M.C.A. 3rd Semester, (MCA 2 Year Programme) w.e.f. 2021-2022 Examination, November–2023

DATA MINING & BIG DATA ANALYSIS

Paper–21MCA23Cl

Time allowed : 3 hours] *[Maximum marks : 80*

Note: ***Question No.1 is compulsory.*** *In addition to Question No. 1, attempt* ***four*** *more questions, by selecting at least* ***one*** *question from each Unit. All questions carry equal marks.*

1. (i) Brief out data mining task primitives must be followed while framing mining query.

(ii) Formulate any two dissimilarity measures available in data mining.

(iii) Mention any four measures taken to implement clustering approaches on high dimensional data.

(iv) Draw the model used for classifying the students's data into urban and rural classes. Assume the data attributes.

(v) Give two examples each for structured and unstructured digital data.

(vi) What do you mean by heartbeat signal generated in Hadoop environment?

(vii) Enlist any four advantages of using Map reduce as an integral component of Hadoop.

(viii) Data Analysts prefer to work with Hive. Justify.

[P. T. O.

Unit-I

2. (i) Define concept hierarchy and its types. Elaborate how concept hierarchy can be used in data reduction with appropriate example?

   (ii) How supervised discretization works differently from unsupervised discretization in data mining task?

3. (i) How appropriate sampling of data can improve the quality of data mining results? Differentiate between Simple Random Sampling without replacement and Simple Random Sampling with replacement.

   (ii) Elaborate how efficient and scalable frequent itemsets can be mined from the given data set by using Apriori algorithm? Explain importance of prune step in improving its efficiency.

Unit-II

4. (i) What important role does bias value and weights play in Multilayer Feed forward neural network model? Explain.

   (ii) Write down different measures that can be used evaluating classifiers along with their formula and examples.

5. (i) How centroid based technique i.e. k-Means method is used for clustering the given data by applying it recursively?

(ii) Describe how DBSCAN method can be used to create arbitrary shaped clusters by using density reachable and density connected points.

Unit-III

6. (i) Detail out the architecture followed by Hadoop for handling a user query and roles performed by Job tracker and Task tracker in handling the query.

   (ii) YARN was introduced in Hadoop 2.0 to enhance the performance of Mapreduce. Comment.

7. (i) Explain how data is compressed and copied from local to HDFS and from HDFS to local in Hadoop environment.

   (ii) For proper Fault Tolerance System, Rack awareness is very important? Support your answer.

Unit-IV

8. (i) Explain the sorting, shuffling and spilling of data carried out in Map reduce phase (examples).

   (ii) Elaborate the Map Reduce types and formats with an appropriate scenario.

9. (i) What all makes Hadoop ecosystem? Brief out working of any five such components.

   (ii) Explain different components of Pig environment and how Pig coding is converted into Mapreduce.