DATA SCIENCE
(MCA2143)

Time Allotted : 2½ hrs **Full Marks : 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 4 (four) from Group B to E, taking one from each group.

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

Group – A

1. Answer any twelve: **12 × 1 = 12**

Choose the correct alternative for the following

(i) Which of the following is NOT a primary method of data acquisition in data science?
(a) Web scraping (b) Sensor data collection
(c) Machine learning modelling (d) Surveys and questionnaires

(ii) Which of the following tasks is typically NOT involved in data wrangling?
(a) Data cleaning
(b) Data visualization
(c) Data transformation
(d) Data integration

(iii) The p.d.f of a random variable X is $f(x) = k(2x - 1), 0 \leq x \leq 2$
The value of k is
(a) 1 (b) $\frac{1}{2}$ (c) $\frac{1}{3}$ (d) $\frac{2}{3}$

(iv) In binomial distribution when $n \to \infty$ and p is very small
(a) It converts into Poisson distribution
(b) It converts into Normal distribution
(c) It converts into Exponential distribution
(d) None of the above

(v) If X is normally distributed with mean zero and unit variance then expectation of X^2
(a) 0 (b) 1 (c) $\frac{3}{7}$ (d) $\frac{4}{3}$

(vi) Which type of machine learning algorithm falls under the category of "unsupervised learning"?
(a) Linear Regression (b) K-means Clustering
(c) Decision Trees (d) Random Forest

(vii) Which of the following statements is not true about SVM?
(a) It has regularization capabilities
(b) It handles non-linear data efficiently
(c) It has much improved stability
(d) Choosing an appropriate kernel function is easy

(viii) Which of the following statements about heat maps in data science is true?
(a) Heat maps are primarily used for visualizing categorical data
(b) Heat maps display data values as colours in a matrix format
(c) Heat maps are only used for 2D data visualization
(d) Heat maps are not useful for identifying patterns or trends in data

(ix) Which of the following functions is used to create a scatter plot in Matplotlib?
(a) plot() (b) bar() (c) scatter() (d) line()

(x) In box plot, data will be divided in how many parts?
(a) 3 (b) 4 (c) 2 (d) as many as we want

Fill in the blanks with the correct word

(xi) In practice, Line of best fit or regression line is found when ____________ .

(xii) The _____________clustering algorithm does not require the assumption of equal-sized clusters.

(xiii) In data science, a histogram is used to represent the _________ distribution of a dataset.

(xiv) In web scraping, the process of navigating through different web pages to extract desired information is commonly achieved using _______.

(xv) Naive Bays Classifiers are a collection of _________ algorithms.

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)(Analyse/IOCQ)]*

3 + 4 + 5 = 12

3. (a) In a study of dietary habits and health outcomes, researchers collected data on the consumption of fruits and incidence of heart disease among a sample of 100 individuals. The following contingency table summarizes their findings:

	High Fruit Consumption	Low Fruit Consumption
Heart Disease	12	28
No Heart Disease	18	42

Calculate the covariance between fruit consumption (coded as 1 for high consumption, 0 for low consumption) and the presence of heart disease (coded as 1 for presence, 0 for absence). Compute the Pearson correlation coefficient between fruit consumption and the presence of heart disease. Interpret the results in the context of the study. *[(CO4)(Analyze/IOCQ)]*

(b) In a study examining the relationship between gender and preferred smartphone operating system, researchers surveyed 500 smartphone users and collected the following contingency table data:

	Android	iOS
Male	150	100
Female	120	130

Using this data, calculate the chi-square statistic for testing the independence of gender and preferred smartphone operating system. *[(CO4)(Apply/IOCQ)]*

6 + 6 = 12

Group - C

4. (a) State the Central Limit theorem.
A person gets Rs.$(2x + 5)$where x denotes the number appearing when abalanced die is rolled once. Then how much money can he expect in the long runper game? *[(CO2)(Evaluate/HOCQ)]*

(b) A missile hits a target with the probability 0.3. How many missiles should befired so that there is at least 80% probability of hitting the target? *[(CO2)(Evaluate/HOCQ)]*

(2 + 4) + 6 =12

5. (a) If four dices are rolled, find the probability of getting a sum of 18 on the upper faces? *[(CO2)(Apply/IOCQ)]*

(b) Find the value of the constant k such that $f(x) = kx(1 - x), 0 < x \leq 1$
$= 0\ , else\ where.$
is a possible density function and compute $P\left(X > \frac{1}{2}\right)$. Also find $E(X)$. *[(CO2)(Evaluate/HOCQ)]*

6 + 6 = 12

Group - D

6. (a) Illustrate the procedure to determine k using Elbow method. How the initial centroids in k means clustering is chosen? *[(CO6)(Analyse/IOCQ)]*

(b) What is known as sigmoid function? Explain with example that why sigmoid function is used in Logistic regression? *[(CO3)(Remember/LOCQ)]*

4 + (6 + 2) = 12

7. (a) Derive the log-likelihood function used in logistic regression for binary outcomes. *[(CO3)(Remember/LOCQ)]*

(b) Briefly explain how SVM classify non linearly separable data points by using kernel function. *[(CO3)(Analyse/IOCQ)]*

6 + 6 = 12

Group - E

8. (a) Utilize different scales of measurement to design appropriate data collection formats for gathering information on a specific research question.
[(CO4)(Apply/IOCQ)]

(b) How does the use of heat map visualization in data science aid in identifying patterns or trends within large datasets, and what are the key considerations for effectively interpreting and analyzing heat maps? *[(CO4)(Analyse/IOCQ)]*

(c) Explain the difference between matplotlib and seaborn in terms of their approach to creating visualizations in Python. *[(CO4)(Understand /LOCQ)]*

4 + 4 + 4 = 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	33.3	47.9	18.8