B.TECH/IT/6TH SEM/INFO 3202/2025

DATA ANALYTICS
(INFO 3202)

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) In Fuzzy C-Means clustering, what does the membership value represent?
(a) The distance from a data point to the cluster center.
(b) The likelihood of a data point belonging to a specific cluster.
(c) The number of clusters the data point belongs to.
(d) The number of iterations the algorithm needs to converge.

(ii) Which of the following algorithms is suitable for handling non-spherical clusters?
(a) K-Means (b) DBSCAN
(c) Hierarchical Clustering (d) K-Nearest Neighbors (KNN)

(iii) Your dataset is a mixture of two gaussian distributions. Identification of two independent Gaussian distributions can be obtained by
(a) Principal component analysis (b) Maximum likelihood method
(c) Expectation-Maximization algorithm (d) Linear regression

(iv) What is the main objective of a Decision Tree algorithm in classification tasks?
(a) To minimize the variance within each leaf node.
(b) To find the optimal decision boundary.
(c) To split the dataset into subsets based on the features.
(d) To maximize the overall classification error.

(v) A 300 mb data file is fragmented into 5 input splits. The number of mappers required will be
(a) 4 (b) 5 (c) 1 (d) Cannot be predicted

(vi) Which of the following is the primary function of Hadoop's HDFS (Hadoop Distributed File System)?
(a) To store large datasets across multiple machines
(b) To perform complex data processing tasks
(c) To query large datasets in real-time
(d) To manage and schedule Hadoop jobs

(vii) Which of the following Hadoop components is responsible for executing MapReduce tasks?
(a) DataNode (b) JobTracker
(c) TaskTracker (d) ResourceManager

(viii) MongoDB is an example of which type of NoSQL database?
(a) Key-Value store (b) Document-based store
(c) Graph database (d) Column-family store

(ix) Which of the following is true about MongoDB indexing?
(a) MongoDB does not support indexing.
(b) MongoDB supports indexing only on primary keys.
(c) MongoDB supports various types of indexes, including compound and geospatial indexes.
(d) MongoDB supports only full-text indexing.

(x) What is the "primary key" equivalent in MongoDB?
(a) Document ID (b) Collection Name
(c) Field Value (d) Index

Fill in the blanks with the correct word

(xi) In DBSCAN, points that are not core points and are not reachable from any other point are called ________.

(xii) In Naive Bayes, the likelihood of the data given a class is computed using ________ distribution for continuous features and a ________ distribution for categorical features.

(xiii) In K-Nearest Neighbors (KNN), the choice of the hyperparameter ________ determines how many nearest neighbors will be considered for classification.

(xiv) In the Perceptron algorithm, the ________ is the weighted sum of the inputs plus a bias term.

(xv) During the training process, a neural network updates its weights using the ________ algorithm, which aims to minimize the error in predictions.

Group - B

2. (a) Create clusters of the following spatial data objects using DBSCAN with minpts = 3 and Epsilon = 3.2.

O1	4	6
O2	7	4
O3	3	3
O4	5	5
O5	4	6
O6	5	7
O7	6	2
O8	6	6

[(CO1,CO3,CO6)(Create/HOCQ)]

(b) Discuss the advantages of DBSCAN over kmean's algorithm. *[(CO1) (Remember/LOCQ)]*

8 + 4 = 12

3. (a) State the limitations of K means clustering algorithm. *[(CO1,CO3) (Remember/LOCQ)]*

(b) Apply K-means clustering algorithm, to cluster the following data point represented in the form of (x_coordinate, y_coordinate). Consider A1, A5 as the initial cluster centroids. Update the centroids twice (i.e., iterate twice to update the centroids or stop if no difference between cluster centroids are achieved earlier)

Data Points: A1(7,5); A2(3,6); A3(15,15); A4(15,14); A5(8, 9); A6(3,7); A7(12,23); A8(3,6); A9(12,23) ; A10(17,27). *[(CO1, CO6) (Apply/IOCQ)]*

4 + 8 = 12

Group - C

4. (a) Construct a Decision Tree (with Gini Index metric) using the weekend spending data, as given in the following Table

Week End	Weather type	Humidity	Money Expended	Decision
Week1	Hot	High	500	Stay In
Week2	Cold	Low	2000	Shopping
Week3	Rainy	Low	1500	Restaurant
Week4	Rainy	High	500	Stay In
Week5	Hot	Low	2000	Restaurant
Week6	Cold	High	1500	Shopping
Week7	Hot	Low	2000	Shopping
Week8	Cold	Low	500	Restaurant
Week9	Cold	High	2000	shopping
Week10	Rainy	High	500	Stay In

[(CO2, CO3,CO6) (Apply/IOCQ)]

(b) Justify for or against the statement "If you neglect outliers while fitting a supervised machine learning model as classifier, you are overfitting your model". *[(CO3) (Evaluate/HOCQ)]*

9 + 3 = 12

5. (a) Consider the following table with predicted results obtained after applying 2 classification model on a dataset, having to two classes malignant (positive class) and benign (negative class). Calculate the TP, FP, TN, and FN of each model and next Compare the sensitivity of the two models and determine whose performance is better.

	Model 1 (Predicted Class)	Model 2 (Predicted Class)	Actual Class
Sample 1	+	+	+
Sample 2	-	+	+
Sample 3	-	-	+
Sample 4	+	-	+
Sample 5	+	-	_
Sample 6	-	+	-
Sample 7	+	-	-

[(CO3,CO6)(Evaluate/HOCQ)]

(b) Justify with example for or against the statement

"The concept of neighbors in higher dimension does not exist". *[(CO3,CO6)(Analyse/IOCQ)]*

7 + 5 = 12

Group - D

6. (a) Discuss overfitting problem with respect to supervised learning. *[(CO2,CO3,CO6)(Analyse/IOCQ)]*

(b) Justify for or against the statement "K-fold cross validation mitigates overfitting problem". What happens when K =N where N is the total number of data samples. *[(CO3,CO6)(Analyse/IOCQ)]*

(c) What do you understand by sensitivity as a performance metric? *[(CO3,CO6)(Analyse/IOCQ)]*

4 + 5 + 3 = 12

7. (a) Describe with the help of a diagram the architecture of Hadoop Distributed File System. *[(CO4)(Understand/LOCQ)]*

(b) Suppose you have a word file with the following text.
"The world is going through a huge crisis. God save the world.
The world is beautiful". The file size is 110 MB. Explain how the file gets broken into input splits and explain the overall steps in mapper and reducer. *[(CO4,CO6) (Demonstrate/HOCQ)]*

5 + 7 = 12

Group - E

8. (a) Explain the advantages as well as the limitations of NO-SQL databases with respect to SQL based databases. *[(CO5) (Understand/LOCQ)]*

(b) Describe the data model followed by a Document oriented database, MongoDb, with the help of an example. *[(CO5) (Understand/LOCQ)]*

5 + 7 = 12

9. (a) Write a MongoDB query to do the following:

(i) insert the following document into a collection called students:

```
{
  "name": "John Doe",
  "age": 22,
  "subjects": ["Math", "Physics"],
  "GPA": 3.8
}
```

(ii) How would you retrieve all documents from the students collection where age > 30?

(iii) Write a query to update the **GPA** of "John Doe" to **4.0**.

(iv) How would you delete a document where name = "Jane Doe" from a collection? *[(CO5)(Evaluate/HOCQ)]*

(b) How do **sharding and replication** work in NoSQL databases? *[(CO5)(Understand/LOCQ)]*

(2 + 2 + 2 + 2) + 4 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	30	35	35