DATA ANALYTICS
(CSBS 4135)

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) What is the primary goal of data analytics?
(a) To gather data from the web
(b) To develop new data source
(c) To uncover insights and support decision-making
(d) To clean data.

(ii) What is the primary purpose of linear regression?
(a) To classify data into categories
(b) To predict a continuous outcome variable
(c) To reduce data dimensionality
(d) To assess data distribution.

(iii) What is the purpose of splitting a dataset into training and test datasets?
(a) To increase the size of the dataset
(b) To optimize data storage
(c) To evaluate the performance of a classification model
(d) To reduce the number of features.

(iv) What does pre-pruning in decision trees involve?
(a) Trimming branches after the tree has been built
(b) Limiting the growth of the tree during its construction
(c) Using cross-validation to prune the tree
(d) Combining multiple decision trees into one.

(v) In the K-nearest Neighbor algorithm, what does the parameter 'K' represent?
(a) The number of features in the dataset
(b) The number of nearest neighbors to consider for classification
(c) The number of decision trees in an ensemble
(d) The number of layers in a neural network.

(vi) How does PCA determine the principal components of a dataset?
(a) By clustering the data into K groups
(b) By calculating the eigenvectors and eigenvalues of the covariance matrix
(c) By applying a hierarchical clustering algorithm
(d) By using fuzzy logic to group data points.

(vii) In K-means clustering, how is the number of clusters determined?
(a) By the silhouette score (b) It is specified by the user
(c) By the distance metric (d) It is automatically calculated.

(viii) In which type of visualization does the color intensity represent the magnitude of values in a matrix format?
(a) Heat map (b) Bubble chart
(c) Gauge chart (d) Force Directed Chart.

(ix) What is the primary purpose of data visualization?
(a) To collect raw data
(b) To clean and preprocess data
(c) To present data in a graphical format that makes it easier to understand
(d) To analyze data using statistical methods.

(x) Which of the following is true about Neural Networks?
(a) They require feature engineering to perform well
(b) They consist of layers of interconnected nodes
(c) They always produce interpretable results
(d) They do not require a large amount of data.

Fill in the blanks with the correct word

(xi) ______________ Matrix is used for measuring classification accuracy.

(xii) __________ is a well-known dimension reduction technique.

(xiii) An example of activation function in ANN is _____________.

(xiv) Linear Regression equation of Y on X is ______________.

(xv) The _______ plot is commonly used to show the distribution of a numeric variable and identify potential outliers.

Group - B

2. (a) Briefly discuss structured data, unstructured data and semi structured data with example. *[(CO1)(Remember/LOCQ)]*

(b) Discuss feature extraction and feature reduction techniques in the light of feature engineering. *[(CO1)(Remember/LOCQ)]*

(c) How data analytics can be utilized in healthcare to enhance patient care and operational efficiency? *[(CO1)(Understand/LOCQ)]*

3 + 4 + 5 = 12

3. (a) Find linear regression equation for the following set of data.

x	2	4	6	8
y	3	7	5	10

[(CO2)(Apply/IOCQ)]

(b) Explain the Tanh activation function in ANN. *[(CO2)(Remember/LOCQ)]*

(c) Explain how does logistic regression differ from linear regression? *[(CO3)(Understand/LOCQ)]*

6 + 3 + 3 = 12

Group - C

4. (a) Given the following dataset predict whether a player should play or not if the weather is sunny using Naïve Bayes' classifier.

Observation No.	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rainy	No
9	Sunny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes

[(CO3)(Apply/IOCQ)]

(b) Discuss the impact of choosing the value of K in KNN classifier. *[(CO3)(Understand/LOCQ)]*

(c) What is an outlier? *[(CO2)(Remember/LOCQ)]*

7 + 3 + 2 = 12

5. (a) Why KNN classification algorithm is a lazy learning classifier? *[(CO3)(Understand/LOCQ)]*

(b) Write down the steps decision tree classification algorithm. *[(CO3)(Remember/LOCQ)]*

(c) Provide an example scenario where a decision tree would be an appropriate classifier. *[(CO3)(Understand/LOCQ)]*

3 + 6 + 3 = 12

Group - D

6. (a) Explain the concept of Fuzzy C-means clustering algorithm. *[(CO4)(Remember/LOCQ)]*

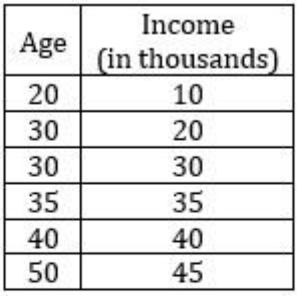
(b) How does Fuzzy C-means clustering differ from the K-means algorithm? *[(CO4)(Understand/LOCQ)]*

(c) Describe briefly the different linkage functions in hierarchical clustering algorithm. *[(CO4)(Remember/LOCQ)]*

4 + 4 + 4 = 12

7. (a) Write down the steps of PCA algorithm. *[(CO5)(Remember/LOCQ)]*

(b) Apply K-Means algorithm on the below given dataset to form two clusters (number of desired clusters K=2):



Age	Income (in thousands)
20	10
30	20
30	30
35	35
40	40
50	45

[(CO4)(Apply/IOCQ)]

(c) Describe briefly the functions of dendrogram in hierarchical clustering. *[(CO4)(Remember/LOCQ)]*

3 + 7 + 2 = 12

Group - E

8. (a) What are the common challenges in data visualization, and how can they be addressed to improve data interpretation? *[(CO4)(Remember/LOCQ)]*

(b) Explain Histogram and Box Plot with example. *[(CO6)(Remember/LOCQ)]*

(c) What is the difference between static and dynamic data visualizations? *[(CO6)(Understand/LOCQ)]*

4 + 5 + 3 = 12

9. (a) You are tasked with visualizing survey data that includes categorical and numerical variables. What visualization techniques would you use, and why? *[(CO6)(Understand/LOCQ)]*

(b) Explain Choropleth with example. *[(CO6)(Remember/LOCQ)]*

(c) What is a gauge chart, and when would it be appropriate to use it? *[(CO6)(Remember/LOCQ)]*

6 + 3 + 3 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	79.2	20.8	0