University of Mumbai

Examination June 2021

Examinations Commencing from 1st June 2021

Program: Computer Engineering

Curriculum Scheme: Rev2016

Examination: TE Semester VI

Course Code: CSC603and Course Name: Data Warehousing and Mining

Time: 2 hour Max. Marks: 80

===

===

Q1.	**Choose the correct option for following questions. All the Questions are compulsory and carry equal marks**
1.	The purpose of the operational system is used to______
Option A:	Run the business in real time and is based on historical data
Option B:	Takes strategic decisions for business
Option C:	Support decision making and is based on historical data
Option D:	Run the business in real time and is based on current data
2.	Which of following describes a data warehouse well?
Option A:	Can be updated by end users.
Option B:	Contains numerous naming conventions and formats.
Option C:	Organized around important subject areas.
Option D:	Contains only current data
3.	Expected amount of information (in bits) needed to assign a class to a randomly drawn object is _________
Option A:	Gain ratio
Option B:	Gini Index
Option C:	Entropy
Option D:	Information Gain
4.	Which of the following achieves data reduction by detecting redundant attributes
Option A:	Data cube aggregation
Option B:	Dimension reduction
Option C:	Data compression
Option D:	Numerosity reduction
5.	The fraudulent usage of credit card-scan be detected using data mining task should be used
Option A:	Prediction
Option B:	Outlier analysis
Option C:	Association analysis
Option D:	Correlation
6.	Given the record of users and movies viewed. Using Jaccard similarity measures, find similarity between {A-B,A-C,B-C }

Users	Movie 1	Movie 2	Movie 3	Movie 4	movie 5
A	1	0	1	0	1
B	0	0	1	0	1
C	0	1	0	0	1

Option A:	{0.67,0.25,0.33}
Option B:	{0.67,0.33,0.25}
Option C:	{0.5,0.33,0.67}
Option D:	{0.5,0.25,0.67}
7.	Five-number summary of a distribution (Minimum, Q1, Median, Q3, Maximum) is displayed by--------------
Option A:	Histogram
Option B:	quantile plot
Option C:	Scatterplot
Option D:	Box plot
8.	If a set is a frequent set and no superset of this set is a frequent set, then it is called _________.
Option A:	maximal frequent set
Option B:	border set
Option C:	lattice
Option D:	infrequent sets
9.	________ is a mining task that examines the web and hyperlinks structure that connect web pages.
Option A:	Web content mining
Option B:	Web structure mining
Option C:	Web usage mining
Option D:	Web link mining
10.	What does Web content mining involve?
Option A:	analyzing the universal resource locator in Web pages
Option B:	analyzing the unstructured content of Web pages
Option C:	analyzing the pattern of visits to a Web site
Option D:	analyzing the PageRank and other metadata of a Web page
11.	A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the suffix pattern is called as
Option A:	Suffix path
Option B:	FP-tree
Option C:	Prefix path
Option D:	Condition pattern base
12.	In star schema, there is one fact table as F1 is connected with four-dimension

	tables D1, D2, D3, D4 then fact table will have how many foreign keys?
Option A:	2
Option B:	4
Option C:	3
Option D:	5
13.	If Mean salary is 54,000 Rs. and standard deviation is 16,000 Rs. then find z score value of 73,600 Rs. salary
Option A:	1.225
Option B:	0.351
Option C:	1.671
Option D:	1.862
14.	The generalization of cross-tab which is represented visually is _____________ which is also called as data cube.
Option A:	Two-dimensional cube
Option B:	Multidimensional cube
Option C:	N-dimensional cube
Option D:	Cuboid
15.	In KDD and Data mining, noise is referred to as
Option A:	Complex data
Option B:	Meta data
Option C:	Error
Option D:	Repeated data
16.	Find the IQR of the data set {3, 7, 8, 5, 12, 14, 21, 13, 18}.
Option A:	6
Option B:	12
Option C:	16
Option D:	10
17.	Which of the following is not a method to estimate a classifier's accuracy
Option A:	Holdout method
Option B:	Random Sampling
Option C:	Information Gain
Option D:	Bootstrap

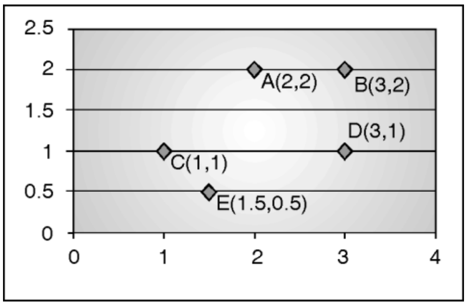
18.	For questions given below consider the data Transactions :
	T1 {F, A, D, B}
	T2 {D, A, C, E, B}
	T3 {C, A, B, E}
	T4 {B, A, D}
	With minimum support is 60% and the minimum confidence is 80%. Which of the following is not valid association rule?

