K. K. Wagh Institute of Engineering Education & Research, Nashik

(An Autonomous Institute From A.Y. 2022-23)

InSem Examination-IISummer2025	
Exam Seat No.:	
Academic Year:2024-2025	Semester:VI
Class:TY	Program:B.Tech
Branch Code:COM/CSD	Pattern:2022
Name of Course:Data Science and Big Data	Course Code:COM223011
Max. Marks:30	Duration:1.15 Hrs.

Instructions: Candidates should read carefully the instructions printed on the Question Paper and on the cover page of the Answer Book, which is provided for their use.

1. This question paper contains 02 pages.
2. Answer to each new question is to be started on a new page.
3. Assume suitable data wherever required, but justify it.
4. Draw the neat labelled diagrams, wherever necessary.
5. The last columns indicates the Course Outcome and level of Blooms Taxonomy of the Question/sub-question.

Marks CO

Question No. 1

1 a) What is the need of data preprocessing? Explain any two techniques involved in data preprocessing. (4) CO1

1 b) What are outliers? Explain any two techniques to detect outliers. (3) CO1

Question No. 2

2 a) Consider following data (in increasing order) for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. answer the following: (8) CO1

(i) Use min-max normalization to transform the value 25 for age onto the range [0.0, 1.0].

(ii) Use z-score standardization to transform the value 25 for age, where the standard deviation of age is 12.94 years.

(iii) Use normalization by decimal scaling to transform the value 35 for age.

OR

2 b) What is a linear binary pattern? Find the LBP for the given matrix. (8) CO1

40	70	40
80	30	60
10	50	10

Question No. 3

3 a) What is Overfitting and Underfitting? What are the problems that could arise in machine learning due to it. (4) CO2

3 b) Differentiate between linear and logistic regression. Provide real-world examples where each is applicable (3) CO2

Question No. 4

4 a) The following table records the number of balls that Rujay took for scoring runs in different matches. In how many balls is he likely to score a century? (use Linear Regression) (8) CO2

Runs Scored(x)	**8**	**35**	**47**	**54**	**11**	**85**	**84**	**93**	**89**	**2**
Number of Balls(y)	**10**	**20**	**31**	**23**	**5**	**47**	**35**	**67**	**73**	**1**

OR

4 b) Explain the following regression evaluation metrics. (8) CO2

Mean Absolute Error (MAE)

Mean Squared Error (MSE)

Root Mean Squared Error (RMSE)

R-squared (Coefficient of Determination)

····· **End of question paper**·····