(2 ½ Hours) **[Total Marks: 60]**

N.B: (1) **All questions are compulsory.**
(2) Figures to the **right** indicate full marks.
(3) **Assume additional data if necessary** but state the same clearly.
(4) Symbols have their usual meanings and tables have their usual standard design unless stated otherwise.
(5) Use of **calculators** and statistical tables are **allowed**. / If required keep it.

Q.1 Attempt **any two** of the following **(12)**

a) What is data reduction? Describe one of the technique of data reduction with an example. **6**

b) What is data science? What are the characteristics of data in data science? **6**

c) Explain Jacquard similarity measures with an example. **6**

d) Explain the challenges faced in handling Big data. **6**

Q.2 Attempt **any two** of the following **(12)**

a) Describe the process of reading a file from HDFS. **6**

b) Write a note on the following commands:- **6**
i. cp
ii. mv
iii. appendToFile

c) Write a note on the following related to mapreduce framework: **6**
i. Mapper
ii. Reducer
iii. Partitioner

d) Illustrate the architecture of Hadoop with its components with the help of a diagram. **6**

Q.3 Attempt **any two** of the following **(12)**

a) Explain the linear regression equation, and state how it is used for prediction in supervised machine learning. **6**

b) Write short notes on: **6**
i. Multicollinearity
ii. Durbin-watson test
iii. Heteroskedascity

c) Write short notes on the following: **6**
Over fitting and under fitting

d) Consider the following table and calculate the regression coefficient of the Linear regression equation if x is internal_Exam and Final_score is the target variable y. Calculate the value of slope or regression coefficient. **6**

internal_Exam	Final_score
7	40.79
0	69.23
1	76.75
8.5	75.66
9.5	55.48
3	67.11
8	67.98
16	85.09

Q.4 Attempt **any two** of the following **(12)**

a) What is clustering? Distinguish between clustering and classification. **6**

b) Distinguish between eager learner and lazy learner. How KNN algorithm is used in the classification. **6**

c) What is confusion matrix? Explain the following terms related to confusion matrix: **6**
i. Precision
ii. Recall

d) Write a short note on Hierarchical clustering method. **6**

Q.5 Attempt **any two** of the following **(12)**

a) Describe the different types of variations in time series with appropriate examples. **6**

b) Describe the semi averaging technique for measuring the underlying trend. **6**

c) What is sentiment analysis? Describe their applications. **6**

d) Write short notes on:= **6**
1. Stemming
2. POS tagging

―――――

DC04480B59A19F5C1B754C7AE818C368