MASTER OF COMPUTER APPLICATIONS (MCA-NEW)

Term-End Examination

December, 2022

MCS-226 : DATA SCIENCE AND BIG DATA

Time : 3 hours *Maximum Marks : 100*

Weightage : 70%

Note : *Question no.* ***1*** *is* ***compulsory*** *and carries 40 marks. Attempt any* ***three*** *questions from the rest.*

1. (a) Explain the following types of data : *6*

(i) Semi-structured data

(ii) Unstructured data

(iii) Qualitative data

(iv) Quantitative data

(b) What is meant by "Probability distribution of continuous random variable" ? Explain with the help of a diagram. Also explain the normal distribution. *6*

(c) What are the characteristics of Hadoop Distributed File System (HDFS) ? Why is it used for Big data processing ? *6*

(d) Explain the characteristics of data streams. *4*

(e) What are NoSQL databases ? Why are they used ? *4*

(f) Explain any one mechanism of filtering of data streams. *4*

(g) Explain the following, with the help of an example, in the context of R programming : *6*

(i) Dataframe

(ii) List

(iii) Vector

(h) What is logistic regression ? Which function of R programming can be used to implement logistic regression ? *4*

2. (a) Explain the characteristics of measurement scales of data. Use these characteristics to define various measurement scales of data. *6*

(b) Explain the steps of significance testing, with the help of an example. *8*

(c) Explain the following terms with the help of an example : *6*

(i) Data pre-processing

(ii) Data curation

(iii) Data cleaning

3. (a) Explain the characteristics of Big data. How does Big data differ from relational data ? *6*

(b) Explain the steps of map-reduce paradigm using the example of word counting. *6*

(c) List the features of any ***two*** of the following : *8*

(i) Apache Spark

(ii) Hive

(iii) Column-based databases

(iv) Graph-based databases

4. (a) How can link analysis be used to compute PageRank ? *4*

(b) Explain the concept of Recommendation System. *6*

(c) Explain how the similarity between two documents can be found. *6*

(d) Explain how the social networks can be represented using a graph. *4*

5. (a) Write an R program to create two 3×3 matrices and multiply them. How is this program different from a similar C program ? *5*

(b) What is a box plot ? List the commands of R programming that can be used to create a box plot. *5*

(c) What is multiple regression ? Write steps about how R programming can be used to create multiple regression model. *5*

(d) What is a decision tree ? Write steps on how R programming can be used for making decision tree. *5*