Roll No. .....

67101-N

MCA 3rd Semester (MCA 2 Year Programme) w.e.f. 2021-2022 Examination – December, 2024

DATA MINING & BIG DATA ANALYSIS

Paper : 21MCA23C1

Time : Three hours] *[Maximum Marks : 80*

Before answering the questions, candidates should ensure that they have been supplied the correct and complete question paper. No complaint in this regard, will be entertained after examination.

Note : Attempt ***five*** questions in all, selecting ***one*** question from each Unit. Question No. **1** is ***compulsory***. All questions carry equal marks.

1. (a) Define Data Mining Systems along with its core components.

(b) State Apriori principle followed in Association mining.

(c) Define supervised and unsupervised learning ? Give ***one*** example for each.

(d) List down any ***four*** desirable characteristics of clustering methods.

(e) Explain the variety and veracity characteristics of Big data.

(f) Differentiate between Hadoop environment and Hadoop ecosystem.

(g) Briefly explain the working of YARN, as a heart of complete Hadoop ecosystem.

(h) Write down any *four* features of Big R making it most appropriate choice for handling big data.

UNIT – I

2. (a) How orientation, focus, type of users, schema used, size of query and data operations differs OLTP from OLAP ?

(b) Define Data Objects. Differentiate between different types of data objects along with *two* examples for each.

3. (a) Elaborate the concept of Candidate set generation (Apriori algorithm). Performance evaluators and frequent sequential patterns related to Association mining.

(b) What do you mean by association rules ? Explain different types of association rules generated after implementing Pattern mining.

UNIT – II

4. (a) Elaborate the working of Decision Tree Induction algorithm, making it greedy and non back tracking classifier.

   (b) Explain how different classifiers can be evaluated by using accuracy and specificity by using Confusion matrix.

5. (a) Describe how K Mean (Partitioning) Clustering detects the outliers and create clusters.

   (b) Differentiate between DIANA and AGNES techniques following hierarchical concepts to generate clusters with the help of labeled diagrams.

UNIT – III

6. (a) Explain how Structured, Semi structured and Unstructured data along with their respective characteristics and examples (*One* for each type).

   (b) Elaborate different phases of Big data analytics life cycle, making it appropriate for batch as well as stream processing.

7. (a) Differentiate between data analytics performed by using Hadoop and different Unix tools.

   (b) Elaborate how data compression and serialization plays a vital role in big data analytics with appropriate examples.

UNIT – IV

8. (a) Describe the following concepts : Job scheduling and shuffling with respect to Map Reduce with their respective importance in processing big data.

   (b) Why Pig is considered as an integral component of Hadoop ecosystem ? Explain the same in terms of its architecture.

9. (a) Elaborate the architecture of HBase and what important role does Region servers play in it.

   (b) Define collaborative filtering and explain its types. How it enhance the data analytics process ?