Option A:	A -> B
Option B:	B -> A
Option C:	D -> A
Option D:	A -> D

19.	To calculate distance between two isotheticrectangles, ___________________is efficient approach and produces cluster of high quality
Option A:	CLARA
Option B:	PAM
Option C:	Spatial mining
Option D:	IR Approximation
20.	Geographers typically model the world with objects located at different places on surface of the earth. Through ________model, the real word entities are represented by lines, points and polygons
Option A:	Vector data model
Option B:	Raster data model
Option C:	Network data model
Option D:	Topology data model

<table>
<tr><td><b>Q2</b></td><td><b>Solve any Four out of Six5 marks each</b></td></tr>
<tr><td>A</td><td><i>Consider Metadata as an equivalent of Amazon book store, where each data element is book. What this meta data will contain. Explain.</i></td></tr>
<tr><td>B</td><td><i>Suppose a group of sales price records has been sorted as follows: 6, 9, 12, 13, 15, 25, 50, 70, 72, 92, 204, 232. Partition them into three bins by equal-frequency (Equi-depth) partitioning method. Perform data smoothing by bin mean.</i></td></tr>
<tr><td>C</td><td><i>Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order):
13, 15, 16, 16, 19, 20, 23, 29, 35, 41, 44, 53, 62, 69, 72
Use min-max normalization to transform the value 45 for age onto the range
[0:0, 1:0].</i></td></tr>
<tr><td>D</td><td><i>Use K-means algorithm to create 3 - clusters for given set of values:
{2, 3, 6, 8, 9, 12, 15, 18, 22}</i></td></tr>
<tr><td>E</td><td><i>Transaction database is given Below. Min Support = 2. Draw FP-Tree.</i>
<table>
<tr><th>TID</th><th>List of item_Ids</th></tr>
<tr><td>T100</td><td>I1, I2, I5</td></tr>
<tr><td>T200</td><td>I2, I4</td></tr>
<tr><td>T300</td><td>I2, I3</td></tr>
<tr><td>T400</td><td>I1, I2, I4</td></tr>
<tr><td>T500</td><td>I1, I3</td></tr>
<tr><td>T600</td><td>I2, I3</td></tr>
<tr><td>T700</td><td>I1, I3</td></tr>
<tr><td>T800</td><td>I1, I2, I3, I5</td></tr>
<tr><td>T900</td><td>I1, I2, I3</td></tr>
</table></td></tr>
<tr><td>F</td><td><i>Write short note on Spatial Clustering Techniques : CLARANS .</i></td></tr>
<tr><td><b>Q3</b></td><td><b>Solve any Two Questions out of Three 10 marks each</b></td></tr>
<tr><td>A</td><td><i>For a Supermarket Chain consider the following dimensions, namely Product, store, time , promotion. The schema contains a central fact tables sales facts with three measures unit_sales, dollars_sales and dollar_cost.</i></td></tr>
</table>

<table>
<tr><td></td><td><em>Design star schema and calculate the maximum number of base fact table records for the values given below :
Time period : 5 years
Store : 300 stores reporting daily sales
Product : 40,000 products in each store(about 4000 sell in each store daily)
Promotion : a sold item may be in only one promotion in a store on a given day</em></td></tr>
<tr><td>B</td><td>Use the data given below. Create adjacency matrix. Use complete link algorithm to cluster given data set. Draw dendrogram.


</td></tr>
<tr><td>C</td><td>Using the following training data set. Create classification model using decision-treeand draw final Tree.

<table>
<tr><th>Tid</th><th>Income</th><th>Age</th><th>Own House</th></tr>
<tr><td>1.</td><td>Very High</td><td>Young</td><td>Yes</td></tr>
<tr><td>2.</td><td>High</td><td>Medium</td><td>Yes</td></tr>
<tr><td>3.</td><td>Low</td><td>Young</td><td>Rented</td></tr>
<tr><td>4.</td><td>High</td><td>Medium</td><td>Yes</td></tr>
<tr><td>5.</td><td>Very high</td><td>Medium</td><td>Yes</td></tr>
<tr><td>6.</td><td>Medium</td><td>Young</td><td>Yes</td></tr>
<tr><td>7.</td><td>High</td><td>Old</td><td>Yes</td></tr>
<tr><td>8.</td><td>Medium</td><td>Medium</td><td>Rented</td></tr>
<tr><td>9.</td><td>Low</td><td>Medium</td><td>Rented</td></tr>
<tr><td>10.</td><td>Low</td><td>Old</td><td>Rented</td></tr>
<tr><td>11.</td><td>High</td><td>Young</td><td>Yes</td></tr>
<tr><td>12.</td><td>medium</td><td>Old</td><td>Rented</td></tr>
</table></td></tr>
</table>