

No. of Printed Pages : 6

MCS-226

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

Term-End Examination

December, 2024

**MCS-226 : DATA SCIENCE
AND BIG DATA**

Time : 3 Hours

Maximum Marks : 100

Weightage : 70%

Note : (i) *Question No. 1 is compulsory and carries 40 marks.*

(ii) *Attempt any **three** questions from the remaining Q. Nos. 2 to 5.*

1. (a) Explain data science with the help of its applications. Also, discuss how structured data is different from semi-structured data. 6

- (b) What is Bayes theorem ? Explain Bayes theorem with the help of an example. 6
- (c) What do you mean by 'Big Data Analysis' ? Explain big data processing using spark ecosystem. 6
- (d) Discuss why data preprocessing is important in data science and big data applications with the help of suitable diagram. Also discuss different phases of data preprocessing. 6
- (e) Explain the term 'Distributed File System' in the context of big data. Also explain the different features of distributed file system. 6
- (f) Explain different types of data structures in R-language. Which function of R-programming can be used to implement linear regression ? Explain linear regression using R-language. 6

(g) What is the use of scatter plot ? How can you draw a scatter plot using R-programming language ? 4

2. (a) What are the *two* measures to define the central tendencies of quantitative data ? Explain with the help of an example. Also, discuss different measures to define the spread or variability of observed quantitative data with the help of examples. 8

(b) Explain the following terms with the help of an example : 8

(i) Sampling

(ii) Dredging

(iii) Simpson's paradox

(iv) Histograms

- (c) What is the use of pair plot ? Explain how do you read a pair plot. 4
3. (a) Explain, how Master/Slave process works in HDFS architecture. Also, differentiate between Apache Hadoop-1 and Hadoop-2 using suitable diagram. 8
- (b) What is key-value pair based NoSQL ? List the benefits of key-value pair based NoSQL. Explain when to use key-value NoSQL database with the help of an example. 6
- (c) Explain the spider trap and dead-end problem in PageRank. Discuss the solutions for the spider trap and dead-end problem. 6
4. (a) What is the purpose of a distance measure ? Differentiate between cosine distance and edit distance with the help of an example. 6

- (b) What is a recommender system ? Discuss the process of content-based recommendations using suitable diagram.

6

- (c) What do you mean by data stream processing ? Which model of data stream processing is useful in finding stock market trends ? Justify your answer.

8

5. (a) Explain the following types of graphs used (using syntax) for visualization in R-programming language :

6

(i) Bar-charts

(ii) Box-plots

(iii) Line-graphs

- (b) What is Logistic Regression ? Write steps about how R-programming can be used to create logistic regression model.

6

- (c) Explain, where do we use random forests algorithm. Write the pseudo-code for random forests algorithm. Also, write steps on how R-programming can be used making a decision tree. 8

x x x x x x x