University of Mumbai Program:
Computer Engineering Curriculum
Scheme: Rev2019
Examination: Third Year Semester: V
Course Code: CSC504 Course Name : Data Warehousing & Mining

Time: 2 hours 30 mins Max. Marks: 80

Q1. All questions compulsory 2 marks each (20 Marks)

Q1.	What is the access rights for a data Warehouse?
Option A:	Read Only
Option B:	Write only
Option C:	Read & Write
Option D:	None
Q2.	What is Transient data?
Option A:	Data in which changes to existing records cause the previous version of the records to be eliminated
Option B:	Data in which changes to existing records do not cause the previous version of the records to be eliminated
Option C:	Data that are never altered or deleted once they have been added
Option D:	Data that are never deleted once they have been added
Q3.	Which Operation treats incorrect or missing data?
Option A:	Pre-processing
Option B:	Interpretation
Option C:	Selection
Option D:	Transformation
Q4.	Summarization of the general characteristics or feature of a target class of data is known as
Option A:	Data Characterization
Option B:	Data Classification
Option C:	Data discrimination
Option D:	Data selection

Q5.	__________ is a technique which is used for data reduction in data mining process
Option A:	Attribute subset selection
Option B:	Correlation
Option C:	Cartesian Product
Option D:	Join
Q6.	For a Confusion Matrix, True Negative= 100, False Positive= 20, False Negative=10, True Positive=200 . Values of Sensitivity and Specifity are:
Option A:	95% and 83.3%
Option B:	100% and 70%
Option C:	70% and 100%
Option D:	86.2% and 74%
Q7.	Outliers effect which algorithm the most?
Option A:	K-means clustering algorithm
Option B:	K-medoids clustering algorithm
Option C:	K-medians clustering algorithm
Option D:	K-modes clustering algorithm
Q8.	What is the output given by Hierarchical Clustering ?
Option A:	final estimate of cluster centroids
Option B:	tree showing how close things are to each other
Option C:	assignment of each point to clusters
Option D:	outliers

Q9.	This method constructs a highly compact data structure to compress the original transaction database while discovering interesting patterns
Option A:	Apriori
Option B:	Classification
Option C:	Clustering
Option D:	Frequent Pattern Growth
Q10.	Clickstream is also known as _____________
Option A:	Web log
Option B:	Buffer Data
Option C:	Rank-sink
Option D:	Hub

Q2. (20 Marks Each)	Solve any Two Questions out of Three 10 marks each
A	Suppose that a data warehouse for Big University consists of the four dimensions student, course, semester, and instructor, and two measures count and avg grade. At the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg grade measure stores the actual course grade of the student. At higher conceptual levels, avg grade stores the average grade for the given combination. (a) Draw a snowflake schema diagram for the data warehouse. (b) Starting with the base cuboid [student,course,semester,instructor], what specific OLAP operations (e.g., roll-up from semester to year) should you perform in order to list the average grade of CS courses for each Big University student.

University of Mumbai Program:
Computer Engineering Curriculum
Scheme: Rev2019
Examination: Third Year Semester: V
Course Code: CSC504 Course Name : Data Warehousing & Mining
Time: 2 hours 30 mins Max. Marks: 80

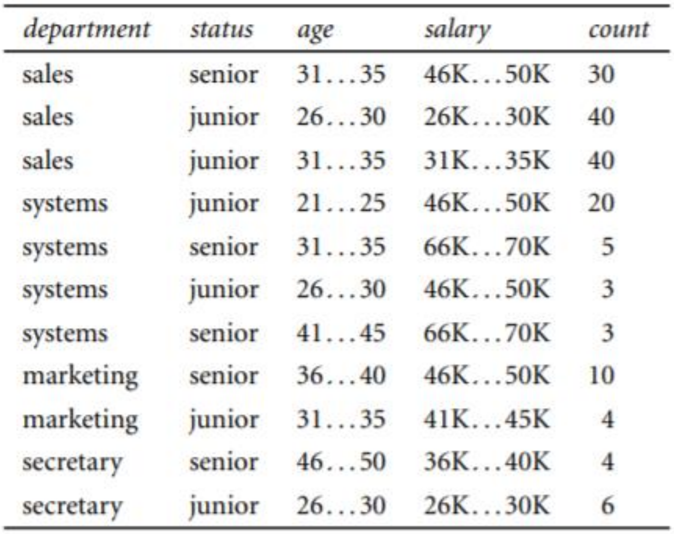
B	Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. (a) What is the mean of the data? What is the median? (b) What is the mode of the data? Comment on the data’s modality (i.e., bimodal, trimodal, etc.). (c) What is the midrange of the data? (d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data? (e) Give the five-number summary of the data. (f) Show a boxplot of the data.
C	In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

University of Mumbai Program:
Computer Engineering Curriculum
Scheme: Rev2019
Examination: Third Year Semester: V
Course Code: CSC504 Course Name : Data Warehousing & Mining
Time: 2 hours 30 mins Max. Marks: 80

Q3. (20 Marks Each)	Solve any Two Questions out of Three	10 marks each

A

The following table consists of training data from an employee database. The data have been generalized. For example, “31 ... 35” for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row



department	*status*	*age*	*salary*	*count*
sales	senior	31…35	46K…50K	30
sales	junior	26…30	26K…30K	40
sales	junior	31…35	31K…35K	40
systems	junior	21…25	46K…50K	20
systems	senior	31…35	66K…70K	5
systems	junior	26…30	46K…50K	3
systems	senior	41…45	66K…70K	3
marketing	senior	36…40	46K…50K	10
marketing	junior	31…35	41K…45K	4
secretary	senior	46…50	36K…40K	4
secretary	junior	26…30	26K…30K	6

Let status be the class label attribute.

(a) Use your algorithm to construct a decision tree from the given data.

B

Consider four objects with two attributes (X,Y).These four objects are to be grouped together into two clusters .Following are the objects with their attribute value. Apply K-means clustering algorithm on given dataset.

Objects	X	Y
A	1	1
B	2	1
C	4	3
D	5	4

C

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

TID	Items bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

(a) Find all frequent itemsets using Apriori
(b) List all the strong association rules (with support s and confidence c)

University of Mumbai Program:
Computer Engineering Curriculum
Scheme: Rev2019
Examination: Third Year Semester: V
Course Code: CSC504 Course Name : Data Warehousing & Mining
Time: 2 hours 30 mins Max. Marks: 80

Q4. (20 Marks Each) Solve any Two Questions out of Three 10 marks each

A

Car theft example: Attributes are color, type, origin and the subject, stolen can be either yes or no.

Car No.	Color	Type	Origin	Stolen
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Apply Naïve-Bayes algorithm on above dataset

B

Use the data given below. Create adjacency matrix. Use Single link algorithm to cluster given data set. Draw Dendrogram.

Object	Attribute(X)	Attribute(Y)
A	2	2
B	3	2
C	1	1
D	3	1
E	1.5	0.5

C

Explain Personalization with an example.