Seat No.: _____ Enrolment No._____

GUJARAT TECHNOLOGICAL UNIVERSITY

ME – SEMESTER – II (New)– EXAMINATION – WINTER-2019

Subject Code: 3722319 **Date: 21-11-2019**

Subject Name: Data Preparation and Analysis

Time: 02:30 PM TO 05:00 PM **Total Marks: 70**

Instructions:

1. **Attempt all questions.**
2. **Make suitable assumptions wherever necessary.**
3. **Figures to the right indicate full marks.**

Q.1 **(a)** Why data preparation is required before applying data mining tool? **[7]**

(b) 1. Define and give example of following type of measurements: **[6]**
(i) Nominal scale (ii) Ordinal scale (iii) Interval scale

2. Define the term: discretization **[1]**

Q.2 **(a)** A traffic inspector has counted the number of automobiles passing a certain point in 100 successive 20-minute time periods. The observations are listed below. **[7]**

23 20 16 18 30 22 26 15 5 18
14 17 11 37 21 6 10 20 22 25
19 19 19 20 12 23 24 17 18 16
27 16 28 26 15 29 19 35 20 17
12 30 21 22 20 15 18 16 23 24
15 24 28 19 24 22 17 19 8 18
17 18 23 21 25 19 20 22 21 21
16 20 19 11 23 17 23 13 17 26
26 14 15 16 27 18 21 24 33 20
21 27 18 22 17 20 14 21 22 19

Construct a histogram and a frequency polygon for this traffic data.

(b) Define the term: outlier. **[1]**
Explain any two methods of outlier detection with suitable example. **[6]**

OR

(b) Explain approaches to dealing with missing values in heterogeneous data. **[7]**

Q.3 **(a)** Explain all steps involved in knowledge discovery process. **[7]**

(b) What are advantages of descriptive and inferential statistics? By taking example, explain it. **[7]**

OR

Q.3 **(a)** What is CRISP-DM? Explain its life cycle. **[7]**

(b) Explain fuzzy k-means clustering. **[7]**

Q.4 **(a)** Define the term: normalization **[1]**
Explain any two methods of normalization with suitable example. **[6]**

(b) What is exploratory data analysis? **[1]**
Explain any three graphical techniques used in exploratory data analysis. **[6]**

OR

Q.4 **(a)** Give brief about any seven temporal visualization styles for time series data. **[7]**

(b) Define the term: correlation. **[1]**
Differentiate between pearson correlation coefficient and spearman's correlation with suitable example. **[6]**

Q.5	**(a)**	1. With suitable example, differentiate between (i) equal-depth binning and (ii) equal-width binning	**[6]**
		2. Define the term: entropy in data	**[1]**
	(b)	Define the term: hierarchical data	**[1]**
		Differentiate between following three information visualization methods for showing hierarchical data. (i) cone tree diagram (ii) botanical tree diagram and (iii) treemap diagram	**[6]**

OR

Q.5	**(a)**	How to handle missing data? Explain with suitable examples.	**[7]**
	(b)	What are scalability issues in data integration? Also propose solution for each.	**[7]**

****