TE(COMP) / SEM V / R-19 / FH-22 / DWM / 10.06.22

QP CODE: 91926

University of Mumbai

Examinations Summer 2022

Examinations Commencing from 17th May 2021 to _____

Program: **_Computer Engineering_____**

Curriculum Scheme: Rev2019

Examination: TE Semester: V

Course Code: **CSC504** and Course Name: **Data Warehousing and Mining**

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.	For the given attribute marks values: 35,45,50,55,60,65,75. Identify the first quartile and third quartile of data.
Option A:	35, 50
Option B:	35,75
Option C:	45,65
Option D:	45,60
2.	OLTP is not
Option A:	Operational database
Option B:	Subject oriented
Option C:	For continuous updates from many sources
Option D:	For high read, write update delete activity
3.	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
4.	In Banking scenario following dimension tables are identified. Point out the inappropriate one.
Option A:	Account
Option B:	Branch
Option C:	Time
Option D:	Transaction
5.	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
6.	Suppose we have the following values for salary (in thousands of dollars),

	in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. The mid range of data would be:
Option A:	50,000
Option B:	70,000
Option C:	77,000
Option D:	52,000
7.	Consider the data transaction given below:T1:{F,A,D,B} T2:{D,A,C,E,B} T3:{C,A,B,E} T4:{B,A,D} With minimum support = 60% and the minimum confidence = 80% , which of the following is not valid association rule?
Option A:	A -----> B
Option B:	B ----->A
Option C:	D -----> A
Option D:	A -----> D
8.	Given the Confusion matrix , the accuracy is

Classes	YES	NO
YES	90	210
NO	140	9560

Option A:	60%
Option B:	100%
Option C:	96.5%
Option D:	35%
9.	Data collection is done by crawling through number of web pages in
Option A:	Data Mining
Option B:	Web mining
Option C:	Text Mining
Option D:	Spatial data mining
10.	What is a Dendrogram?
Option A:	A tree diagram used to illustrate the arrangement of clusters in hierarchical clustering
Option B:	A tree diagram used to illustrate the arrangement of clusters in partitional clustering
Option C:	A type of hierarchical clustering
Option D:	A type of bar chart diagram to visualize K-means clusters.

Question 2	Solve any Two out of Three 10 marks each
A	The college wants to record the marks for the courses completed by students using the dimensions: a) Course b) Student c) Time and a measure of Aggregate marks. Create a cube and describe following operations: i) roll up ii) Drill down iii) Slice iv) Dice
B	Discuss the different steps involved in data preprocessing.
C	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

The table below shows the six data points. Apply Agglomerative clustering to find clusters. Use Euclidian distance measure. consider single linkage.

	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

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 Six 5 marks each
A	Explain major issues in data mining.
B	Differentiate between OLTP and OLAP.
C	Explain web usage mining in detail.
D	What are the various methods for estimating classifiers accuracy.
E	Use k means algorithm to create 3-clusters for given set of values:{2,3,6,8,9,12,15,18,22}
F	What are the various Issues regarding Classification and Prediction?

2/2