Qn. Set Code-1

Semester: 6th
Programme: B.Tech
Branch: CSSE

SPRING END SEMESTER EXAMINATION-2024

6th Semester B.Tech

DATA MINING AND DATA WAREHOUSING

IT 3031

(For 2021 & Previous Admitted Batches)

Time: 2 Hours 30 Minutes | 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.
All parts of a question should be answered at one place only.

SECTION-A

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

   (a) Differentiate between interval attribute and ratio attribute.

   (b) *If* p and q, are the binary attributes find SMC and Jaccard coefficient.

   $p = 1\ 1\ 1\ 0\ 0\ 1\ 0\ 1\ 0\ 1$

   $q = 0\ 1\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 1$

   (c) Differentiate between OLAP and OLTP.

   (d) Brief the term frequent itemset and closed itemset.

   (e) Calculate the mean and Standard deviation of the following data

   [18,20,22,16,14,12,8,10,11,6].

   (f) What are the assumptions taken during linear regression?

   (g) What are the requirements of clustering in data Mining?

(h) Differentiate between snowflake schema and fact constellation schema.

(i) Given two objects represented by the tuples (24, 12, 40, 16) and (22, 9, 37, 8). Compute the Euclidean distance between the two objects.

(j) What are the steps of text mining Techniques?

SECTION-B

2. (a) State the difference between Symmetric vs. Skewed data and determine the five-number summary, tukey fence, and outlier (if any) for the data set given below. [4]

{5, 18, 7, 14, 18, 12, 4, 16, 10, 6, 6, 12}

Draw the box plot to describe the distribution of data in the data set.

(b) Suppose a company wants to decide on a promotion based on the years of work experience of its employees. So, it needs to analyze a database that looks like this: [4]

Name	Years
A	5
B	6
C	7
D	4
E	3
F	8
G	12
H	10
I	11
J	13

Use the following three methods to normalize the given data

a. min-max normalization by setting min=0 and max=1

b. Z-Score normalization

c. Decimal Scaling

3. (a) Is gender independent of education level? A random sample of 395 people were surveyed and each person was asked to report the highest education level they obtained. The data that resulted from the survey is summarized in the following table: [4]

	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

Perform a chi square to find are gender and education level dependent at 5% level of significance? In other words, given the data collected above, is there a relationship between the gender of an individual and the level of education that they have obtained? (For the above case the critical value of chi square with 3 degree of freedom is 7.815)

(b) Explain different types of schemas used in the multi-dimensional model in Data Warehousing. [4]

SECTION-C

4. (a) Apply apriori 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% also mentioned the method to improve the efficiency of apriori.

(b) The sales of a company (in crores) for each year are shown in the table below. [4]

Year (x)	2013	2016	2018	2020	2022
Sales (y)	10	25	35	50	70

Find the least square regression line y=ax+b.

Use this to estimate the sales of the company in the year 2024.

5. (a) Using KNN find the default status of sumit. [4]

Customer	Age	Loan	Default
Suman	25	4000	N
Arya	30	4000	N
Sarthak	35	8000	N
Rohit	23	2000	N
Hardik	26	2500	Y
Vineet	31	1800	Y
Suryansh	22	9000	Y
Sagun	40	4500	N
Suneha	42	5600	N
Yash	45	7000	Y
Sumit	38	1400	?

The value of k=5.Also sate the disadvantages of KNN.

(b) What is the difference between probabilistic and non-probabilistic sampling? [4]

Describe different types of probability sampling.

6. (a) What is the major assumption that the Naive Bayes classifier follows? [4]

T_Id	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	120K	No
2	No	Married	110K	No
3	No	Single	70K	No
4	Yes	Married	80K	No
5	Yes	Divorced	90K	No
6	No	Married	200K	YES
7	No	Divorced	150k	No
8	Yes	Single	130K	YES
9	No	Single	75K	YES

Given the above dataset for tax evaders, use a Naive bayes classifier to predict whether the following customer will evade tax.

Y = (Refund = No, Marital Status=Married, Income = 100K)

(b) Consider 12 points {X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12} with the following coordinates into two clusters: [4]

X1(1, 1), X2(1, 3), X3(2, 2), X4(3, 1), X5(4, 4), X6(8, 1), X7(8, 3), X8(9, 2), X9(10, 1), X10(10, 3), X11(3,6) and X12(4,8)

Suppose initial cluster center for each cluster are X1, X2 respectively. Use k-means algorithm to show the clusters.

SECTION-D

7. (a) Calculate Gini Index for past trend, open interest, trading volume and return in the following manner with the example data and build the decision tree. [4]

Past Trend	Open Interest	Trading Volume	Retur n
Positive	Low	High	Up
Negative	High	Low	Down
Positive	Low	High	Up
Positive	High	High	Up
Negative	Low	High	Down
Negative	High	High	Down
Negative	Low	High	Down
Positive	Low	Low	Down
Positive	High	High	Up
Positive	Low	High	Up

(b) Consider the following prediction summary of a model. [4]

N=100	Actual - Yes	Actual - No
Predicted - Yes	60	10
Predicted - No	5	25

Define and calculate the following performance metrics from the above summary

i. precision

ii. recall

iii. f-measure

iv. misclassification rate

8. (a) For the given Data Set find the cluster using a complete link technique. Use Euclidean distance and draw the dendrogram. [4]

Sample No	X	Y
1	1	4
2	3	5
3	6	8
4	5	7
5	4	6
6	6	2

(b) What is meant by the term "missing or erroneous values" in a dataset? Can outliers be treated as erroneous values? [4]

Describe various methods for addressing missing/erroneous values, and explain how outliers are handled.

*****