DATA SCIENCE
(CSEN 5141)

Time Allotted : 3 hrs **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

(i) A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college.
Which of the following statement is true in following case?
(a) Feature F1 is an example of nominal variable
(b) Feature F1 is an example of ordinal variable
(c) It doesn't belong to any of the above category
(d) None of the above.

(ii) How do we perform Bayesian classification when some features are missing?
(a) We integrate the posteriors probabilities over the missing features
(b) We ignore the missing features
(c) We assuming the missing values as the mean of all values
(d) Drop the features completely.

(iii) Suppose, you are given three variables X, Y and Z. The Pearson correlation coefficients for (X, Y), (Y, Z) and (X, Z) are C1, C2 & C3, respectively. Now, you have added 2 in all values of X (i.e. new values become X + 2), subtracted 2 from all values of Y (i.e. new values are Y-2) and Z remains the same. The new coefficients for (X, Y), (Y, Z) and (X, Z) are given by D1, D2 & D3, respectively. How do the values of D1, D2 & D3 relate to C1, C2 & C3?
(a) D1= C1, D2 < C2, D3 > C3 (b) D1 = C1, D2 = C2, D3 = C3
(c) D1> C1, D2 < C2, D3 =C3 (d) D1 < C1, D2 < C2, D3 < C3.

(iv) A scientist takes a big bucket of water from a lake and counts how many species of bacteria, bugs, and other creepy crawlies he finds in the bucket. Identify the sample in this situation.
(a) The sample is "all the species that live in the lake"
(b) The sample is "the species that are in the bucket"
(c) The sample is "the number of species in the lake"
(d) The sample is "the number of species found in the bucket".

(v) Which method shows hierarchical data in a nested format?
(a) Area charts (b) Scatter Plots (c) Treemaps (d) Population pyramid.

(vi) Which of the following is true about Manhattan distance?
(a) It can be used for continuous variables
(b) It can be used for categorical variables
(c) It can be used for categorical as well as continuous
(d) None of these.

(vii) Data has been collected on visitors' viewing habits at a bank's website. Which technique is used to identify pages commonly viewed during the same visit to the website?
(a) Clustering (b) Classification (c) Association Rules (d) Regression.

(viii) Which of the following approach should be used to ask Data Analysis question?
(a) Find out the question which is to be answered
(b) Find only one solution for particular problem
(c) Find out answer from dataset without asking question
(d) None.

(ix) Which of the following statements is/are true about "Type-1" and "Type-2" errors? 1) Type1 is known as false positive and Type2 is known as false negative. 2) Type1 is known as false negative and Type2 is known as false positive. 3) Type1 error occurs when we reject a null hypothesis when it is actually true.
(a) Only (1) (b) (2) and (3) (c) (1) and (3) (d) Only (3).

(x) What is the mean of test scores? {70, 70, 80, 85, 85, 90, 95, 95, 100, 100}
(a) 85, 95, and 100 (b) 30 (c) 87 (d) None.

Group – B

2. (a) How is data science different from big data and analytics? [(CSEN5141.2)(Understand/LOCQ)]
(b) What is data cleansing? What are the important steps of data cleansing? [(CSEN5141.1)(Understand/LOCQ)]
(c) Explain descriptive, predictive, and prescriptive analytics. [(CSEN5141.2)(Understand/LOCQ)]

2 + 4 + 6 = 12

3. (a) During the data preprocessing step, how should one treat missing/null values? Explain with example. [(CSEN5141.2)(Understand/IOCQ)]
(b) What is multicollinearity? (Two features that have a correlation > 0.8) How will this affect your model? [(CSEN5141.3)(Analyze/IOCQ)]
(c) What is resample? Why it is required? Which Sampling Distribution is used for resample? [(CSEN5141.4)(Remember/LOCQ)]

3 + (2 + 3) + (1 + 1 + 2) = 12

Group – C

4. (a) The length of time, in hours, it takes an "over 40" group of people to play one soccer match is normally distributed with a mean of 2 hours and a standard deviation of 0.5 hours. A sample of size n = 50 is drawn randomly from the population.

