Seat No.: _____ Enrolment No._____

GUJARAT TECHNOLOGICAL UNIVERSITY

BE - SEMESTER–VI (NEW) EXAMINATION – SUMMER 2023

Subject Code:3160714 Date:12-07-2023

Subject Name:Data Mining

Time:10:30 AM TO 01:00 PM Total Marks:70

Instructions:

1. **Attempt all questions.**
2. **Make suitable assumptions wherever necessary.**
3. **Figures to the right indicate full marks.**
4. **Simple and non-programmable scientific calculators are allowed.**

			MARKS
Q.1	**(a)**	What is market basket analysis? Precisely explain the meaning of the following association rule: computer → antivirus_software [support = 60%, confidence = 60%]	**03**
	(b)	In real-world data, tuples with missing values for some attributes are a common occurrence. List and describe various methods for handling this problem.	**04**
	(c)	With the help of a suitable diagram, describe the steps involved in data mining when viewed as a process of knowledge discovery.	**07**
Q.2	**(a)**	Give a short example to show that items in a strong association rule are not always interesting.	**03**
	(b)	Briefly describe how partitioning technique may improve the efficiency of Apriori algorithm.	**04**
	(c)	Discuss how frequent itemsets can be generated using FP-Growth algorithm with the help of the following transactions. Let minimum support threshold is 2.	**07**

Transaction ID	Item IDs
T1	I1, I2, I5
T2	I2, I4
T3	I2, I3
T4	I1, I2, I4
T5	I1, I3
T6	I2, I3
T7	I1, I3
T8	I1, I2, I3, I5
T9	I1, I2, I3

OR

	(c)	A database has the following six transactions.	**07**

Transaction ID	Items
T1	HotDogs, Buns, Ketchup
T2	Chips, Coke
T3	Coke, Chips, HotDogs
T4	Ketchup, Chips
T5	Buns, HotDogs
T6	HotDogs, Chips, Coke

Find all frequent itemsets and also generate the strong association rules using Apriori algorithm. Let minimum support threshold is 33.34% and minimum confidence threshold is 60%.

Q.3 **(a)** Describe any three primitives for specifying a data mining task. **03**

(b) The following table shows the midterm and final exam grades obtained by students in a database course. **04**

x (Midterm exam)	y (Final exam)
72	84
50	63
81	77
74	78
94	90
86	75
59	49
83	79
65	77
33	52
88	74
81	90

Use the method of least squares to find an equation for the prediction of a student's final exam grade based on the student's midterm grade in the course. Predict the final exam grade of a student who received 86 grade in the midterm exam.

(c) What is noise? Describe the possible reasons for noisy data. Explain the different techniques to remove the noise from data. **07**

OR

Q.3 **(a)** Discuss outlier analysis as a data mining functionality with the help of an example. **03**

(b) Explain how classification rules are extracted from a decision tree with the help of an example. **04**

(c) Explain in detail - min-max normalization method. Use this method to normalize the following group of data by setting min = 0 and max = 1. 200, 400, 600, 1000 **07**

Q.4 **(a)** Differentiate classification and clustering. **03**

(b) Discuss data matrix and dissimilarity matrix with respect to clustering. **04**

(c) Apply ID3 classification algorithm on the following data and construct a decision tree. Show all the stepwise calculations clearly. **07**

age	income	student	credit_rating	Class: buys_computer
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes

senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

OR

Q.4 **(a)** Discuss cross-validation method for evaluating the accuracy of a classifier. **03**

(b) How k-means clustering method differs from k-medoids clustering method? Discuss major drawbacks of k-means clustering method. **04**

(c) Predict class label of the tuple X = (**age** = youth, **income** = medium, **student** = yes, **credit_rating** = fair) with the help of Naive Bayesian classification method and the following data. Show all the stepwise calculations clearly. **07**

age	**income**	**student**	**credit_rating**	**Class: buys_computer**
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

Q.5 **(a)** Discuss web structure mining. **03**

(b) Discuss multimedia mining. **04**

(c) Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters: **07**

A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9)

The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. With the help of k-means algorithm calculate,

(i) The three cluster centers after the first round execution

(ii) The final three clusters

OR

Q.5 **(a)** Discuss agglomerative hierarchical clustering method in brief. **03**

(b) Explain the any four typical requirements of clustering in data mining. **04**

(c) What is web mining? Explain web usage mining in detail. **07**

****