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

 ****

**Presidency University**

**Bengaluru**

 **School Of Computer Science and Engineering & Information Science**

**Summer Term Examinations, Aug 2024**

**Date**: 06.08.2024

**Time**: 1.00-4.00pm

**Max Marks**: 100

**Weightage**: 50%

**Odd Semester**: 2023 - 24

**Course Code**: CSE2021

**Course Name**: DATA MINING

**Department: CSE**

 **Instructions:**

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Q.No** | **Questions** | **Marks** | **CO** | **RBT** |
| 1 | 1. Define data mining and state how it differs from data base querying.
 | 4 | CO1 | L1 |
| 1. Differentiate the two data mining tasks namely classification and regression with a suitable examples for each
 | 6 | CO1 | L2 |
| 1. Explain the various stages involved in the process of KDD with a neat illustration
 | 10 | CO1 | L3 |
| OR |
| 2 | 1. List any 4 applications of data mining tasks
 | 4 | CO1 | L1 |
| 1. Describe any 3 issues faced by data mining tasks.
 | 6 | CO1 | L2 |
| 1. Explain the various types of data mining tasks with an example for each.
 | 10 | CO1 | L3 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 3 | 1. Differentiate the two attribute types namely nominal and ordinal with a suitable example for each.
 | 4 | CO2 | L1 |
| 1. Explain the distance measures namely i)Euclidean and ii)Manhattan
 | 6 | CO2 | L2 |
| 1. Transform the values of income [12000, 20000, 25000, 30000, 45000, 60000, 73600, and 98000] into the range of 0 to 1 using the normalization techniques namely, min-max and decimal scaling. Write the disadvantage(s) of min-max normalization.
 | 10 | CO2 | L3 |

OR

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 4 | 1. Differentiate missing and noisy data with an example for each.
 | 4 | CO2 | L1 |
| 1. Explain the two normalization techniques namely min-max and decimal scaling.
 | 6 | CO2 | L2 |
| 1. Explain the two discretization methods namely i) equal width and ii)equal frequency binning. The observed values of ‘Age’ in a database is between 0 to 100. Discretize this attribute into 5 bins using equal width binning. Write the bin intervals.
 | 10 | CO2 | L3 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 5 | 1. What is a transaction data base? Give an example for the same.
 | 4 | CO3 | L1 |
| 1. Explain the two steps involved in generating association rules from a transaction data base.
 | 6 | CO3 | L2 |
| 1. Explain the candidate generation method used by Apriori algorithm. Also explain the Pseudo-code the rule generation algorithm.
 | 10 | CO3 | L3 |

OR

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 6 | 1. What is the anti-monotone/downward closure property of an itemset?
 | 4 | CO3 | L1 |
| 1. Define the two metrics namely support and confidence of association rules with an example for each.
 | 6 | CO3 | L2 |
| 1. For the Transaction Database given below, find the frequent itemsets using Apriori Algorithm with min\_support count of 2 .

|  |  |
| --- | --- |
| Transaction ID | Items |
| T100 | 1,3,4 |
| T200 | 2,3,5 |
| T300 | 1,2,3,5 |
| T400 | 2,5 |

  | 10 | CO3 | L3 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 7 | 1. Define training, validation and test data sets with respect to classification.
 | 4 | CO4 | L1 |
| 1. Define the three measures of node impurity.
 | 6 | CO4 | L2 |
| 1. Using Gini Index as the measure of node impurity, find the Gini of the split on “Owns Home”, and “Married” and hence the best splitting node using the following data set.

|  |  |  |
| --- | --- | --- |
| Owns Home | Married | Risk Class |
| YES | YES | B |
| NO | NO | A |
| YES | YES | C |
| YES | NO | B |
| NO | YES | C |
| NO | NO | A |
| NO | NO | B |
| YES | NO | A |
| NO | YES | C |
| YES | YES | C |

 | 10 | CO4 | L3 |

OR

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 8 | 1. Write the minimum and maximum values of Gini index for a data set with ‘n’ number of classes.
 | 4 | CO4 | L1 |
| 1. Define confusion matrix and ROC curve of a classifier.
 | 6 | CO4 | L2 |
| 1. Find the best split (between a1 and a2) in the following data set, according to gini index.

 | 10 | CO4 | L3 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 9 | 1. Define the two objectives of a good clustering technique.
 | 4 | CO5 | L1 |
| 1. What are the requirements of a good clustering algorithm?
 | 6 | CO5 | L2 |
| 1. The weight and pH of four different medicines are given in the table below. Using k-means clustering, group these medicines into 2 clusters with medicine A and medicine B as the initial cluster centroids. **Show the detailed working , the cluster membership and the cluster centroids after iteration 1 only**.

|  |  |  |
| --- | --- | --- |
| **Medicine** | **Weight** | **pH** |
| A | 1 | 1 |
| B | 2 | 1 |
| C | 4 | 3 |
| D | 5 | 4 |

  | 10 | CO5 | L3 |

OR

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 10 | 1. List the various clustering approaches.
 | 4 | CO5 | L1 |
| 1. What are the requirements of clustering in data mining?
 | 6 | CO5 | L2 |
| 1. The petal\_length and petal\_width of four different flower species are given in the table below. Using k-means clustering, group these species into 2 clusters with F1 and F2 as the initial cluster centroids. **Show the detailed working , the cluster membership and the cluster centroids after iteration 1 only**.

|  |  |  |
| --- | --- | --- |
| **Flower** | **Petal\_length (in cm)** | **Petal\_Width****(in cm)** |
| F1 | 7 | 5 |
| F2 | 12 | 3 |
| F3 | 10 | 8 |
| F4 | 5 | 2 |

  | 10 | CO5 | L3 |