M.Sc.(IT)(Sem-II) July-2023

Information Technology : Paper I - Big Data Analytics (R2020)

(Time: 2 hours) **[Total Marks: 60]**

N. B.: (1) **All** questions are **compulsory**.
(2) Make **suitable assumptions** wherever necessary and **state the assumptions** made.
(3) Answers to the **same question** must be **written together**.
(4) Numbers to the **right** indicate **marks**.
(5) Draw **neat labeled diagrams** wherever **necessary**.
(6) Use of a **Non-programmable** calculator is **allowed**.

Q.1 **Attempt *any two* of the following:** **12**

a What are the challenges with big data? Explain in short.

b Explain the Classification of Analytics with respect to the Second School of Thought.

c What are the requirements of technologies to meet the challenges of big data? Also, explain the Responsibilities of Data Scientists.

d What are the different phases of the Data Analytics Lifecycle? Explain each in detail with a neat diagram.

Q.2 **Attempt *any two* of the following:** **12**

a What is K-means clustering? Describe the steps to find k clusters using the k-means algorithm.

b What is the role of support in the apriori algorithm? Also, explain how the Apriori property works with a neat diagram.

c What is Linear regression? Explain in detail. Also, explain any two of its use cases.

d Apply the Ordinary Least Squares (OLS) technique to estimate the parameters of the linear regression model with a neat diagram.

Q.3 **Attempt *any two* of the following:** **12**

a How to predict whether customers will buy a product or not? Explain with respect to the decision tree.

b Explain a probabilistic classification method based on Naive Bayes' theorem.

c What is the critical problem in using the Term frequency? How can it be fixed?

d What is sentiment analysis? How it can be carried out? Explain it in detail.

Q.4 **Attempt *any two* of the following:** **12**

a How to refactor the data science pipeline into an iterative model? Explain all its phases with a neat diagram.

b Write a short note on Hadoop Distributed File System.

c Write a short note on job chaining with a neat diagram.

d Write in brief about Spark. Also, write and explain its primary components.

Q.5 **Attempt *any two* of the following:** **12**

a Write in brief about the design pattern. Explain each of its categories.

b Write the entire procedure with appropriate commands for importing data from MySQL to Hive.

c Explain Spark SQL interface architecture with a neat diagram.

d Which different types of filters can be used in HBase? Explain its entire procedure with appropriate commands.
