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

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

WINTER-2025	
Exam Seat No.:	
Academic Year:2025-2026	Semester:V
Class:TY	Program:B.Tech
Branch Code:CSD	Pattern:2023
Name of Course:Data Science and Big Data	Course Code:2313306B
Max. Marks:60	Duration:2.30 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 page(s).
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			
1a)	Explain how Principal Component Analysis (PCA) helps in reducing the number of features in a dataset.	(6)	CO1
Question No. 2			
2a)	Write a short note on Ridge regression.	(6)	CO2
Question No. 3			
3a)	Describe the main idea behind the Support Vector Machine (SVM) algorithm. How does it find the best decision boundary?	(8)	CO3
	OR		
3b)	Explain the following evaluation metrics used in classification problems and discuss when each is most appropriate to use: 1.Accuracy 2. Precision 3.Recall 4.F1-Score	(8)	CO3
3c)	Differentiate between binary and multiclass classification problems. Provide examples from real-world applications and discuss how model design varies.	(8)	CO3
	OR		
3d)	Explain the concept of balanced and imbalanced multiclass classification. Discuss the impact of class imbalance on model performance with suitable examples.	(8)	CO3
Question No. 4			
4a)	Discuss how the choice of 'K' affects the performance of K-Means clustering. What methods help determine the optimal K?	(8)	CO4

OR

4b)	Compare agglomerative and divisive hierarchical clustering in terms of process and computational efficiency.	(8)	CO4
4c)	Explain the importance of evaluating clustering quality. What metrics are commonly used?	(8)	CO4

OR

4d)	Explain the concept of cluster analysis and its role in unsupervised learning. Use examples to illustrate its applications.	(8)	CO4

Question No. 5

5a)	Discuss the factors contributing to the exponential growth of data in recent years.	(8)	CO5

OR

5b)	Explain the five key characteristics of Big Data (The 5 V's)	(8)	CO5
5c)	Describe the key stages of the Data Analytics Lifecycle and their interdependencies.	(8)	CO5

OR

5d)	Identify and explain major sources of Big Data in today's digital ecosystem.	(8)	CO5

····· End of question paper·····