University of Mumbai

Examinations Summer 2022

Program: **Computer Engineering**

Curriculum Scheme: Rev 2016 (Choice Base Credit Grading System)

Examination: TE Semester VI

Course Name: Data Warehousing and Mining (Paper code- 88903)

Time: 2-hour 30 minutes Max. Marks: 80

Q1.	**Choose the correct option for following questions. All the Questions are compulsory and carry equal marks**
1.	Correcting the customer flat number is
Option A:	Type1 change
Option B:	Type 2 change
Option C:	Type 3 change
Option D:	Type 4 change
2.	A company would like to improve its sales by analyzing its past data. Which of the following tasks will occupy maximum time required for the return on investment?
Option A:	identifying sources of data
Option B:	ETL process
Option C:	data analysis
Option D:	preparing reports
3.	Which of the application is not a data mining application?
Option A:	fraud detection
Option B:	Intrusion detection
Option C:	Customer segmentation
Option D:	IRCTC query
4.	Suppose you have a variable, economic status with three categories (low, medium and high). This variable falls under which of the following types ?
Option A:	Categorical
Option B:	Ordinal
Option C:	Binary
Option D:	Continuous
5.	Fields like various flags and textual fields that were left in the original data structures are referred to as _____
Option A:	Spam dimensions
Option B:	Miscellaneous dimensions
Option C:	Junk dimensions
Option D:	Snowflake dimensions
6.	The most granular fact table yields _____, which are pre calculated summaries.
Option A:	flags
Option B:	multiway tables
Option C:	groups
Option D:	aggregates

7.	You are given reviews of food quality of few restaurants as Good, Average or Poor. Finding reviews of a new restaurant is an example of
Option A:	Classification
Option B:	Regression
Option C:	Clustering
Option D:	Association mining
8.	Which of the following is the process of fetching all the web pages linked to a web site?
Option A:	Indexing
Option B:	Crawling
Option C:	Processing
Option D:	Linking
9.	---------- finds groups that are very different from each other, but whose members are similar to each other
Option A:	Clustering
Option B:	Grouping
Option C:	Stratified
Option D:	Classification
10.	Find the Inter-Quartile Range (IQR) of the data set {3, 7, 8, 5, 12, 14, 21, 13, 18}.
Option A:	6
Option B:	12
Option C:	**10**
Option D:	16

Question 2	**Solve any Two out of Three**	**10 marks each**
A	1	Consider a data warehouse for order analysis application where there are three dimensions 1. Customer 2.Order date and 3.Product. Consider a measure: quantity sold. Create a cube and illustrate the following OLAP operations. 1. Roll up 2. Drill down 3. Slice 4. Dice and 5. Pivot
B	2	The table below shows the six data points. Apply Agglomerative clustering to find clusters. Use Euclidian distance measure. Consider single linkage algorithm.

	X	y
D_1	0.4	0.53
D_2	0.22	0.38
D_3	0.35	0.32
D_4	0.26	0.19
D_5	0.08	0.41
D_6	0.45	0.30

C — 3 — Consider the following dataset S, which contains observations of several cases of sunburn:

Name	Hair	Height	Weight	Dublin	Result
Sarah	Blonde	Average	Light	No	Sunburned
Dana	Blonde	Tall	Average	Yes	None
Alex	Brown	Short	Average	Yes	None
Annie	Blonde	Short	Average	No	Sunburned
Emily	Red	Average	Heavy	No	Sunburned
Pete	Brown	Tall	Heavy	No	None

John	Brown	Average	Heavy	No	None
Katie	Brown	Short	Light	Yes	None

Unseen sample X= <brown,tall,average,No> Predict the result value as sunburned or None.

Question 3 **Solve any Two out of Three** **10 marks each**

A

Suppose we want ID3 to evaluate car database as whether the car is acceptable or not. The target classification is "Should we accept car?" which can be acceptable or unacceptable.

Buying_Price	Maintenance_Price	Lug_Boot	Safety	Evaluation?
High	High	Small	High	Unacceptable
High	High	Small	Low	Unacceptable
Medium	High	Small	High	Acceptable
Low	Medium	Small	High	Acceptable
Low	Low	Big	High	Acceptable
Low	Low	Big	Low	Unacceptable
Medium	Low	Big	Low	Acceptable
High	Medium	Small	High	Unacceptable
High	Low	Big	High	Acceptable
Low	Medium	Big	High	Acceptable
High	Medium	Big	Low	Acceptable
Medium	Medium	Small	Low	Acceptable
Medium	High	Big	High	Acceptable
Low	Medium	Small	Low	Unacceptable

B

A database has four transactions .Let min sup =60% and min conf=80%.

TID	Date	Items purchased
T100	21/04/2022	{K,A,D,B}
T200	21/04/2022	{D,A,C,E,B}
T300	22/04/2022	{C,A,B,E}
T400	23/04/2022	{B,A,D}

Find all the frequent item sets using apriori algorithm and also list all the strong association rules.

C

What is web structure mining? List the approaches used to structure the web pages to improve on the effectiveness of search engines and crawlers. Explain page rank technique in detail

Question 4	Solve any Four Questions out of Five 5 marks each
A	What are the different modes of the data loading process? Explain in brief.
B	Differentiate between OLTP and OLAP.
C	Explain web usage mining in detail.
D	What are the various methods for estimating classifiers accuracy.
E	Explain k-means clustering algorithm. Suppose the data for clustering is {2,4,10,12,3,20,30,11,25} consider k=2, cluster the given data using above algorithm.