Department of Computer Science and Engineering, S V N I T, Surat
Mid-Semester Examinations, Sept. - Oct. 2022

B. Tech. IV (CSE) – 7th Semester
Course: Data Warehousing and Mining (CS441)

Date: 29th September 2022 Time: 14.00 hrs to 15.30 hrs Max Marks: 30

Instructions: 1.Please start the answer to each question on new page ONLY of your answer sheets.
2. Please write your correct exam no without fail on the answer sheets as well as the question papers.

Q. 1 Answer the following [Any Two]: [06]

a) Describe the data mining architecture and the purpose of each component.

b) What is a data quality issue in data analysis? List feature selection techniques and explain any two methods.

c) Explain the data preprocessing techniques in detail. List two methods for each technique.

Q. 2 Answer the following: [12]

1. Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result
   a) Calculate the mean, median and standard deviation of age and %fat.
   b) Draw the boxplots for age and %fat.
   c) Normalize the two variables based on z-score normalization.
   d) Calculate the correlation coefficient. Are these two variables positively or negatively correlated?

2. Employee dataset contains an attribute AGE with values in increasing order: { 13, 15, 16, 16, 19, X, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, Y, 45, 46, 52, 70}.
   a) Explain imputation using KNN.
   b) Find the value of the X and Y for above data using KNN.
   c) Find the values of X and Y using Median value replacement.

   OR

   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 using both binning methods.

3. Build a Decision Tree for classification using the training data for the following table. Divide the Height attribute into 3 ranges as follows: < 1.6, 1.6-1.8, > 1.8.

	Gender	Height	Class
1	F	1.58	Tall
2	M	1.58	Medium
3	M	1.7	Medium
4	F	1.65	Tall
5	F	1.85	Tall
6	F	1.4	Short
7	M	1.4	Short
8	M	1.7	Medium
9	F	1.75	Tall
10	M	1.82	Tall
11	F	1.6	Tall

Q. 3 Consider following transaction table and support=30% to answer the following: **[12]**

TID	Items
1	{I1, I3, I6}
2	{I4, I6}
3	{I1, I2, I5, I7}
4	{I2, I8, I5, I1}
5	{I1, I2, I4, I5, I3}
6	{I1, I2, I5 }
7	{I6, I1, I3}
8	{I3, I2, I5}
9	{I6, I7}
10	{I1, I3, I2, I4, I5}

a) Generate the FP-Tree and also generate conditional FP-Tree for the item having least support in f-list.

b) Fill the blank and justify the statement for the given example: "The change in dataset format from horizontal to vertical, may impact the _____ of Association rule mining approach".

c) Pruning ratio is defined as the percentage of itemsets not considered to be a candidate because (1) they are not generated during candidate generation or (2) they are pruned during the candidate pruning step. What is the pruning ratio of the Apriori algorithm for the given table?

d) Discover all frequent and closed itemsets. What is the percentage of closed itemsets compared to frequent itemsets?

OR

d) Discover the association rules of k=3 having the 100% confidence.

*