Enrolment No./Seat No_____

GUJARAT TECHNOLOGICAL UNIVERSITY

BE- SEMESTER– VI EXAMINATION – SUMMER 2026

Subject Code:3160714 **Date:29-05-2026**

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.**

Q.1	**(a)**	Define Schema. Explain the following schemas with suitable example. 1) Star 2) Snowflakes 3) Constellations	**03**
	(b)	A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data – Justify.	**04**
	(c)	Describe the steps involved in data mining when viewed as a process of knowledge discovery.	**07**
Q.2	**(a)**	What is Concept Hierarchy? List and explain types of Concept Hierarchy.	**03**
	(b)	Differentiate OLAP and OLTP.	**04**
	(c)	Explain the major issues in data mining	**07**
		OR	
	(c)	Explain the need for data smoothing during pre-processing and discuss data smoothing by Binning.	**07**
Q.3	**(a)**	Briefly explain Linear and Non-linear regression.	**03**
	(b)	Explain measures for finding rule interestingness (support,confidence) with example.	**04**
	(c)	Compare association and classification. Briefly explain associative classification with suitable example.	**07**
		OR	
Q.3	**(a)**	Discuss following terms. 1) Supervised learning 2) Correlation analysis	**03**
	(b)	Write a note on incremental Association Rule Mining.	**04**
	(c)	Explain Hash-based technique to improve efficiency of Apriori algorithm.	**07**
Q.4	**(a)**	Define “clustering”? Mention any two applications of clustering.	**03**
	(b)	Draw the topology of a multilayer, feed-forward Neural Network.	**04**
	(c)	Write and discuss the algorithm which is used to generate frequent itemsets using an iterative level-wise approach based on candidate generation.	**07**
		OR	
Q.4	**(a)**	Explain web mining using example.	**03**
	(b)	What is classification? Explain classification as a two step process with diagram.	**04**
	(c)	Discuss the main features of Hadoop Distributed File System.	**07**
Q.5	**(a)**	Explain Hadoop storage – HDFS.	**03**
	(b)	How K-Mean clustering method differs from K-Medoid clustering method?	**04**
	(c)	Explain mapreduce. Explain any example using mapreduce.	**07**

OR

Q.5	**(a)**	Explain data mining application for fraud detection.	**03**
	(b)	Dicuss the Applications of Distributed and parallel Data Mining.	**04**
	(c)	What is web log? Explain web structure mining and web usage mining in detail.	**07**

****

Enrolment No./Seat No_____

GUJARAT TECHNOLOGICAL UNIVERSITY

BE - SEMESTER–VI EXAMINATION – SUMMER 2025

Subject Code: 3160714 Date: 28-05-2025

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 the difference between discrimination and classification?	**03**
	(b)	Describe the steps involved in data mining when viewed as a process of knowledge discovery.	**04**
	(c)	Draw and explain the data mining architecture.	**07**
Q.2	**(a)**	Discuss issues to consider during data integration.	**03**
	(b)	Describe various methods for handling the missing values in the dataset.	**04**
	(c)	Use the methods below to normalize the following group of data: 200, 300, 400, 600, 1000 (a) min-max normalization by setting min = 0 and max = 1 (b) Calculate z-score normalization (c) z-score normalization using the mean absolute deviation instead of standard deviation (d) normalization by decimal scaling	**07**
		OR	
	(c)	For the following data (in increasing order) for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. (a) Use smoothing by bin means to smooth these data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data. (b) How might you determine outliers in the data? (c) What other methods are there for data smoothing?	**07**
Q.3	**(a)**	What is meant by association rule?	**03**
	(b)	Write a short note on: Web content mining.	**04**
	(c)	Explain about the Apriori algorithm for finding frequent item sets with an example.	**07**
		OR	
Q.3	**(a)**	What is meant by Maximal Frequent Item Set?	**03**
	(b)	Write a short note on: Web usage mining.	**04**
	(c)	Which patterns are interesting? Explain based on Pattern Evaluation Methods.	**07**

Q.4	**(a)**	What is the difference between "supervised" and unsupervised" learning?	**03**
	(b)	What are the attribute selection measures? Explain any one in details.	**04**
	(c)	Discuss about Decision tree induction algorithm with an example.	**07**

OR

Q.4	**(a)**	What do you meant by Bayesian Classification?	**03**
	(b)	Describe the issues regarding classification and prediction.	**04**
	(c)	Discuss about Neural Network-Based Algorithms.	**07**

Q.5	**(a)**	What are the requirements of clustering?	**03**
	(b)	Discuss about Outlier Detection.	**04**
	(c)	Discuss about k-nearest neighbor classification algorithm with an example?	**07**

