AUTUMN MID SEMESTER EXAMINATION-2022

School of Computer Engineering
Kalinga Institute of Industrial Technology, Deemed to be University

DATA MINING AND DATA WAREHOUSING

[IT 3031]

Time: 1 1/2 Hours **Full Mark: 20**

Answer any four Questions including Q.No.1 which is Compulsory.
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.

1. Answer all the questions. **[1 x 5]**

   a) List out the at least 4 types of databases and also state the kind of data mining techniques can be applied on those databases.
   b) Compute the Jaccard similarity between the following two binary vectors
   {x = 0101010001, y = 0100011000}
   c) Explain the difference between **slice** and **dice** OLAP operation with suitable example.
   d) Define closed frequent itemset and maximal frequent itemset with suitable example.
   e) Which visualization technique is best suited for understanding the distribution two variables are same or not?

2. For the following data set find the five-number summary, IQR, Tukey fence, and outlier (if any). Draw the box plot to describe the distribution of data in the data set.
   18, 34, 76, 29, 15, 41, 46, 25, 54, 38, 20, 32, 43, 22 **[5 Marks]**

3. Suppose a group of 12 sales price records has been sorted as follows: 5, 10, 11, 13, 15, 35, 50, 55, 72,92, 204, 215.Partition them into three bins by each of the following methods then smooth out the bin values by median and by boundary for both the partitioning **[5 Marks]**
   (a) Equal-frequency partitioning (b) Equal-width partitioning

4. Suppose the price and weight of 10 products in a shop is given in the below table. **[5 Marks]**

Price	10.4	8.2	20.5	57.3	195	60	33	54	130	220
Weight	0.7	1.9	2.1	2.6	3.1	1.1	2.7	1.8	4.7	9.4

   a. Normalize the weight variables based on the min-max normalization (min = 10, max = 20)
   b. Normalize the price variables with z-score normalization
   c. Calculate the Pearson correlation coefficient. Are these two variables positively or negatively correlated?

5. A transaction table shown below, and assuming a minimum level of support mini_sup = 60% and a minimum level of confidence min_conf = 80%. **[5 Marks]**

TID	Date	Item Bought
T100	11/09/2022	{ B,A,D}
T200	12/09/2022	{ C,A,B,E}
T300	13/09/2022	{ D,A,C,E,B}
T400	15/09/2022	{ K,A,D,B}

   (a) Find all frequent itemsets using the Apriori algorithm.

   (b) List all of the strong association rules, along with their support and confidence value.

***** Best of Luck *****