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) A density based clustering algorithm is
(a) PAM (b) STIRR (c) ROCK (d) DBSCAN

(ii) Which of the following is finally produced by Hierarchical Clustering?
(a) Final estimate of cluster centroids
(b) Tree showing how close things are to each other
(c) Assignment of each point to clusters
(d) All of the mentioned.

(iii) With respect to clustering technique which of the following statement is / are true.
Statement 1- K means algorithm is good in handling outliers
Statement 2- Fuzzy C Means does not require the value of number of clusters, as initial parameter
Statement 3- Border Points are those which have at least one neighbour which is core object
(a) 1 & 3 (b) All (c) 1 & 2 (d) Only 3

(iv) In decision tree C4.5 uses which parameter for selecting the best splitting attribute.
(a) Information Gain (b) Gain Ratio
(c) Gain Matrix (d) Gini Index

(v) Which of the following is required by K-means clustering?
(a) Defined distance metric (b) Number of clusters
(c) Initial guess as to cluster centroids (d) All of the mentioned.

(vi) In hadoop ecosystem data is stored as 64/128 bytes of block size in
(a) HDFS (b) HIL files (c) Both (d) None of the above

(vii) The expected value or _____ of a random variable is the center of its distribution.
(a) mode (b) median (c) mean (d) bayesian inference

(viii) Which of the following characteristic of big data is relatively more concerned to data analytics?
(a) Velocity (b) Variety
(c) Volume (d) None of the mentioned.

(ix) Categorical attribute that can be ordered is
(a) Ordinal attribute (b) Nominal attribute
(c) Boolean Attribute (d) All of these

(x) Which of the following is example of hard clustering?
(a) Fuzzy C means (b) K means
(c) A means (d) T means.

Fill in the blanks with the correct word

(xi) The sum squared error in kmeans is determined by _____ method.

(xii) The _____ represents an outcome where the model correctly predicts the positive class.

(xiii) In Naïve Bayes class with highest _____ probability is the class of an unknown dataset.

(xiv) _____ is an example of document based NoSQL database.

(xv) The task tracker in Hadoop is a _____ service.

Group - B

2. (a) Clusters of the following spatial data objects using DBSCAN with minimum number of neighbours required to become a core object is 2 and Epsilon is 2.2

01	3	4
02	6	5
03	2	4
04	8	7
05	19	16
06	15	17
07	4	3
08	18	16

[(CO1,CO3)(Apply/IOCQ)]

(b) What are core, border and noise objects in DBSCAN algorithm

[(CO1)(Understand/LOCQ)]

8 + 4 = 12

3. (a) State the objective function of K means clustering algorithm.

[(CO1)(Understand/LOCQ)]

(b) A textile company in New York state, USA, must decrease expenses by minimizing delivery costs. One way to do that is to relocate warehouses closer to their distributors. The company employs 10 distributors across the state of New York. The following demonstration simulates how an operations manager could segment distributors into **two clustered** geographies using the **KMeans function** and then identify two optimal warehouse locations central to the identified centroid function. Update the centroids twice (i.e., iterate twice to update the centroids or stop if no difference between cluster centroids are

achieved earlier). Distributors locations are as follows. Consider distributor **D1** and distributor **D10** as initial centroids.
D1(17,15); D2(13,26); D3(15,16); D4(15,14); D5(8, 9); D6(13,17);
D7(22,23); D8(3,6); D9(22,26) ; D10(4,6) *[(CO1,CO3)(Evaluate/HOCQ)]*

4 + 8 = 12

Group - C

4. (a) The following dataset consists of four features age, income level, whether the person has permanent job. The class label is whether the person is a defaulter with respect to loan or not. Construct a decision tree using Information Gain. Show only 2 levels of the tree.

Age	Income	Permanent Job	Loan -Defaulter
youth	Low	no	Yes
youth	Low	yes	No
Middle-aged	Low	yes	No
Middle-aged	High	yes	No
Middle-aged	High	no	Yes
Old	High	yes	No
Old	High	no	Yes
Youth	High	yes	No
Youth	High	no	Yes

[(CO2,CO3,CO6)(Evaluate/HOCQ)]

(b) Explain the principle of Fuzzy C means clustering technique.

[(CO1)(Understand/LOCQ)]

8 + 4 = 12

5. (a) Explain working principle of Naive Bayes classification technique.

[(CO2)(Understand/LOCQ)]

(b) Justify the following in for or against the statement:
(i) KMeans algorithm performs poorly in the presence of outliers
(ii) ID3 has been modified in C4.5 classification technique
(iii) ROCK clustering technique can work on Categorical attribute
(iv) DBSCAN can handle outliers efficiently. *[(CO3)(Analyse/IOCQ)]*

4 + 8 = 12

Group - D

6. (a) Consider a library of an University. Apply a MapReduce based technique to find out the count of each unique words present in the books corresponding to "Design Analysis of Algorithms, by dividing the text in books into input splits, and provided to the mappers as input. Explain each step of execution with the help of a diagram, keeping in mind that a mapper takes key value as input and generates key value as output. *[(CO5,CO6)(Analyse/IOCQ)]*

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

6 + 6 = 12

7. (a) Consider a blood management system which has classified its donor as compatible or non-compatible donors. The prediction results obtained by two classification model on six samples is provided in the table below. 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	Yes	Yes	Yes
Sample 2	No	No	Yes
Sample 3	Yes	Yes	Yes
Sample 4	Yes	No	Yes
Sample 5	No	No	No
Sample 6	Yes	Yes	No
Sample 7	yes	No	No

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

(b) Explain the steps how in Fuzzy C means the membership matrix gets updated. *[(CO1)(Understand/LOCQ)]*

(c) State hadoop's master services and slave services and the relationship among them. *[(CO4)(Understand/LOCQ)]*

6 + 4 + 2 = 12

Group - E

8. (a) Describe the characteristics of NoSQL databases. *[(CO5)(Remember/LOCQ)]*

(b) Describe the benefits of NoSQL databases. *[(CO5)(Remember/LOCQ)]*

(c) Describe the following terms in the context of HBase architecture.
(i) Master server (ii) Region server (iii) Zookeeper. *[(CO5) (Remember/LOCQ)]*

4 + 4 + 4 = 12

9. (a) Write down the MongoDB commands for the following operations.
(i) Creating Database
(ii) Creating Collection
(iii) Finding Documents in a Collection
(iv) Adding Documents to a Collection. *[(CO5)(Apply/IOCQ)]*

(b) Compare the concept of normalizing data with document references and denormalizing data with embedded documents in the context of MongoDB database. *[(CO5)(Analyse/IOCQ)]*

(c) Describe the concept of capped collections in the context of MongoDB database. *[(CO5)(Understand/LOCQ)]*

4 + 4 + 4 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	43.75	31.25	22.9

Course Outcome (CO):

After the completion of the course students will be able to

1. Apply the different clustering algorithms to cluster real life datasets.
2. Apply appropriate classification algorithm to classify an unknown dataset.
3. Analyze the performance of the Clustering or Classification Algorithms.
4. Identify the need of Big Data Paradigms, and will be able to Store and Process Data on Hadoop Distributed File System.
5. Identify the need of No-SQL Databases and be able to Convert Relational Model to different No-SQL Data Models.
6. Create Appropriate Classifiers or Clustering Models for Analyses of Big Data using Hadoop Eco System.

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