JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST-2 EXAMINATION- 2025

B.Tech-VII Semester (CSE/IT)

COURSE CODE (CREDITS): 19B1WCI731 (2) MAX. MARKS: 25

COURSE NAME: Computational Data Analysis

COURSE INSTRUCTORS: Ekta Gandotra MAX. TIME: 1 Hour 30 Min

Note: (a) All questions are compulsory.
(b) The candidate is allowed to make Suitable numeric assumptions wherever required for solving problems
(c) Calculator is allowed.

<table>
<tr><th>Q. No.</th><th>Question</th><th>CO</th><th>Marks</th></tr>
<tr><td>Q1.</td><td>Consider the following dataset:
<table>
<tr><th>Employee_ID</th><th>Department</th><th>Project Completed (Yes/No)</th></tr>
<tr><td>E101</td><td>Sales</td><td>Yes</td></tr>
<tr><td>E102</td><td>HR</td><td>No</td></tr>
<tr><td>E103</td><td>IT</td><td>Yes</td></tr>
<tr><td>E104</td><td>Sales</td><td>No</td></tr>
<tr><td>E105</td><td>IT</td><td>Yes</td></tr>
</table>
a. Calculate the Information Gain of the Department feature with respect to the target variable Project Completed.
b. Evaluate why selecting Employee_ID as the top feature based on Information Gain may not be appropriate for predicting whether an employee completed a project.</td><td>4</td><td>3
2</td></tr>
<tr><td>Q2.</td><td>A random sample of 30 students was surveyed, and each student was asked whether they attended a coaching class. The results are summarized below:
<table>
<tr><th>Attended Coaching</th><th colspan="2">Passed Exam</th></tr>
<tr><th></th><th>Yes</th><th>No</th></tr>
<tr><td>Yes</td><td>12</td><td>6</td></tr>
<tr><td>No</td><td>4</td><td>8</td></tr>
</table>
Apply the Chi-Square (χ^2) Test of Independence to check whether attending coaching is significantly associated with passing the exam. (Note: The critical value of χ^2 with 1 degree of freedom is 3.841 at 5% level of significance).</td><td>4</td><td>5</td></tr>
</table>

<table>
<tr><td>Q3.</td><td>Consider the following two-dimensional dataset consisting of four points:
(2,0), (0,1), (3,4), (5,2)

a. Compute the first principal component (PC1) of this dataset using the PCA algorithm.
b. Project the given points onto PC1.
c. Calculate the proportion of total variance explained by PC1 and PC2.</td><td>4</td><td>

2
2
2</td></tr>
<tr><td>Q4.</td><td>Consider the following dataset. Apply the K-Nearest Neighbor algorithm with K = 3 to predict the rent of a house having size of 1600 Sq. Ft. and 3 occupants. Use Euclidean distance as distance metric. Perform any necessary preparation of the features before computing distances.

<table>
<tr><th>Size (Sq. Ft.)</th><th>Occupants</th><th>Rent (₹/month)</th></tr>
<tr><td>550</td><td>1</td><td>8000</td></tr>
<tr><td>750</td><td>2</td><td>10000</td></tr>
<tr><td>1200</td><td>3</td><td>15000</td></tr>
<tr><td>2000</td><td>5</td><td>23000</td></tr>
<tr><td>1800</td><td>4</td><td>21000</td></tr>
<tr><td>950</td><td>2</td><td>12000</td></tr>
</table></td><td>2</td><td>5</td></tr>
<tr><td>Q5.</td><td>a. What is the purpose of the kernel function in SVM? How does it help the model to handle non-linearly separable data?
b. Give two limitations of using kernel functions in SVM.
c. Explain how logistic regression handles categorical data. Illustrate your answer with an example.</td><td>2</td><td>1

1
2</td></tr>
</table>