

DATA SCIENCE (MCAP 2262)

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) The mean and variance of a binomial distribution with parameter $6, \frac{2}{3}$ are
 (a) $6, \frac{4}{3}$ (b) $4, \frac{4}{3}$ (c) $0, \frac{2}{3}$ (d) $9, \frac{4}{3}$
- (ii) 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) $\frac{2}{3}$
- (iii) The effectiveness of an SVM depends upon
 (a) Selection of a kernel trick (b) Kernel parameters
 (c) Soft margin parameter C (d) All of these
- (iv) Correlation coefficient is used in statistics to measure
 (a) How strong the relationship between two variables
 (b) How strong the relationship between two values
 (c) Experiments performed in the EDA process
 (d) Required in percentile based outlier removal.
- (v) Which of the algorithm is most sensitive to the outliers?
 (a) k-means clustering algorithm (b) k-medians clustering algorithm
 (c) k-modes clustering algorithm (d) k-medoids clustering algorithm.
- (vi) In binomial distribution when $n \rightarrow \infty$ 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.
- (vii) Which data acquisition method involves obtaining information directly from individuals or groups?
 (a) Web scraping (b) Sensor data collection
 (c) Machine learning modelling (d) Surveys and questionnaires.

- (viii) Why is data pre processing important in data science?
 (a) It ensures data is in a format suitable for analysis
 (b) It helps in reducing computational complexity
 (c) It improves the accuracy of machine learning models
 (d) All of the above.
- (ix) 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.
- (x) Which visualization type is most appropriate for comparing the relationship between two continuous variables?
 (a) Box plot (b) Line plot (c) Heat map (d) Bubble chart.

Fill in the blanks with the correct word

- (xi) K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest _____.
- (xii) Naive Bays Classifiers are a collection of _____ algorithms.
- (xiii) Scatter plots are ideal for visualizing the _____ relationship between two continuous variables.
- (xiv) BeautifulSoup and Scrapy are popular Python libraries used for _____ in web scraping projects.
- (xv) Data acquisition is crucial in data science as it lays the foundation for _____ and subsequent analysis to derive insights and make informed decisions.

Group - B

2. (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.

[[C04](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

3. (a) What are the primary considerations when selecting data sources for acquisition in a data science project, and how do they impact the subsequent analysis?
[[CO1](Analyze/IOCQ)]
- (b) How does the choice of data cleaning techniques impact the reliability and accuracy of insights derived from a dataset during the data wrangling process?
[[CO1,2](Remember/LOCQ)]
- (c) Suppose you have a dataset containing the ages of individuals from a certain population. You want to bin this data into age groups for analysis. The following are the age ranges and corresponding frequencies of individuals in each range:

Age Range	Frequency
0-10	25
11-20	45
21-30	60
31-40	55
41-50	40
51-60	30

Calculate the mean age for each binning range, rounded to the nearest integer. Then, find the overall mean age for this population dataset, also rounded to the nearest integer.

[[CO3](Apply/IOCQ)]

3 + 3 + 6 = 12

Group - C

4. (a) A random variable has the following p.m.f

X	0	1	2	3	4	5	6	7
P(X=x)	0	k	2k	2k	3k	K ²	2k ²	7k ² +k

(i) Determine k. (ii) Find $P(3 < X \leq 6)$.
[[CO4](Apply/IOCQ)]

- (b) A car hire firm has 2 cars which it hires out day by day. The number of demands for a car on each day is distributed as a Poisson distribution with average number of demands per day 1.5. Evaluate the proportion of days on which some demand is refused. [given $e^{-1.5} = 0.223$].
[[CO4,6](Apply/HOCQ)]

(4 + 4) + 4 = 12

5. (a) The length of the species of a fish is normally distributed with mean 5.3 cm and sd 2.5 cm. A fish is selected at random. Find the probability that a fish selected at random having length

(i) less than 4.5 cm. (ii) between 2.8 cm and 4.9 cm.
[[CO4,5](Apply/IOCQ)]

- (b) During a rainy season on a tropical island the length of a shower has an exponential distribution with average length of shower equal to $\frac{1}{2}$ minutes.

Calculate the probability that the shower will last more than 3 minutes. If it has already lasted for 2 minutes, find the probability that it will last for at least one more minute?

[[CO4,5](Apply/IOCQ)]

(3 + 3) + (3 + 3) = 12

Group - D

6. (a) Explain in short that how SVM can classify nonlinearly separable data points using kernel function? [[CO3](Understand/LOCQ)]
 (b) Explain the terms precision, recall and accuracy. [[CO3](Understand/LOCQ)]
6 + (2 + 2 + 2) = 12
7. (a) Illustrate the procedure to determine k using Elbow method. How the initial centroids in k means clustering is chosen? [[CO3](Understand/LOCQ)]
 (b) What is known as sigmoid function? Explain with example that why sigmoid function is used in Logistic regression? [[CO3](Remember/LOCQ)]
7 + 5 = 12

Group - E

8. (a) What are some common interpretations and insights that can be gained from analyzing the quartiles, median, and whiskers of a boxplot in data science? [[CO4](Understand/LOCQ)]
 (b) Given a dataset, which visualization method would be most appropriate for representing the geographical distribution of data? Justify your choice. [[CO3,4](Understand/LOCQ)]
 (c) Critique the use of colour in data visualization, discussing both its advantages and potential pitfalls. [[CO4,5](Evaluate/HOCQ)]
4 + 4 + 4 = 12
9. (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](Analyze/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

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	40.62	48.95	8.33

Course Outcome (CO):

After the completion of the course students will be able to

- MCA2262.1 Learn and understand how data is collected, managed and stored.
- MCA2262.2 Understand and describe the key concepts in data science and popular techniques used by data scientists.
- MCA2262.3 Build skills in data management.
- MCA2262.4 Demonstrate proficiency with statistical analysis of data.
- MCA2262.5 Develop ability to build and assess data-based models.
- MCA2262.6 Apply the concepts and algorithms to real-world data.

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