Qn. Set Code-1

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

SPRING END SEMESTER EXAMINATION-2023

6th 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) State the difference between nominal attribute and ordinal attribute.

(b) Find the dissimilarity between two vectors **'x'** and **'y'**, using Cosine Similarity.
The 'x' vector has values, **x = {3, 2, 0, 5}**
The 'y' vector has values, **y = {1, 0, 0, 0}**

(c) Find the Interquartile range (IQR) of the following data set D:
D= {45,47,52,52,53, 55,56,58,62, 80}

(d) Differentiate between data marts and data warehouse.

(e) State the different techniques to improve the efficiency of Apriori algorithm.

(f) Explain the role of lift value in an association rule.

(g) List out the critical assumptions of linear regression.

(h) Briefly explain the prior probability and posterior probability.

(i) Define agglomerative and divisive hierarchical clustering.

(j) List out the challenges with multimedia database.

SECTION-B

2. (a) Consider a model that can predict either a person has the disease or not from a sample data set having total number of samples N=100. Calculate (i) Misclassification rate (ii) Precision (iii) Recall (iv) Accuracy for the observed confusion matrix of the model. [4]

N=100	Actual: No	Actual: Yes
Predicted: No	65	3
Predicted: Yes	8	24

(b) With suitable example describe briefly about different type of missing value in data. [4]

3. (a) A group of twelve children participated in a psychological study designed to assess the relationship between age (x years) and average total sleep time (ATST) in y minutes and listed in given table. [4]

Child	Age (x years)	ATST (y minutes)
A	4.4	586
B	6.7	565
C	10.5	515
D	9.6	532
E	12.4	478

Calculate the value of the correlation coefficient between x and y. Assess the statistical significance of your value and interpret your results.

(b) Briefly describe about the different type of non-probability sampling. [4]

SECTION-C

4. (a) Consider a Big Bazar scenario where the product set is P = {Rice, Pulse, Oil, Milk, Apple}. The database comprises six transactions where 1 represents the presence of the product and 0 represents the absence of the product. [4]

Transaction ID	Rice	Pulse	Oil Milk	Apple
t1	1	1	1	0
t2	0	1	1	1
t3	0	0	0	1
t4	1	1	0	1
t5	1	1	1	0
t6	1	1	1	1

Using apriori algorithm find all frequent item sets and association rules. if the minimum support is more than 50% and the minimum confidence is 80%.

(b) Explain Frequent Itemset. Closed Itemset and Maximal Itemset and establish the relation between them. [4]

5. (a) Consider a survey on people opinion to know whether a special paper tissue is good or not based on the two attributes (acid durability and strength). Given five training samples in the table below: [4]

X_1=Acid durability	X_2= Strength	Y=Class
8	9	Bad
6	5	Bad
3	4	Good
1	5	Good
8	8	Bad

Now the factory produces a new paper tissue that pass laboratory test with X_1= 3 and X_2= 7. Find out the class of the new tissue using K-NN where K=3.

(b) Draw the multi-layer perceptron network for the given data and write generalized steps of back propagation algorithm. [4]

$X=[X_1, X_2, X_3]$, $W_h=[\{a_1, a_2, a_3\}, \{b_1, b_2, b_3\}]$, $Wo=[k_1, k_2]$, $b_h=[c_1, c_2]$, $b_o=[d]$

6. (a) Discuss the steps of Naïve Bayes classification. [4]

(b) Construct a decision tree using Gini Impurity for the given training data in the table. [4]

Buy Computer data			
Income	**Student**	**Credit rating**	**Buys computer**
High	No	Fair	No
High	No	Excellent	No
High	No	Fair	Yes
Medium	No	Fair	Yes
Low	Yes	Fair	Yes
Medium	No	Excellent	No
Low	Yes	Excellent	No

SECTION-D

7. (a) Cluster the following six points (with (x, y)representing locations) into two clusters: [4]

A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4)

The Initial cluster centers are: A1(2, 10), A4(5, 8) and use Euclidean distance method.

(b) Draw the dendogram using complete link agglomerative clustering to group the data described by the following distance matrix. [4]

	A	B	C	D	E	E
A	0	1	4	5	7	8
B		0	2	6	9	10
C			0	3	11	12
D				0	8	7
E					0	6
F						0

8. (a) Discuss the comparison between data mining and Web mining with a suitable example. [4]

(b) Differentiate between spatial and temporal data mining with a suitable example. [4]

*****