

No. of Printed Pages : 5

MCS-226

**MASTER OF COMPUTER
APPLICATIONS
(MCA-NEW)**

Term-End Examination

June, 2025

**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) What are the different types of data in Data Science ? Briefly explain each type. 5

D-3036/MCS-226

P. T. O.

- (b) Explain conditional probability with equation and suitable example. 5
- (c) What is a Euclidean distance measure ? How does it differ from cosine distance ? 5
- (d) Explain Data Streams. Justify the statement, “Data Stream is a challenging task in Data Science”. 5
- (e) What is Apache Spark ? What are the features of Apache Spark that differ from Hadoop ? 5
- (f) Explain NoSQL. What are the differences between RDBMS and NoSQL. 5
- (g) What are the factors in R programming ? Give characteristics of factors. 5

- (h) What is JSON File in R ? How to convert JSON into a data frame ? 5
2. (a) Differentiate between Big Data and Data Warehouse, with suitable explanation for each. 5
- (b) What is data cleaning ? What are the methods of data cleaning ? 5
- (c) What do you mean by Box Plot ? Explain clearly how the Box plot differs from Scatter plot. What is the utility of Box Plot in Data Science ? Explain in detail. 10
3. (a) What is Big Data ? What are the characteristics of Big Data ? 5

- (b) What is HDFS ? Write steps to load data into HDFS format. 5
- (c) What do you understand by the term ‘Finding Similar Documents’ ? What are the various concepts of document similarity analysis ? Compare Minhashing and locality sensitive hashing for document similarity with suitable illustrations. 10
4. (a) Write a code in R programming to perform concatenation of the following three strings : 5
- “Helo”, “,” “Learning is Fun”
- (b) What do you understand by HIVE ? Explain the components of HIVE architecture with diagram. 5

- (c) What is Link Spamming ? Illustrate link spam with a suitable example. What are the possible solutions to combat link spamming ? 10
5. Write short notes on the following : 5×4=20
- (a) Linear Regression
 - (b) Support Vector Machines
 - (c) Vector in R programming
 - (d) Time Series Analysis
 - (e) Partitioning *vs.* Pruning

× × × × ×