27. a. The ABC laboratory was interested in determining whether there was a correlation between height and weight, such that as height grows, so does weight. For each of the groups, they took a sample of 1000 persons and calculated the average height of that group and the details are given below: 10 3 2 2

Height in (cms)	Weight (in kg)
130	55
135	56
140	62
142	63
147	63
156	61

Compute the regression coefficients and come up with the conclusion that any such relationship exists.

(OR)

b.i. Compare simple linear regression and multiple linear regression. 5 4 3 2

ii. Construct program in R to implement simple linear regression. 5 4 3 2

28. a. Suppose there are 2 categories, category A and category B and we have a new data point X1, so this data point will lie in which of these categories? Identify the algorithm that is used to provide the solution to this problem and illustrate with your own example. 10 3 4 1

(OR)

b. Demonstrate Bayes theorem in detail with necessary examples. If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set? Do we expect the test prediction accuracy of QDA relative to LDA improve, decline or unchanged? (In case of increase in sample size)? Why? 10 4 4 1

29. a.i. Infer the steps that are required to perform cross-validation. 5 3 5 4

ii. Compare K-fold cross validation and Loocv. 5 3 5 4

(OR)

b. Apply best subset/ forward stepwise and backward to 2 models and compare the results. Also implement best subset select in R programming? Assume number of predictors on your own. 10 3 5 4

30. a. Illustrate in detail about implementation of principal component analysis in R and also explain the steps that are required for the given data: 10 3 6 4

Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

(OR)

b.i. Predict the reason for using the decision tree and implement the tree in R. 5 3 6 4

ii. Show the implementation of boosting in R. 5 3 6 4

* * * * *

Reg. No. | | | | | | | | | | | | | | |

B.Tech. DEGREE EXAMINATION, NOVEMBER 2022
Sixth/ Seventh Semester

18CSO106T – DATA ANALYTICS USING OPEN SOURCE TOOL
(For the candidates admitted from the academic year 2018-2019 to 2019-2020)

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** should be answered in answer booklet.

Time: 2½ Hours Max. Marks: 75

PART – A (25 × 1 = 25 Marks)

Answer **ALL** Questions

Marks BL CO PO

1. Which of the following is employed in R language statistical analysis? 1 1 1 5
   (A) RStudio (B) Studio
   (C) Heck (D) KStudio

2. R typically treats numbers as ______ precision real numbers. 1 2 1 5
   (A) Single (B) Double
   (C) Real (D) Imaginary

3. Which use a specific name or number to index either rows or columns? 1 2 1 5
   (A) Data sets (B) Data frames
   (C) Data (D) Functions

4. The infinite loop is started by ______ from the beginning. 1 1 1 5
   (A) Never (B) Repeat
   (C) Break (D) Set

5. Which of the following is used to produce a sequential vector: 1 1 1 5
   C((1,2,3,4,5,6,7,8))?
   (A) Seq (8) (B) Seq (10)
   (C) Seq (15) (D) Seq (12)

6. When using a simple linear regression model with a single independent variable, how many coefficients must you estimate? 1 2 1 5
   (A) 1 (B) 2
   (C) 3 (D) 4

7. Confidence intervals and hypothesis tests are calculated under the premise that the errors are independent, normally distributed, have a mean of zero and ______. 1 1 3 2
   (A) Mean (B) Variance
   (C) SD (D) KNN

8. If the slope of the regression equation $y_0 = b_0 + b_1x$ is positive, then 1 2 3 2
   (A) As x increases y decreases (B) As x decreases y increases
   (C) As x increases so does y (D) Either (A) or (C) is correct

9. In R, which function is used for linear regression? 1 1 3 2
   (A) lm (formula, data) (B) lv (formula, data)
   (C) lrm (formula, data) (D) Regression.line (formula, data)

10. Which of the following classes of problem in machine learning are frequently encountered? 1 2 5 4

(A) Regression (B) Classification

(C) Progression (D) Both (A) and (B)

11. A multiple regression model has 1 2 3 2

(A) Only one independent variable (B) More than one independent variable

(C) More than one dependent variable (D) Cannot be determined

12. Which of the following scenarios will cause LDA to fail? 1 2 3 2

(A) If the discriminatory information is in the mean, but not is the variance of the data

