B) simple regression (C) logistic regression (D) reverse regression method	1	1	2
8.	Express what does K stand for in K means algorithm (A) Number of clusters (B) Number of data (C) Number of iterations (D) Number of attributes	1	1	2
9.	A multiple regression model has: (A) Cannot be determined (B) Only one independent variable (C) More than one dependent variable (D) More than one independent variable	1	1	3
10.	If the predicted logit in a logistic regression is 0, what is the transformed probability? (A) 0 (B) 1 (C) 0.5 (D) 0.05	1	1	3

11. Which evaluation measures cannot be used to compare the results of a logistic regression with the intended outcome? 1 1 3

(A) AUC-ROC
(B) Accuracy
(C) Logloss
(D) Mean-Squared-Error

12. Logistic Regression is a Machine Learning algorithm that is used to predict the probability of a _____ 1 1 3

(A) categorical independent variable.
(B) categorical dependent variable.
(C) numerical dependent variable.
(D) numerical independent variable.

13. How does a ridge regression estimator's bias-variance decomposition compare to that of an ordinary least squares regression? 1 1 4

(A) Ridge has larger bias, smaller variance
(B) Ridge has larger bias, larger variance
(C) Ridge has smaller bias, larger variance
(D) Ridge has smaller bias, smaller variance

14. K-fold cross-validation is 1 1 4

(A) quadratic in K
(B) linear in K
(C) exponential in K
(D) cubic in K

15. Identify the penalty term for the Lasso regression 1 1 4

(A) the absolute sum of the coefficients
(B) the square of the magnitude of the coefficients
(C) the square root of the magnitude of the coefficients
(D) the sum of the coefficients

16. Infer the penalty term for the Ridge regression? 1 1 4

(A) the square root of the magnitude of the coefficients
(B) the sum of the coefficients
(C) the square of the magnitude of the coefficients
(D) the absolute sum of the coefficients

17. Principal Component Analysis reduces the dimension by finding a few________ 1 1 5

(A) Octagonal linear combination
(B) Pentagonal Linear Combination
(C) Orthogonal linear combinations
(D) Hexagonal linear combination

18. End Nodes are represented by __________ 1 1 5

(A) Circles
(B) Triangles
(C) Squares
(D) Disks

19. A _________ is defined as a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. 1 1 5

(A) Neural Networks
(B) Trees
(C) Graphs
(D) Decision tree

20. When you use the boosting algorithm, you always consider the weak learners. Which of the following is the main reason for having weak learners? 1 1 5

(A) To prevent overfitting
(B) To prevent under fitting
(C) to prevent overfitting and underfitting
(D) to prevent neither overfitting nor under fitting

PART - B (5 × 4 = 20 Marks) Marks BL CO

Answer **any 5** Questions

21. Create a scatter plot using the 'plot()' function of R to visualize the relationship between the 'Sepal.Length' and 'Petal.Length' columns in the 'iris' dataset. 4 2 1

22. Using R programming, create a data frame 'student_data' with two columns: 'Name' and 'Age.' Populate it with information for three students. Also, Use the 'summary()' function to provide a summary of 'student_data.' 4 2 1

23. In a multiple linear regression model, what does the coefficient for an independent variable represent? And, If the coefficient for a particular predictor is close to zero, how does it affect the interpretation of that predictor's impact on the dependent variable? 4 2 2

24. Explain the fundamental concept of Bayes' Theorem and its significance in probability theory. Provide a brief example or scenario where Bayes' Theorem can be applied, and describe how it helps in updating probabilities. 4 2 3

25. Explain the primary objective of Linear Discriminant Analysis (LDA) in machine learning. Describe the key difference between LDA and Principal Component Analysis (PCA) when it comes to dimensionality reduction. Provide a brief example or scenario where LDA can be effectively applied for classification or feature selection. 4 2 3

26. Describe the main objective of fitting a lasso model on a training set, and highlight the importance of selecting λ through cross-validation. 4 2 4

27. Illustrate the practical use of PCA with a real-world example where it can help reduce data dimensionality while retaining essential information. 4 2 5

PART - C (5 × 12 = 60 Marks)

Answer **all** Questions

Marks BL CO

28. (a) Create an R program to assemble a list that comprises a vector, a matrix, and another list. Assign names to the elements within this list. Then, determine the total number of objects within the list. Subsequently, access the first and second elements of the list, and subsequently remove the third element. 12 3 1

(OR)

(b) Using the provided variables in the Iris Dataset, including Sepal Length, Sepal Width, Petal Length, Petal Width, and Species:
(i) What is the recommended method for importing the data into R before commencing the analysis?
(ii) Could you please provide examples of various formats for importing the data into R?

29. (a) Explain the concept of Simple Linear Regression. What does the regression equation represent, and how do you interpret its coefficients? Also, Implement Simple Linear Regression in R to model the relationship between "X" and "Y" from your dataset. Provide the R code for loading data, fitting the regression model, and displaying the summary results. Interpret the key output, including the coefficients and their significance. 12 3 2

(OR)

(b) Describe the core idea of Multiple Linear Regression. What does the regression equation represent in this context, and how do you interpret the coefficients for multiple predictors? Also, Demonstrate how to conduct a Multiple Linear Regression analysis in R to predict the response variable "Y" based on multiple predictors. Provide the R code for loading the dataset, building the regression model, and summarizing the results. Explain the significance of predictors and any considerations for model assessment.

30. (a) Explore the application of LDA for dimensionality reduction in a dataset. Apply LDA to reduce dimensionality for the given dataset, which consists of two classes: 12 3 3

Class 1 (C1):
Data points: X1 = {(3,2), (2,4), (2,3), (3,6), (5, 5)}

Class 2 (C2):
Data points: X2 = {(9,10), (6,8), (9,5), (8,7), (10,8)}

(OR)

(b) Suppose you are working on a spam email classification task. You have a dataset with 1000 emails, where 400 are spam (S) and 600 are not spam (NS). Additionally, you have two features: "Contains the word 'free'" (F) and "Contains the word 'discount'" (D).

Out of the 400 spam emails, 300 contain the word 'free,' and 200 contain the
word
'discount.' Out of the 600 non-spam emails, 50 contain the word 'free,' and
100
contain the word 'discount.'

Calculate the following using Bayes' Theorem:

(i) The probability that an email is spam given that it contains the word 'free' (P(S|F)).
(ii) The probability that an email is spam given that it contains the word 'discount' (P(S|D)).

Show your calculations and explain the steps you take to arrive at the answers.

31. (a) Perform model selection using three different approaches: best subset selection, forward stepwise selection, and backward stepwise selection, each with a maximum of two models. Then, compare the results obtained from these approaches. Additionally, implement stepwise subset selection in R programming, considering an arbitrary number of predictors. 12 3 4

(OR)

(b) Explain the process of implementing k-fold cross-validation in R and provide a step-by-step guide with relevant examples. Additionally, discuss the strengths and weaknesses of k-fold cross-validation compared to the validation set method and leave-one-out cross-validation (LOOCV).

32. (a) Apply boosting, bagging, and random forests to a dataset of your choice. Ensure that you train these models on a designated training set and assess their performance using a separate test set. How do the results in terms of accuracy compare to simpler techniques like linear or logistic regression? Ultimately, which of these methods demonstrates the. most favorable performance? 12 3 5

(OR)

(b) Explain the purpose of utilizing a Decision Tree and implement it in R. Additionally, demonstrate the implementation of Boosting in R.

*****