BIRLA INSTITUTE OF TECHNOLOGY, MESRA, RANCHI
(END SEMESTER EXAMINATION)

CLASS: MTech / Pre-PhD **SEMESTER : II/NA**
BRANCH: CSE **SESSION : SP/2023**

SUBJECT: IT516 DATA MINING AND DATA ANALYSIS

TIME: 3 Hours **FULL MARKS: 50**

INSTRUCTIONS:
1. The question paper contains 5 questions each of 10 marks and total 50 marks.
2. Attempt all questions.
3. The missing data, if any, may be assumed suitably.
***4*. Before attempting the question paper, be sure that you have got the correct question paper.**
5. Tables/Data hand book/Graph paper etc. to be supplied to the candidates in the examination hall.

			CO
Q.1(a)	A poker-dealing machine is supposed to deal cards at random, as if from an infinite deck. In a test, you counted 1600 cards, and observed the following: **Spades 404** **Hearts 420** **Diamonds 400** **Clubs 376** Could it be that the suits are equally likely? Or are these discrepancies too much to be random? The critical value (using $\alpha = 0.05$) is 7.815.	[5]	**CO1, B3**
Q.1(b)	Consider the two-dimensional patterns (2, 1), (3, 5), (4, 3), (5, 6), (6, 7), and (7, 8). Compute the principal component using PCA Algorithm.	[5]	**CO1, B3**
Q.2(a)	Design the architecture of a data mining system and explain its major components	[5]	**CO2,B3**
Q.2(b)	Suppose that a data warehouse for **ABC University** consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg grade. When 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. (i) Draw a snowflake schema diagram for the data warehouse. (ii) Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations on multidimensional data? (iii) If each dimension has five levels (including all), such as "student < major < status < university < all", how many cuboids will this cube contain (including the base and apex cuboids)?	[5]	**CO2,B3**
Q.3(a)	The following contingency table summarizes supermarket transaction data, where hot dogs refer to the transactions containing hot dogs, hotdogs' refers to the transactions that do not contain hot dogs, hamburgers refers to the transactions containing hamburgers, and hamburgers' refers to the transactions that do not contain hamburgers. **hotdog hotdog'** **hamburgers 2,000 500** **hamburgers' 1,000 1,500** (a) Suppose that the association rule "hot dogs ⇒ hamburgers" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong? (b) Based on the given data, is the purchase of hot dogs independent of the purchase of hamburgers? If not, what kind of correlation relationship exists between the two?	[5]	**CO3,B3**
Q.3(b)	Comparison of four Pattern Evaluation Measures (Lift, Chi-square, all_confidence, and cosine) using the above contingency table in Q3(a).	[5]	**CO3,B4**

PTO

Q.4(a) Consider the following set of training examples: [5] **CO4,B3**

Instance	Classification	a1	a2	a3
1.	+	T	T	1.0
2.	+	T	T	6.0
3.	-	T	F	5.0
4.	+	F	F	4.0
5.	-	F	T	7.0
6.	-	F	T	3.0
7.	-	F	F	8.0
8.	+	T	F	7.0
9.	-	F	T	5.0

(i) What is the entropy of this collection of training examples with respect to the positive class?
(ii) What are the information gains of a1 and a2 relative to these training examples?

Q.4(b) For a3(from the above table), which is a continuous attribute, compute the information gained for every possible split. [5] **CO4,B6**

Q.5(a) Briefly outline how to compute the dissimilarity between objects described by the following types of variables: [5] **CO5,B3**
(i) Numerical (interval-scaled) variables
(ii) Asymmetric binary variables
(iii) Categorical variables
(iv) Ratio-scaled variables
(v) Nonmetric vector objects

Q.5(b) Explain DBSCAN Clustering algorithm with eps=0.6 and MinPoints=4 of the given data set consisting of 14 points. [5] **CO5,B6**

X	Y	Distance from (1,2)
1	2	0
3	4	2.8
2.5	4	2.5
1.5	2.5	0.7
3	5	3.6
2.8	4.5	3.08
2.5	4.5	2.9
1.2	2.5	0.53
1	3	1
1	5	3
1	2.5	0.5
5	6	5.6
4	3	3.1

:::::28/04/2023 E:::::