Reg. No.															

B.Tech. DEGREE EXAMINATION, MAY 2022

Fourth to Seventh Semester

18CSO106T – DATA ANALYSIS 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. The term data analysis is defined by the statistician (A) John Tukey (B) William S (C) Hans Peter Luhn (D) William John	1	1	1	1
2. _____ refers to data which can be ranked, has consistent units and has a true zero. Eg:age Some statistics software packages may refer to cardinal and ratio data are 'scale' (A) Nominal data (B) Ordinal data (C) Cardinal / Interval data (D) Ratio data	1	1	1	1
3. _____ are a more generalized form of a matrix (A) Factors (B) Matrices (C) Vectors (D) Data frames	1	1	1	1
4. Which of the following is valid syntax for if else statement in R? (A) If(condition>) {# # do something} else {# # do something else} (B) If(condition>) {# # do something} elseif {# # do something else} (C) If(<condition>) {# # do something} elseif {# # do something else} (D) If(<condition>) {# # do something} else {# # do something else}	1	1	1	1
5. Point out the correct statement? (A) The value NaN represents undefined value (B) NaN can also be thought of as a missing value (C) Number Inf represents infinity in R (D) 'raw' objects are commonly used directly in data analysis	1	1	1	1
6. In practice, line of best fit or regression line is found when + (A) Sum of residuals $\left(\Sigma\left(y-h\left(x\right)\right)\right)$ is minimum (B) Sum of the absolute value of residuals $\left(\Sigma\left(y-h\left(x\right)\right)\right)$ is maximum (C) Sum of square of residuals $\left(\Sigma\left(y-h\left(x\right)\right)2\right)$ is maximum (D) Sum of square of residuals $\left(\Sigma\left(y-h\left(x\right)\right)2\right)$ is minimum	1	1	2	2

7. In the mathematical equation of linear regression $y = \beta_1 + \beta_2 + \varepsilon, (\beta_1, \beta_2)$ refers to 1 2 2 2

(A) (x-intercept, slope) (B) (slope, x-intercept)
(C) (y-intercept, slope) (D) (slope, y-intercept)

8. Consider the following learning algorithms: 1 2 2 1

(i) Logistic regression
(ii) Back propagation
(iii) Linear regression

Which of the following options represents classification algorithms?

(A) Only (i) and (ii) (B) Only (i) and (iii)
(C) Only (ii) and (iii) (D) (i), (ii) and (iii)

9. Which of following metrics can be used for evaluating regression models? 1 2 2 1

(i) R squared
(ii) Adjusted R squared
(iii) F statistics
(iv) RMSE / MSE / MAE

(A) (ii) and (iv) (B) (i) and (ii)
(C) (i), (ii), (iii) and (iv) (D) (i), (iii) and (iv)

10. If the absolute value of your calculated t-statistic exceeds the critical value from the standard normal distribution you can 1 2 2 1

(A) Safely assume that your regression results re significant
(B) Reject the null hypothesis
(C) Reject the assumption that the error terms are homoskedastic
(D) Conclude that most of the actual values are very close to the regression line

11. Which of the following methods do we use to best fit the data in logistic regression? 1 2 3 1

(A) Least square error (B) Maximum likelihood
(C) Jaccard distance (D) Both (A) and (B)

12. _____ measures the model prediction error. It corresponds to the average difference between the observed known values of the outcome and the predicted value by the model. 1 2 3 1

(A) R-square (B) Root mean squared error
(C) Residual sum of squares (D) Ordinary least squares

13. The _____ function produces a matrix that contains all the pairwise correlations among the predictors in a dataset. The first command below gives an error message because the _____ variable is qualitative. 1 1 3 1

(A) Pair ( ) and Direction (B) Predict ( ) and Unidirection
(C) Cor ( ) and Direction (D) Lda ( ) Cor

14. If two variables, x and y, have a very strong linear relationship, then 1 2 3 1

(A) There is an evidence that x causes a change in y
(B) There is an evidence that y causes a change in x
(C) There might not be any causal relationship between x and y
(D) None of these alternatives is correct

15. Ridge regression takes _____ value of variables. 1 1 3 1

(A) Squared value of variables (B) Absolute value of variables
(C) Cube value of variables (D) Root value of variables

16. Which of the following is not a step involved in leave-one-out cross validation? 1 2 4 2

(A) Leave out one data point and build the model on the rest of dataset
(B) Test the model against the next subset and record the test error associated with the prediction
(C) Repeat the process for all data points
(D) Compute the overall prediction error by taking the average of all these test error estimates recorded at step 2