(B) If the discrimination information is in the mean and variance of the data

(C) If the discriminatory information is not in the mean, but in the variance of the data

(D) If the discriminatory information is neither in mean nor variance of the data

13. Which of these is untrue? 1 1 3 2

(A) If we take the weighted sum of inputs as the output as we do in linear regression, the value can be more than 1 but we want between 0 and 1

(B) Logistic regression is a generalized linear regression because we don't output the weighted sum of input directly, but pass it through a function

(C) The value of the sigmoid function always lies between 0 and 1

(D) Logistic regression is used to determine the value of a continuous dependent variable

14. Identify the hypothesis of logistic regression? 1 2 4 1

(A) To limit the cost function between 0 and 1

(B) To limit the cost function between –1 and 1

(C) To limit the cost function between –infinity and +infinity

(D) To limit the cost function between 0 and +infinity

15. Recognize the cost function for logistic regression from the following 1 1 3 2

(A) Sigmoid function (B) Logistic function

(C) Both (A) and (B) (D) Linear function

16. Which of the following examples of cross validation is appropriate? 1 2 5 4

(A) Selecting variables to include or a model

(B) Comparing predictors

(C) Selecting coefficients (D) Both (A) and (B)

17. Consider selecting a model parameter using co-fold cross validation? Which method is best for selecting a final model to use and calculate its error? 1 2 5 4

(A) Pick any of the 10 models you built for your model; use its error estimate on the held-out data

(B) Train a new model on the full data set, using the parameter you found, use the average CV error as its estimate

(C) Average all of the 10 models you got, use the average CV error as its error estimate

(D) Average all of the 10 models you got, use the error time combined model gives on the full training set

18. What does it mean for the ridge regression if the regularization parameter is set to 0? 1 2 4 1

(A) Large coefficients are not penalized

(B) Over fitting problems are not accounted for

(C) The loss function is not the same as OLS loss function

(D) Both (A) and (B)

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

(A) Ridge has larger bias, larger variance

(B) Ridge has larger bias, smaller variance

(C) Ridge has smaller bias, larger variance

(D) Ridge has smaller bias, smaller variance

20. K-fold cross-validation is 1 1 5 4

(A) Linear in K (B) Quadratic in K

(C) Cubic in K (D) Exponential in K

21. Decision nodes are represented by ______. 1 1 5 4

(A) Disks (B) Squares

(C) Circles (D) Triangles

22. Principal component analysis reduces the dimension by finding a few ___. 1 2 5 4

(A) Hexagonal linear combination (B) Orthogonal linear combination

(C) Octagonal linear combination (D) Pentagonal linear combination

23. If you are building a model using a random forest-bagging based approach. Which among the following is possible? 1 1 6 4

(i) Number of trees should be as large as possible

(ii) Have interpretability after using random forest

(A) (i) only (B) (ii) only

(C) (i) and (ii) (D) Cannot be determined

24. Which of the following is true in this situation when applying bagging on regression trees? 1 1 6 4

(i) Build the N regression with N bootstrap

(ii) Take the average of N regression tree

(iii) Each tree has a high variance with low bias

(A) (i) and (ii) (B) (ii) and (iii)

(C) (i) and (iii) (D) (i), (ii) and (iii)

25. Which result does hierarchical clustering ultimately produce? 1 2 6 4

(A) Final estimate of cluster centroids

(B) Tree showing how close things are to each other

(C) Assignment of each point to clusters

(D) Predicting the best fit

PART – B (5 × 10 = 50 Marks)

Answer **ALL** Questions

Marks BL CO PO

26. a. Use functions in R to perform the following task on the given set of data: 10 3 1 5

Num = [10, 22, 31, –42, 51, –25, 28, –27, 30]

(i) Segregate the odd and even numbers

(ii) Segregate positive and negative numbers

(OR)

b. Assume you are given with a collection of data that contains list of items of the same type. Write an R program to perform the following operations 10 4 1 5

(i) Identify the second highest value in a given set of data and also identify whether a given set of data contains a specified element

(ii) Sort the list of items in descending order