OR

Q.5	**(a)**	State the categories of clustering methods?	**03**
	(b)	Explain the concept hierarchy.	**04**
	(c)	Explain in detail about partitioned Clustering method.	**07**

****

Enrolment No./Seat No_____

GUJARAT TECHNOLOGICAL UNIVERSITY

BE - SEMESTER–VI (NEW) EXAMINATION – SUMMER 2024

Subject Code: 3160714 Date:22-05-2024

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)**	Define data mining. Describe three challenges to data mining regarding data mining methodology and user interaction issues.	**03**
	(b)	Explain the steps in knowledge discovery.	**04**
	(c)	Explain the various data mining issues.	**07**
Q.2	**(a)**	What are the smoothing techniques available to remove noise?	**03**
	(b)	Discuss normalization in detail.	**04**
	(c)	In real-world data, tuples with *missing values* for some attributes are a common occurrence. Describe various methods for handling this problem.	**07**
		OR	
	(c)	Discuss data discretization and concept hierarchy generation.	**07**
Q.3	**(a)**	How are association rules mined from large databases?	**03**
	(b)	Give the difference between Boolean association rule and quantitative association rule.	**04**
	(c)	What are the limitations of the apriori approach for mining? Briefly describe the techniques to improve the efficiency of apriori algorithm.	**07**
		OR	
Q.3	**(a)**	Describe two interesting measures for association rules.	**03**
	(b)	How Meta rules are useful in constraint based association mining.	**04**
	(c)	Write an algorithm for finding frequent item-sets using candidate generation.	**07**
Q.4	**(a)**	What are the difference between supervised learning and unsupervised learning?	**03**
	(b)	Write down short note on backpropagation.	**04**
	(c)	What is information gain? Explain the steps required to generate a decision tree from a training data set.	**07**
		OR	
Q.4	**(a)**	Differentiate between linear regression and nonlinear regression.	**03**
	(b)	Explain various methods of evaluating accuracy of classifier.	**04**
	(c)	Write a short on: web content mining.	**07**
Q.5	**(a)**	Explain temporal mining.	**03**
	(b)	Differentiate between partitioning and hierarchical methods for clustering.	**04**
	(c)	Explain following clustering algorithm in details: 1) CLARA 2) BIRCH	**07**
		OR	
Q.5	**(a)**	List out the applications of distributed and parallel data mining.	**03**
	(b)	Illustrate strength and weakness of k-mean in comparison with k-medoid algorithm.	**04**
	(c)	Explain the typical requirements of clustering in data mining.	**07**

****

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**

****

Seat No.: _____ Enrolment No._____

GUJARAT TECHNOLOGICAL UNIVERSITY

BE - SEMESTER–VI (NEW) EXAMINATION – SUMMER 2022

Subject Code:3160714 Date:08/06/2022

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 are the types of data?	**03**
	(b)	Compare descriptive and predictive data mining	**04**
	(c)	Draw and explain the data mining architecture.	**07**
Q.2	**(a)**	What is dimensionality reduction?	**03**
	(b)	What are the types of concept hierarchies?	**04**
	(c)	What is Data Cleaning? Describe various methods of Data Cleaning.	**07**
		OR	
	(c)	Discuss issues to be considered during data integration.	**07**
Q.3	**(a)**	What is meant by association rule?	**03**
	(b)	How is association rules mined from large databases?	**04**
	(c)	Explain the various criteria for the classification of frequent pattern mining.	**07**
		OR	
Q.3	**(a)**	List two interesting measures for association rules.	**03**
	(b)	What is meant by multidimensional association rules?	**04**
	(c)	Write short notes on Maximal Frequent Item Set &Closed Frequent Item Set.	**07**
Q.4	**(a)**	What is an outlier?	**03**
	(b)	What is Bayesian theorem?	**04**
	(c)	Demonstrate how Bayesian classification helps in predicting class membership probabilities.	**07**
		OR	
Q.4	**(a)**	Differentiate classification and prediction.	**03**
	(b)	What is the difference between “supervised” and unsupervised” learning scheme.	**04**
	(c)	Explain the issues regarding the classification and prediction.	**07**
Q.5	**(a)**	What is temporal mining?	**03**
	(b)	Explain web usage mining.	**04**
	(c)	Discuss the K-means clustering algorithm using examples.	**07**
		OR	
Q.5	**(a)**	What is multimedia mining?	**03**
	(b)	Explain web content mining.	**04**
	(c)	What do you meant by Clustering? Explain the requirements used in Clustering?	**07**

****