Sardar Vallabhbhai National Institute of Technology, Surat

Computer Science and Engineering Department

End Semester Examination, Nov. – Dec. 2022

[U19CS012]

B. Tech. IV (CSE) – 7th Semester

Course: Data Warehousing and Mining (CS441) (20 mks → 30 mins)

Date:15th December 2022 Time:14.00 hrs to 17.00hrs Max Marks: 100

Instructions: 1.Please start the answer to each question on new page ONLY of your answer sheets.
2. Please write your correct exam no without fail on the answer sheets as well as the question papers.

Q. 1 Answer the following questions [Any Three]: (3 mks each) **[18]**

1. For all the data mining techniques, write the data mining applications forinternet search engine company.

integrity correctness Believability
A Consistency Accuracy BICAT A Authenticity

2. Mention the data quality issues. For anyone of the data quality issue, explain the reasons of occurring and how to deal with it.
3. A teacher wants to use association analysis to analyze test results of 100 students. The test consists of 100 questions with four possible answers each. How would you convert this answer data into a form of data suitable for ANY ONE data mining technique?
4. Explain one of the measures other than the support and confidence which can improve the importance of association rules.

Q. 2 Answer the following questions [Any Two]: (2.5 mks / each) **[14]**

1. Find the five number summary for the following data set:
   10,11,12,25,25,27,31,33,34,34,35,36,43,50,59
   Draw a box plot and histogram plot for the given data set and explain which one is the advantageous.
2. Following is the frequency table for Attribute A1 of the dataset. What will be the Entropy of attribute A1?

	A1=1	A1=2	A1=3	Row Sum
Class=1	2	1	1	4
Class=2	2	2	1	5
Class=3	4	5	6	15
Column Sum	8	8	8	24

3. Plot a Dendrogram using single linkage Agglomerative clustering for the following data.

min

	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.71	0				
P3	5.66	4.95	0			
P4	3.61	2.92	2.24	0		
P5	4.24	3.54	1.41	1.00	0	
P6	3.20	2.50	2.50	0.50	1.12	0

Σ activa
f(x)

Q. 3 Answer the following questions [Any Three]: (3 mks each) **[18]**

1. Compare the advantages and disadvantages of eager classification (e.g., decision tree, Bayesian, neural network) versus lazy classification (e.g. k-nearest neighbor)
2. What is Posterior probability and prior probability? Explain in detail how it is used to predict the class label.
3. Explain classification with algorithm Back propagation. MSE ↓ activa bias
4. What are the limitations of K-means clustering? Explain Clustering Using Representatives method to overcome above limitations.
5. Why is it important to consider density when clustering a dataset? How does DBSCAN use different points to cluster the dataset?

Q. 4 Answer the following questions: [20]

1. For the given dataset, Apply the naive bayes algorithm and predict the outcome for the car where color=Red, Type = SUV, Origin = Domestic

Color	Type	Origin	Stolen?
red	sports	domestic	yes
red	sports	domestic	no
red	sports	domestic	yes
yellow	sports	domestic	no
yellow	sports	imported	yes
yellow	SUV	imported	no
yellow	SUV	imported	yes
yellow	SUV	domestic	no
red	SUV	imported	no
red	sports	domestic	yes

2. Find clusters of given objects in a data set using k-medoids clustering algorithm for single iteration. Take the value of K=2 and O2 and O8 as initial medoids.

Data Points	A1	A2
O1	2	6
~~O2~~	~~3~~	~~4~~
O3	3	8
O4	4	7
O5	6	2
O6	6	4
O7	7	3
~~O8~~	~~7~~	~~4~~
O9	8	5
O10	7	6

Q. 5 Answer the following questions: [30]

(15)(15)

1. For the given transactions, do as per the instructions:

T1: a1, a2, a3, ..., a99, a100
T2: a51, a52, a53, ..., a99, a100
T3: a1, a2, a3, ..., a99, a100
T4: a1, a2, a3, ..., a50

(2.5) a. Justify the statement: The introduction/truncation of null transactions does not affect the support and confidence.

b. For absolute support>=3, discover the closed and maximal item sets.

c. For absolute support=1, discover the 1-frequent items and largest possible k-frequent items, and also mention the equation that can show possible candidates.

2. Consider the multinationale-commerce companywhich manages the data up to the city level of product sales.

a. Prepare the data warehouse schema and explain the selection of your schema with its advantages.

b. Write the OLAP query to find the country level data from the city level cuboids. Also mention the SQL query for the same.

c. Show the concept hierarchy for your schema.Calculate the count of cubes for your concept hierarchy.

*