|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Roll No. |  |  |  |  |  |  |  |  |  |  |  |  |



**PRESIDENCY UNIVERSITY**

**Bengaluru**

|  |
| --- |
| **End - Term Examinations – JANUARY 2025** |
| **Date:** 11 – 01- 2025 **Time:** 01:00 pm – 04:00 pm |

|  |  |  |
| --- | --- | --- |
| **School:** SOIS | **Program:** BCA | |
| **Course Code :** CSA2007 | **Course Name :** Data Mining | |
| **Semester**: III | **Max Marks**: 100 | **Weightage**:50% |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CO - Levels** | **CO1** | **CO2** | **CO3** | **CO4** | **CO5** |
| **Marks** | **24** | **262** | **24** | **26** |  |

**Instructions:**

1. *Read all questions carefully and answer accordingly.*
2. *Do not write anything on the question paper other than roll number.*

**Part A**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Answer ALL the Questions. Each question carries 2marks. 10Q x 2M=20M** | | | | |
| **1** | Outline 4 applications of data mining. | **2 Marks** | **L1** | **CO1** |
| **2** | State whether data mining is misnomer? Justify your answer. | **2 Marks** | **L1** | **CO1** |
| **3** | List properties of attributes with suitable example. | **2 Marks** | **L1** | **CO2** |
| **4** | Compute Smoothing bin by means for the given data: 3, 6, 9, 12, 14, 17, 19, 22, 25, 28, 32, 35. Let Bin Size: 4 | **2 Marks** | **L2** | **CO2** |
| **5** | Recite the methods for handling noisy data. | **2 Marks** | **L1** | **CO2** |
| **6** | State the differences between FP growth and Apriori method | **2 Marks** | **L1** | **CO3** |
| **7** | A dataset has 7 transactions, and the item set {A,B} appears in 3 transactions. Compute the support for {A,B}? | **2 Marks** | **L2** | **CO3** |
| **8** | Describe dendogram with a suitable diagram? | **2 Marks** | **L1** | **CO4** |
| **9** | Differentiate between intra-cluster and inter-cluster with an example. | **2 Marks** | **L2** | **CO4** |
| **10** | List 4 common causes of outliers in a data set. | **2 Marks** | **L1** | **CO4** |

**Part B**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Answer the Questions Total 80 Marks.** | | | | | |
| **11.** | **a.**  **b.** | Elucidate 10 applications of data mining.  Elucidate the functionalities of data mining. | **10 Marks**  **10 Marks** | **L2**  **L2** | **CO1**  **CO1** |
| **or** | | | | | |
| **12.** | **a.**  **b.** | What types of data can be mined using data mining? Defend.  Explain the procedures of data mining as a process of knowledge discovery of data bases with a neat diagram. | **10 Marks**  **10 Marks** | **L2**  **L2** | **CO1**  **CO1** |
|  |  |  |  |  |  |
| **13.** | **a.**  **b.** | Entropy Based Discretization to find the best split for the following data. (0,Y), (4,Y), (12,Y), (16,N), (16,N), (18,Y), (24,N), (26,N), (28,N). If S has to be permitted into 2 intervals ‘S1’ and ‘S2’ using two split point 14 and 21. Interpret the Best Split Point.  Suppose two stocks X and y have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14). If the stocks are affected by the same industry trends, will their prices rise or fall together? Justify using covariance. | **10 Marks**  **10 Marks** | **L3**  **L3** | **CO2**  **CO2** |
| **or** | | | | | |
| **14.** | **a.**  **b.** | **Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):**  (a) Compute the Euclidean distance.  (b) Compute the Manhattan distance.  (c) Compute the Minkowski distance, using h = 3  Rachel told Eric that the reason her car insurance is less expensive is that female drivers get in fewer accidents than male drivers. Specifically, she says that male drivers are held responsible in 65% of accidents involving drivers under 23. If Eric does some research of his own and discovers that 46 out of the 85 accidents he investigates involve male drivers, Compute his data support or refute Rachel’s hypothesis? Assume level of significance as 0.05.  **Note: DF(1,0.05)=3.841; DF(2,0.05)=5.991.** | **10 Marks**  **10 Marks** | **L2**  **L2** | **CO2**  **CO2** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **15.** | **a.** | Database consists of nine transactions taken from the Electronic store. Enumerate all the frequent itemsets using Apriori algorithm with minimum support threshold S=2 and minimum confidence = 70%. | **20 Marks** | **L3** | **CO3** |
| **Or** | | | | | |
| **16.** | **a.** | A retail company wants to analyze the purchasing patterns of its customers to find out the frequently bought items together. The company has a dataset of transactions, where each transaction contains a list of products bought by a customer. Apply FP- Growth algorithm to identify frequent itemsets. Assume min support =3 | **20 Marks** | **L3** | **CO3** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **17.** | **a.** | An e-commerce company is analyzing customer data to understand whether a product will be "Accepted" or not based on certain attributes. Apply ID3 algorithm to construct Decision Tree. The dataset below contains information about the customers: | **20 Marks** | **L3** | **CO4** |
| **Or** | | | | | |
| **18.** | **a.**  **b.** | Classify different types of Outliers with a suitable diagram.  An online store wants to group products into two clusters based on their prices: low-priced products and high-priced products. Apply K-Means Clustering algorithm on the given data: 2,3,4,10,11,12,20,25,30  Assume Initial Clusters: 4, 12 | **10 Marks**  **10 Marks** | **L2**  **L3** | **CO4**  **CO4** |

**\*\*\*\*\* BEST WISHES \*\*\*\*\***