Qn. Set Code-1

Semester: 5th
Programme: B.Tech
Branch: CSE, IT

AUTUMN END SEMESTER EXAMINATION-2022

5th Semester B.Tech

DATA MINING AND DATA WAREHOUSING

IT-3031

(For 2021 (L.E), 2020 & Previous Admitted Batches)

Time: 3 Hours Full Marks: 50

Answer any SIX questions.
Question paper consists of four SECTIONS i.e. A, B, C and D.
Section A is compulsory.
Attempt minimum one question each from Sections B, C, D.
The figures in the margin indicate full marks.
Candidates are required to give their answers in their own words as far as practicable and all parts of a question should be answered at one place only.

SECTION-A

1. Answer the following questions. [1 × 10]

(a) What are the dimensions to measure quality of data.(at least 4 dimensions)

(b) Explain the use of concept hierarchy with a suitable example.

(c) How is data warehouse different from a data base?

(d) Which OLAP operations are responsible for computation of all of the data relationship for one or more dimension?

(e) How snowflake schema is different from star schema?

(f) In Apriori algorithm, if 1 item-sets are 100, then how many number of candidate 2 item-sets are possible out of it.

(g) Write any four limitations of hierarchical clustering.

(h) Define nominal, ordinal variables with suitable example.

(i) Our School of Computer Engineering has total 360 students of four different electives. These are ML: 36, BigData: 54, AI: 90, DMDW: 180. How many from each elective will include in a stratified random sampling of size 20?

(j) If a classifier correctly classify 80 tuples out of 120 tuples of test data what will be error rate of the classifier?

SECTION-B

2. (a) Explain briefly how the knowledge is extracted from the KDD process. List out at least two different data mining techniques to carry out each data mining task. [4]

(b) What is a data warehouse? Give the steps for the design and construction of Data Warehouses and explain with a three-tier architecture diagram. Suppose that a data warehouse consists of four dimensions customer, product, salesperson and sales time, and the three measure sales Amt(in rupees), GST(in rupees) and payment_type(in rupees). Draw the different classes of schemas that are popularly used for modelling data warehouses and explain them. [4]

3. (a) Consider the following table, which presents a simplified example of sold items (#Computers, #Mobile) observed at six different time points for the same organization. [4]

Time Point	#Computers	#Mobiles
T1	15	98
T2	12	87
T3	7	75
T4	16	95
T5	10	90
T6	10	85

Identify and apply the method will helps to check whether these two numeric attributes of selling are related to each other or not?

(b) Discuss the steps of any two algorithms [4]

i. Genetic algorithm

ii. Text Mining

iii. Web Mining

SECTION-C

4. (a) Apply a priori algorithm to the following data set. Assume the transactions are: [4]

Trans ID	**Items Purchased**
101	Apple, Orange, Litchi, Grapes
102	Apple, Mango
103	Mango, Grapes, Apple
104	Apple, Orange Litchi, Grapes
105	Pears, Litchi
106	Pears
107	Pears, Mango
108	Apple, Orange, Strawberry, Litchi, Grapes
109	Strawberry, Grapes
110	Apple, Orange, Grapes

Consider the minimum support count 3. Find all strong association rules with minimum confidence 80%.

(b) Calculate Root Mean Squared Error rate (RMSE) and Mean Absolute Error (MAE) rate for the test set. [4]

Y	8	6	7	5	3
Y’	9	5	7	2	6

5. (a) Draw the multi layer perceptron for the given data. Calculate the output at output layers and update the parameters of the neural network using back propagation algorithm with sigmoid activation function and learning rate is 0.8. [4]

X=[1, 1, 0], W_h=[{0.3, 0.4,0.5}, {0.7,0.2,0.4}], W_o=[0.5, 0.6], b_h=[0.4, 0.5], b_o=[0.2]

(b) Consider the given confusion matrix for a binary classification problem. [4]

Class Labels	P(+)	P(-)
A(+)	360	40
A(-)	80	20

Calculate

(i) True positive rate (ii) Precision

(iii) Specificity (iv) Accuracy

6. (a) Draw single link and complete link dendogram to represent the cluster for the given distance matrix. [4]

	P1	P2	P3	P4	P5
P1	0				
P2	9	0			
P3	3	7	0		
P4	6	5	9	0	
P5	12	9	2	8	0

(b) Consider the given data set of species. [4]

Name	**Give Birth**	**Can Fly**	**Live in Water**	**Have Legs**	**Class**
Human	Yes	No	No	Yes	Mammals
Python	No	No	No	No	Non-Mammals
Salman	No	No	Yes	No	Non-Mammals
Whale	Yes	No	Yes	No	Mammals
Frog	No	No	Sometimes	Yes	Non-Mammals
Komodo	No	No	No	Yes	Non-Mammals
Bat	Yes	Yes	No	Yes	Mammals
Pigeon	No	Yes	No	Yes	Non-Mammals
Cat	Yes	No	No	Yes	Mammals
Leopard shark	Yes	No	Yes	No	Non-Mammals
Turtle	No	No	Sometimes	Yes	Non-Mammals
Penguin	No	No	Sometimes	Yes	Non-Mammals
Porcupine	Yes	No	No	Yes	Mammals
Eel	No	No	Yes	No	Non-Mammals
Salamander	No	No	Sometimes	Yes	Non-Mammals
Gila Monster	No	No	No	Yes	Non-Mammals
Platypus	No	No	No	Yes	Mammals
Owl	No	Yes	No	Yes	Non-Mammals
Dolphin	Yes	No	Yes	No	Mammals
Eagle	No	Yes	No	Yes	Non-Mammals

Predict the class label of it using Bayesian classification technique.

Give Birth	**Can Fly**	**Live in Water**	**Have Legs**	**Class**
Yes	No	Yes	No	?

SECTION-D

7. (a) Given a survey data containing continuous performances in three different activities about the students who are like to opt end-semester examination in two operational modes, either in On-line or Off-line. Using manhattan distance, show how the KNN classifier (let K=3) with majority voting would classify the following student X with {Activity1=5, Activity2 =7, Activity3 =8} [4]

ID	**Activity1**	**Activity2**	**Activity3**	**Opt-to**
1	4	4	5	On-line
2	6	8	8	On-line
3	9	5	5	On-line
4	7	6	8	Off-line
5	7	8	7	Off-line

(b) Consider the same training data without the class labels (Opt-to). You are asked to labeling or grouping the data points into 2 groups (Off-line and On-line). Choose K-means clustering for generating the final clusters and updated centroids after two iterations. Consider initial centroids are ID-2 and ID-4. [4]

8. (a) Predict the value of Y if X=15 using linear regression. [4]

X	2	4	6	8
Y	3	7	5	10

(b) Consider the dataset and find out which attribute is suitable for best split point in C4.5 decision tree induction. [4]

#	Attribute			Shape
	Color	**Outline**	**Dot**	
1	Green	Dashed	No	Triangle
2	Green	Dashed	Yes	Triangle
3	Yellow	Dashed	No	Square
4	Red	Dashed	No	Square
5	Red	Solid	No	Square
6	Red	Solid	Yes	Triangle
7	Green	Solid	No	Square
8	Green	Dashed	No	Triangle
9	Yellow	Solid	Yes	Square
10	Red	Solid	No	Square
11	Green	Solid	Yes	Square
12	Yellow	Dashed	Yes	Square
13	Yellow	Solid	No	Square
14	Red	Dashed	Yes	Triangle

*****