University of Mumbai

Program: Computer Engineering

Curriculum Scheme: Rev2019 Examination: TE Semester V

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

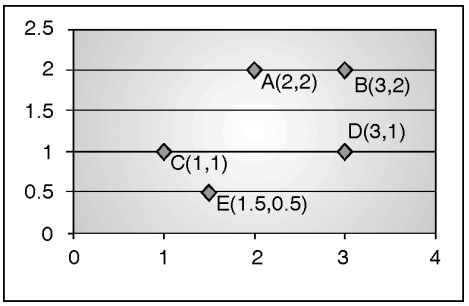
Time: 2 hour 30 minutes Max. Marks: 80

==========

Q1.	**Choose the correct option for following questions. All the Questions are compulsory and carry equal marks**
1.	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
2.	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
3.	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
4.	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
5.	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
6.	_____ 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
7.	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
8.	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
9.	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
10.	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

Q2	**Solve any Two Questions out of Three** **10 marks each**
A	*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.* *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*
B	Use the data given below. Create adjacency matrix. Use **complete** link algorithm to cluster given data set. Draw dendrogram.



C

llowing training data set. Create classification model using decision-tree and draw final Tree.

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

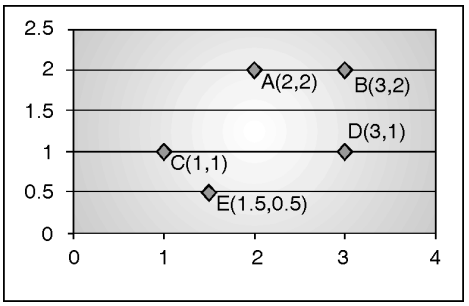
Q3	**Solve any Two Questions out of Three** **10 marks each**

A

Transaction database is given Below. Min Support = 2. Draw FP-Tree and find frequent patterns.

TID	*List of item_Ids*
T100	*I1, I2, I5*
T200	*I2, I4*
T300	*I2, I3*

T400	*I1, I2, I4*
T500	*I1, I3*
T600	*I2, I3*
T700	*I1, I3*
T800	*I1, I2, I3, I5*
T900	*I1, I2, I3*

B	Use the data given below. Create adjacency matrix. Use **Single** link algorithm to cluster given data set. Draw dendrogram. 
C	Suppose that the data for Analysis includes the attribute salary. We have the following values for salary(in thousands of dollars), shown in increasing order: 30, 36,47,50, 52,52,56,60,63,70,70,110. (i) What are the mean, median, mode and midrange of the data? (ii) Find theftrst quartile (Ql) and the third quartile (Q3) of the data. (iii)Show a boxplot of the data.

Q4	**Solve any Two Questions out of Three** **10 marks each**
A	Why is entity-relationship modeling technique not suitable for the data warehouse? How is dimensional modeling different?
B	Explain Page Rank Technique in detail with example
C	Demonstrate Multilevel and multidimensional association rule mining with examples of each.