Find the probability that the sample mean is between 1.8 hours and 2.3 hours. [(CSEN5141.6)(Analyze/HOCQ)]

(b) An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size n = 25 are drawn randomly from the population. Find the probability that the sample mean is between 85 and 92. [(CSEN5141.6)(Analyze/HOCQ)]

6 + 6 = 12

5. (a) What are the fundamental differences between Feature Selection and Dimensionality Reduction Technique. [(CSEN5141.4)(Remember/LOCQ)]

(b) The data below gives the marks obtained by 10 pupils taking Maths and Physics tests,

Pupil	A	B	C	D	E	F	G	H	I	J
Maths marks out of 30 (x)	20	23	8	29	14	11	11	20	17	17
Physics marks out of 30 (y)	30	35	21	33	33	26	22	31	33	36

Is there a connection between the marks gained by ten pupils, A, B, C ..., J in Maths and Physics tests? Identify the correlation using Pearson's correlation. [(CSEN5141.6)(Analyze/HOCQ)]

(c) Define Precision, Recall, Accuracy and ROC Curve. [(CSEN5141.2)(Analyze/IOCQ)]

4 + 4 + 4 = 12

Group – D

6. (a) Develop a predictive model using Naive Bayes algorithm and training data given in the following table. The model will be able to predict the probability that a car will be stolen or not.

Instance No	Colour	Type	Origin	Stolen
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	Sports	Imported	Yes

Use the developed model to predict whether the following cars will be stolen or not.
Car1: <colour = red, Type = SUV, Origin = Domestic>
Car2: <colour = red, Type = SUV, Origin = Imported>
[(CSEN5141.6)(Analyze/HOCQ)]

(b) How do you find RMSE and MSE in a linear regression model? [(CSEN5141.2)(Remember/LOCQ)]

10 + 2 = 12

7. (a) Suppose the fraction of undergraduate students who smoke is 15% and the fraction of graduate students who smoke is 23%. If one-fifth of the college students are

graduate students and the rest are undergraduates, what is the probability that a student who smokes is a graduate student? [(CSEN5141.4)(Evaluate/IOCQ)]

(b) Given the information in part (a), is a randomly chosen college student more likely to be a graduate or undergraduate student? [(CSEN5141.4)(Evaluate/IOCQ)]

(c) Repeat part (b) assuming that the student is a smoker. [(CSEN5141.4)(Evaluate/IOCQ)]

4 + 4 + 4 = 12

Group – E

8. (a) One of the formal definitions of visualization says *"... is the process of extracting salient features from the sets of data and displaying the features in an intuitive and expressive way*". Comment on all the italicized terms in the definition. [(CSEN5141.3)(Understand/IOCQ)]

(b) Define Visual Analytics. Why and when do we use graph database? [(CSEN5141.2)(Remember/LOCQ)]

(c) What is Visualization Lifecycle? [(CSEN5141.2)(Remember/LOCQ)]

5 + (2 + 3) + 2 = 12

9. (a) One of the formal definitions of visualization says "Visual Encoding Design — Use expressive and effective encodings" — Explain with the help of examples. [(CSEN5141.5)(Apply/IOCQ)]

(b) Give examples for any two of the following: [(CSEN5141.5)(Apply/IOCQ)]
(i) Structural visualization
(ii) Temporal visualization
(iii) Geospatial visualization
(iv) Multidimensional visualization.

6 + 6 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	30.20	42.70	27.08

Course Outcome (CO):

On completion of the course the student should be able to:
CSEN5141.1: Understand how data is collected, managed and stored for data science.
CSEN5141.2: Understand the key concepts in data science, including their real-world applications and some of the popular techniques used by data scientists.
CSEN5141.3: Apply data pre-processing techniques.
CSEN5141.4: Evaluate EDA, inference and regression techniques.
CSEN5141.5: Apply data visualization in big-data analytics.
CSEN5141.6: Analyze data science concepts and methods to solve real-world problems.

*LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question.