17. Which of the following is true about the tuning parameter in the Lasso model? 1 2 4 2

(A) Accounts for the amount of expansion of data values about a central point (B) Results in a trade-off between bias and variance in resulting estimators
(C) Increases with variance (D) Does not increase with bias

18. Suppose we fit "Lasso regression" to a dataset, which has 100 features, (x_1, x_2,....x_{100}). Now, we resale one of these feature by multiplying with 10 (say that feature is x_1), and then refit Lasso regression with the same regularization parameter. Now, which of the following options will be correct? 1 2 4 4

(A) It is more likely for x_1 to be included in the model (B) It is more likely for x_1 to be excluded from the model
(C) Can't say (D) None of these

19. Which of the following step / assumption in regression modeling impacts the trade-off between under-fitting and over-fitting the most. 1 2 4 4

(A) The polynomial degree (B) Whether we learn the weights by matrix inversion
(C) The use of a constant-term (D) The non-polynomial degree

20. Let's say a 'Linear regression' model perfectly fits the training data (train error is zero). Now which of the following statement is true? 1 2 4 2

(A) You will always have test error zero (B) You can not have test error zero
(C) None of the above (D) You can have test error zero

21. Based on the cues, choose the most appropriate answer: 1 2 5 1

(i) It is a set of nested clusters that are arranged as a tree
(ii) Requires the computation and storage of an n×n distance matrix

(A) K-mean clustering (B) Hierarchical clustering
(C) K-fold cross validation (D) Regression tree

22. Decision trees can be used if the input and output variables are 1 1 5 1

(A) Categorical (B) Continuous
(C) Both (A) and (B) (D) None of the above

23.	_____ is a special type of bagging applied to decision trees. (A) PCA (B) Bagging (C) Boosting (D) Random forest	1	1	5	1
24.	Which of the following need not be tuned using cross-validation to avoid overfitting in random forest algorithm? (A) Minimum size of terminal nodes (B) Maximum size of terminal nodes (C) Maximum number of terminal nodes (D) None of the above	1	1	5	4
25.	Observe the code snippet given below and answer: What should be filled in place of method to fit a linear regression with backward solution? step.model <-train (Inputdatafile_., method = "_____", tuneGrid = data.frame (nvmax = 1:8), trcontrol = train.control) (A) Leap forward (B) Leap backward (C) Feed backward (D) Leap seq	1	1	5	4

PART – B (5 × 10 = 50 Marks)

Answer **ALL** Questions

		Marks	BL	CO	PO
26. a.i.	How to change a Data frame's row and column names?	3	3	1	1
ii.	Explain the types of data and measurement scales. Nominal, ordinal, interval and ratio.	7	4	1	1
	(OR)				
b.i.	Difference between Array vs matrix in R programming.	4	4	1	1
ii.	Create three vectors x,y,z with integers and each vector has 3 elements. Combine the three vectors to becomes a 3×3 matrix A where each column represents a vector. Change the row names to a, b, c.	6	3	1	1
27. a.i.	Mention the library lies used for linear regression in R.	3	3	2	2
ii.	What are the assumptions of linear regression?	4	3	2	2
iii.	With a syntax / code snippet, explain how MGARCH BEK is performed using R programming language.	3	3	2	1
	(OR)				
b.i.	What is the Null and Alternate Hypothesis?	4	3	2	1
ii.	Discuss briefly on k-fold cross validation method and its pros and cons.	6	4	2	2
28. a.i.	How to choose a regression model that is best fit for a given data?	5	3	3	1
ii.	Explain factors regression in R.	5	3	3	1
	(OR)				

b.i.	Can logistic regression be used for more than two classes?	3	4	3	1
ii.	What is the ROC curve in logistic regression?	4	4	3	1
iii.	Explain how and why AUC ROC be used for regression. If not why?	3	4	3	1
29. a.i.	Which of the following is/are one of the important step(s) to pre-process the text in NLP based projects? Justify (A) Stemming (B) Stop word removal (C) Object standardization	6	3	4	1
ii.	Explain if containers can be nested in Bootstrap? If not, why.	4	3	4	1
	(OR)				
b.i.	Is leave one out-cross validation a better method than k-fold cross validation? Explain with suitable scenarios.	6	4	4	2
ii.	Comment on the variance of leave-one-out cross-validation.	4	4	4	2
30. a.i.	Explain the basics of decision trees-regression trees, classification trees.	7	3	5	1
ii.	Why is Euclidean distance preferred over Manhattan distance in the k-means algorithms?	3	4	5	1
	(OR)				
b.i.	Explain in detail about (A) Fitting classification trees in R (B) Linear models	5 3	4	5	4
ii.	What are the uses of principal components?	2	3	5	1

* * * * *