Reg. No															

Minor CERTIFICATION EXAMINATION, NOVEMBER 2023

First Semester

18CSE318T - FOUNDATIONS OF DATA SCIENCE

(For the candidates admitted during the academic year (2020-2021 & 2021-20222))

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.	What exactly is data science? (A) The study of computer programming (B) The art of making data beautiful (C) The field of extracting knowledge and insights from data (D) The process of storing data in databases	1	1	1
2.	What is the name of the procedure for preparing data for analysis by cleaning and converting it? (A) Data extraction (B) Data visualisation (C) Data preprocessing (D) Data Engineering	1	1	1
3.	Which of the following doesn't fall under the umbrella of data science? (A) Communication building (B) Operationalize (C) Model planning (D) Discovery	1	1	1
4.	What is the most used R function for importing data from a CSV file? (A) read.csv() (B) load() (C) import.csv() (D) read.data()	1	1	1
5.	What is Exploratory Data Analysis's (EDA) main objective? (A) To build predictive models (B) To clean and preprocess data (C) To understand the data's main characteristics (D) To make data visually appealing	1	2	2
6.	Which of the following is not a regular EDA technique? (A) Histograms (B) Linear regression (C) Box plots (D) Scatter plots	1	2	2
7.	What EDA method is effective for finding outliers in a dataset? (A) Correlation analysis (B) Histograms (C) Principal Component Analysis (PCA) (D) Scatter plots	1	2	2
8.	What is an EDA correlation coefficient used to measure? (A) The strength and direction of a linear relationship between two variables (B) The presence of outliers in a dataset (C) The spread or variability of a dataset (D) The mean of a dataset	1	2	2

9. The fundamental difference between a linear regression model and a generalized linear regression model lies in the following. 1 2 3

(A) The errors in the linear regression are normally distributed, while they can have a more general distribution for the generalized linear model

(B) The errors in the linear regression model are homoskedstic while they are heteroskedstic in a generalized linear model

(C) The generalized linear model is not used for continuous dependent variables, while that is not the case with the linear regression model

(D) The linear regression model is easy to estimate, while the generalized linear regression model is not easy to estimate.

10. The K-nearest neighbours (KNN) algorithm has the following characteristics: 1 2 3

(A) It has slow training phase

(B) It has a fast classification phase

(C) Makes no assumptions about the data distribution

(D) Produces a predictive model

11. The below code snippet shows the accurate implementation of a function that does data normalization. 1 3 3

(A) normalise <- function(x) { return ((x - max(x)) / (max(x) - min(x))) }

(B) normalise <- function(x) { return ((x - min(x)) / (max(x) - min(x))) }

(C) normalise <- function(y) { return ((x - min(x)) / (max(x) - min(x))) }

(D) normalise <- function(x) { return ((x - min(x)) / (min(x) - max(x))) }

12. The above code snippet demonstrates the implementation of a K-nearest neighbours (KNN) model. The model is executed using the following variables: "train" represents the training data, "test" represents the testing data, "train_labels" stores the labels associated with the training data, and the classification is performed based on the 7 nearest neighbours. 1 3 3

(A) wbcd_test_pred <- knn(train =train, test = test, cl =train_labels, k = 21)

(B) wbcd_test_pred <- knn(train =train, test = test, cl = train_labels, k = 7)

(C) wbcd_test_pred <- knn(train = wbcd_train, test = wbcd_test, cl = wbcd_train_labels, k = 21)

(D) wbcd_test_pred <- knn(train = test, test = train, cl = test, k = 21)

13. The act of testing the assumed hypothesis for rejection while thinking it to be true is referred to as what? 1 1 4

(A) Null Hypothesis

(B) Statistical Hypothesis

(C) Simple Hypothesis

(D) Composite Hypothesis

14. Consider the hypothesis H_0 where $\phi_0 = 5$, against H_1, where $\phi_1 > 5$. The test is? 1 4 4

(A) Right tailed

(B) Left tailed

(C) Center tailed

(D) Cross tailed

15. The occurrence of a Type 1 error is seen when 1 2 4

(A) We reject Ho if it is True

(B) We reject Ho if it is False

(C) We accept Ho if it is True

(D) We accept Ho if it is False

16. Which of the following options is appropriate for use with the knitr package? 1 1 5

(A) Reports

(B) Data preprocessing documents

(C) Technical manuals

(D) Design document

17. Which of the following statements is used to import the knitr library? 1 1 5

(A) library(knitr)

(B) import knitr

(C) lib(knitr)

(D) package nitr

18. The document generated by the knitr package often has which of the following extension? 1 1 5

(A) .md

(B) .rmd

(C) .html

(D) .tsv

19.	Which of the following represents the accurate sequence of conversion? (A) .md->.Rmd->.html (B) .Rmd->.md->.html (C) .Rmd->.md->.xml (D) .md->.Rmd->xml	1	1	5
20.	What are the possible global choices for figures in knitr? (A) fig.height (B) fig.size (C) fig.breadth (D) fig.width	1	1	5

PART - B (5 × 4 = 20 Marks)

Answer **any 5** Questions

		Marks	BL	CO
21.	Formulate the syntax for defining matrices in R.	4	2	1
22.	Explore the steps involved in data cleaning.	4	1	2
23.	In the context of problem mapping, how does unsupervised learning vary from supervised learning?	4	2	3
24.	Construct ANOVA table and discuss its types.	4	2	4
25.	Examine the presentation of your results to project sponsors.	4	1	5
26.	What is a confusion matrix? Explain in detail with the help of an example.	4	1	4
27.	Summarize the steps in the presentation of a project on visualization and evaluation of model.	4	2	5

PART - C (5 × 12 = 60 Marks)

Answer **all** Questions

		Marks	BL	CO
28.	(a) Write about the following (i)Descriptive statistics (ii)Contingency tables (iii)Data frames	12	2	1
	(OR)			
	(b) Explain how to import and export data into R?			
29.	(a) What are the different plots used to implement visualization on multiple variables?Explain.	12	4	2
	(OR)			
	(b) Briefly describe data cleaning and visualization in data science.			
30.	(a) How to implement the Apriori algorithm using R programming? Explain.	12	3	3
	(OR)			
	(b) Describe the Naive Bayes theorem. Write R code to implement the Naive Bayes algorithm.			
31.	(a) A group of 5 patients treated with medicine. A is of weight 42,39,38,60 &41 kgs. The second group of 7 patients from the same hospital treated with medicine B is of weight 38, 42, 56, 64, 68, 69, & 62 kgs. Find whether there is any difference between medicines.	12	5	4
	(OR)			
	(b) The following information displays the quantity of worms isolated from the GI regions of four groups of muskrats after an anthelmintic investigation using carbon tetrachloride. Perform a two-way ANOVA analysis.			

I	II	III	IV
338	412	124	389
324	387	353	432
268	400	469	255
147	233	222	133
309	212	111	265

32.	(a) Explain various steps in deploying models into production, documenting work, and building effective presentations.	12	2	5
	(OR)			
	(b) Discuss how to present the results of your projects to end user.			

* * * * *