JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -3 EXAMINATION- 2025

B. Tech-I Semester (CSE)

COURSE CODE (CREDITS): 22M11CI112 (3) MAX. MARKS: 35

COURSE NAME: INTRODUCTION TO DATA SCIENCE

COURSE INSTRUCTORS: Dr Nancy Singla MAX. TIME: 2 Hours

***Note:*** *(a) All questions are compulsory.*

*(b) The candidate is allowed to make suitable numeric assumptions wherever required for solving problems*

*(c) Use of calculator is allowed.*



| Q. No | Question | CO | Marks |
|---|---|---|---|
| Q1. | (a) Explain the bias–variance tradeoff in supervised learning. Why can a model with very low training error still perform poorly on unseen data?<br>(b) Which of the following classifiers will be preferred while working with small and large datasets. Explain.<br>• High bias/low variance classifiers<br>• Low bias/high variance classifiers | CO5 | [5+5] |
| Q2. | (a) For the given data, compute two clusters using K-means algorithm for clustering where initial cluster centers are (1.0, 1.0) and (5.0, 7.0). Execute for two iterations.<br>(b) Explain how would you choose the number of clusters in k-means?<br>(c) Describe a real-world scenario where k-means clustering would be useful? | CO6 | [5+3+2] |
| Q3. | (a) You are given a data set consisting of variables having more than 30% missing values. Let us say, out of 50 variables, 8 variables have missing values higher than 30%. How will you deal with them?<br>(b) For each of the following scenarios, state which probability distribution is most suitable and give a brief reason.<br>i. A call center receives an average of 12 calls per hour. You want to model the number of calls in the next hour.<br>ii. The heights of adult males in a large population are being | CO3 | [2+3] |

| Record Number | A | B |
|---|---|---|
| R1 | 1.0 | 1.0 |
| R2 | 1.5 | 2.0 |
| R3 | 3.0 | 4.0 |
| R4 | 5.0 | 7.0 |
| R5 | 3.5 | 5.0 |
| R6 | 4.5 | 5.0 |
| R7 | 3.5 | 4.5 |

| | | | |
|---|---|---|---|
| | analyzed.<br>iii. A company sends promotional emails. Each email has a 10% chance of being opened. You want to model the number of opens out of 200 emails. | | |
| Q4. | You are given the following five training instances:<br>• $x_1 = 2, x_2 = 1, y = 4$<br>• $x_1 = 6, x_2 = 3, y = 2$<br>• $x_1 = 2, x_2 = 5, y = 2$<br>• $x_1 = 6, x_2 = 7, y = 3$<br>• $x_1 = 10, x_2 = 7, y = 3$<br>We want to model this function using the K-nearest neighbor regressor model. Predict the value of y corresponding to (x1, x2) = (3, 6) considering k as 2 and 3. | CO6 | [5] |
| Q5. | A dataset contains product reviews stored as string objects. Describe a preprocessing pipeline to prepare this data for machine learning classification using a BoW model. Explain why each step is necessary. | CO5 | [5] |