No. of Printed Pages : 6

MCS–226

MASTER OF COMPUTER APPLICATIONS (MCA-NEW)

Term-End Examination

December, 2023

MCS-226 : DATA SCIENCE AND BIG DATA

Time : 3 Hours *Maximum Marks : 100*

Weightage : 70%

Note : *Question No. 1 is compulsory. Attempt any* ***three*** *questions from the rest.*

---

1. (a) How does sampling differ from population ? Also, discuss the relation of the terms 'statistic' and 'parameter', with sampling and population, respectively. 5

(b) What is conditional probability ? Write the equation for conditional probability and describe its components with a suitable example. 5

(c) What is a BoxPlot ? What are whiskers in any BoxPlot ? Briefly discuss the utility of BoxPlots in Data Science. 5

(d) What is MapReduce ? Explain the Map function and Reduce function in the Map-Reduce architecture with a suitable block diagram. 5

(e) What is Apache Spark ? Give main features of Apache Spark framework. 5

(f) Differentiate between Data Stream Management System (DSMS) and Data Base Management System (DBMS). 5

(g) Explain the term 'Link Analysis'. Briefly discuss the purpose of link analysis in data science. How link analysis can be used for World Wide Web (WWW) ? 5

(h) Briefly describe the following data structures of 'R' programming with the help of an example for each : 5

(i) Vector

(ii) List

2. (a) What is Data Cleaning ? List and briefly discuss the best practices used for data cleaning and data preparation. 5

(b) What is Logistic Regression ? Give its utility. Also, write and discuss the function of 'R' used to construct the model for logistic regression. 5

(c) What is a Random Forest ? How does it differ from Decision tree ? Explain the role of partitioning, pruning and tree selection process in the working of decision tree. 10

3. (a) Explain the following types of data, used in data science. Give suitable example for each : 10

(i) Structured data

(ii) Semi-structured data

(iii) Unstructured data

(iv) Data streams

(b) What is Binomial Distribution ? Write the formula for binomial probability distribution and use it to produce the probability distribution for the case of three tosses of the coin. 5

(c) What is a scatter plot ? Give uses and best practices for scatter plot. 5

4. (a) Explain word count problem with suitable example. Give pseudo-code for word count problem in MapReduce. 5

(b) Briefly discuss the purpose of the following components of Apache Spark : 10

(i) Spark core

(ii) Spark SQL

(iii) Spark Streaming

(iv) MLib

(v) Graph X

(c) Compare Ad-hoc Queries and Standing Queries of data streams. Give example for each. 5

5. Write short notes on the following : 4×5=20

(a) Sensitive PageRank

(b) Time series analysis

(c) Clustering and its types in 'R' programming

(d) Data Science life cycle

(e) HBase and its utility in Data Science