DATA SCIENCE
(MCAP 2262)

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) For larger values of 'n', Binomial Distribution _____
(a) loses its discreteness (b) tends to Poisson Distribution
(c) stays as it is (d) gives oscillatory values

(ii) A random variable X has the following p.m.f

x_i	1	2	3
f_i	$\frac{1}{2}$	$\frac{1}{3}$	k

Then the value of k is
(a) 1 (b) 0 (c) 5/6 (d) 1/6.

(iii) If $y = 2 - 0.2x$ then the value of y intercept is equal to
(a) -0.2 (b) 2 (c) $2x$ (d) all the above.

(iv) Which of the following will be Euclidean Distance between the two data point A(1,3) and B(2,3)?
(a) 1 (b) 2 (c) 8 (d) 4.

(v) Which of the following machine learning algorithm can be used for imputing missing values of both categorical and continuous variables?
(a) K-NN (b) Random Forest
(c) Linear Regression (d) All of (a), (b) and (c).

(vi) If the mean of the distribution is 5, then the value of $E(2X - 1)$ is
(a) 10 (b) 1 (c) -4 (d) -9.

(vii) For two random variables X and Y, E(XY)=E(X).E(Y) holds if X and Y are
(a) Independent (b) Uncorrelated.
(c) Continuous variate (d) None of these.

(viii) The effectiveness of an SVM depends upon
(a) Selection of Kernel (b) Kernel Parameters.
(c) Soft Margin Parameter C (d) All of the above.

(ix) Naive bayes classifiers are a collection _____ of algorithms
(a) Classification (b) Clustering
(c) Regression (d) All of the above.

(x) Movie Recommendation systems are example of
1. Classification
2. Refinement Learning
3. Regression
(a) 2 (b) 2 and 3 (c) 1 and 3 (d) None of these.

Group- B

2. (a) Explain one-hot encoding with a suitable example. [(CO2)(Remember/LOCQ)]
(b) What's the difference between Feature Engineering vs. Feature Selection? [(CO2)(Understand/LOCQ)]
(c) What is Cross-Validation and why is it important in supervised learning? [(CO1)(Analyze/IOCQ)]
3 + 4 + (2 + 3) = 12

3. (a) When Data Cleaning is necessary in the dataset? What are the step we used for data cleaning? [(CO1)(Identifying/LOCQ)]
(b) How can we handle missing data if it is more in the dataset? [(CO4)(Analyse/IOCQ)]
(c) What do you mean by time series data? [(CO4)(Clarifying/LOCQ)]
(3 + 3) + 4 + 2 = 12

Group - C

4. (a) A consignment of 15 record players contains 4 defectives. The record players are selected at random one by one without replacement and examined. What is the probability that the 9th one examined as the last defective. [(CO2)(Evaluative/HOCQ)]
(b) The probability that a man hitting a target is $\frac{1}{3}$.
(i) If he fires 5 times, what is the probability of hitting the target at least twice?
(ii) How many times must he fire so that the probability of his hitting the target at least once is more than 90%. [(CO2)(Understand/LOCQ)]
4 + (4 + 4) = 12

5. (a) Evaluate the mean and variance of the following distribution :
$f_i = \frac{1}{1+\mu}\left(\frac{\mu}{1+\mu}\right)^i$ where $\mu > 0, x_i = i$ $(i = 0,1,2,3, \ldots.)$. [(CO2)(Evaluate/HOCQ)]
(b) If X is normally distributed with zero mean and unit variance, then find the expectation of X^2. [(CO2)(Remember/LOCQ)]
6 + 6 = 12

Group – D

6. (a) Explain Linear Regression. Write a python code for Prediction using any Regression technique. [(CO6)(Remember/LOCQ)]
   (b) Define Precision, Recall and Accuracy. [(CO1)(Remember/LOCQ)]
   (3 + 3) + 6 = 12

7. (a) How to determine k using the Elbow Method? How to choose initial centroids in K means clustering. [(CO6)(Execute/LOCQ)]
   (b) Why is the odd value of "K" preferable in KNN algorithm? Why should we not use KNN algorithm for large datasets? [(CO6)(Remember/LOCQ)]
   (c) What is Cover's theorem? Explain with a suitable example. [(CO6)(Analyze/IOCQ)]
   (2 + 3) + (2 + 2) + 3 = 12

Group - E

8. (a) Define Visual Analytics. Why and when do we use Graph database? [(CO6)(Remember/LOCQ)]
   (b) What is Z-Score method .Explain the application with a suitable example? [(CO6)(Analyse/IOCQ)]
   (c) What type of data is best visualized with a heat map? [(CO4)(Analyse/IOCQ)]
   (2 + 3) + (2 + 3) + 2 = 12

9. (a) What is a scatter plot? For what type of data scatter plot is usually used for? [(CO4)(Remember/LOCQ)]
   (b) When analyzing a histogram, what are some of the features to look for? [(CO4)(Understand/LOCQ)]
   (c) What do you mean by normalization of Data? Is it necessary for Machine Learning? [(CO4)(Analyse/IOCQ)]
   (2 + 2) + 3 + (3 + 2) = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	60.42	25	14.58

Course Outcome (CO):

After the completion of the course students will be able to

1. Learn and understand how data is collected, managed and stored.
2. Understand and describe the key concepts in data science and popular techniques used by data scientists.
3. Build skills in data management.
4. Demonstrate proficiency with statistical analysis of data.

5. Develop ability to build and assess data-based models.
6. Apply the concepts and algorithms to real-world data.

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