Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL

A Constituent unit of MAHE, Manipal

VI SEMESTER B.TECH. (COMPUTER SCIENCE & ENGINEERING)

END SEMESTER EXAMINATIONS, MAY 2024

SUBJECT: DATA WAREHOUSE AND DATA MINING [CSE 4060]

REVISED CREDIT SYSTEM
(--/05/2024)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer **ALL FIVE** questions.
- Missing data may be suitably assumed.

1A.	Summarize how user interaction poses an issue in Data Mining	**4**
1B.	Evaluate the four data sources from which data comes into data warehouses	**4**
1C.	With the help of an example present a detailed outline of how STAR schema optimizes navigation	**2**
2A.	Examine any 5 features of relational database systems for enforcing quality control	**5**
2B.	Present a detailed outline of the different OLAP models of Data Warehouses	**3**
2C.	Suppose that a patient record table (Table Q16) contains the attributes name, gender, fever, cough, test-1, test-2, test-3, and test-4, where name is an object identifier, gender is a symmetric attribute, and the remaining attributes are asymmetric binary. Determine which of the two individuals are unlikely to have similar diseases Table 2C Relational Table Where Patients Are Described by Binary Attributes *name* *gender* *fever* *cough* *test-1* *test-2* *test-3* *test-4* Jack M Y N P N N N Jim M Y Y N N N N Mary F Y N P N P N ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮	**2**
3A.	For the given dataset, assuming minimum support is set to a value of 2, determine all frequent itemsets using the FP-Growth algorithm. Show the detailed steps by constructing the Conditional (Sub-)Pattern Bases and also show the conditional FP-tree associated with the conditional node I3 using pictorial representation.	**5**

	<table><tr><th>TID</th><th>List of item_IDs</th></tr><tr><td>T100</td><td>I1, I2, I5</td></tr><tr><td>T200</td><td>I2, I4</td></tr><tr><td>T300</td><td>I2, I3</td></tr><tr><td>T400</td><td>I1, I2, I4</td></tr><tr><td>T500</td><td>I1, I3</td></tr><tr><td>T600</td><td>I2, I3</td></tr><tr><td>T700</td><td>I1, I3</td></tr><tr><td>T800</td><td>I1, I2, I3, I5</td></tr><tr><td>T900</td><td>I1, I2, I3</td></tr></table>	
3B.	Formulate the Apriori algorithm to discover frequent itemsets for mining Boolean association rules	**3**
3C.	For the dataset shown in Question 3A extract frequent itemsets using Vertical Data Format. Assume minimum support =2	**2**
4A.	For the following dataset determine the best split for the attribute income and age using Gini Index Class-Labeled Training Tuples from the *AllElectronics* Customer Database <table><tr><th>RID</th><th>age</th><th>income</th><th>student</th><th>credit_rating</th><th>Class: buys_computer</th></tr><tr><td>1</td><td>youth</td><td>high</td><td>no</td><td>fair</td><td>no</td></tr><tr><td>2</td><td>youth</td><td>high</td><td>no</td><td>excellent</td><td>no</td></tr><tr><td>3</td><td>middle_aged</td><td>high</td><td>no</td><td>fair</td><td>yes</td></tr><tr><td>4</td><td>senior</td><td>medium</td><td>no</td><td>fair</td><td>yes</td></tr><tr><td>5</td><td>senior</td><td>low</td><td>yes</td><td>fair</td><td>yes</td></tr><tr><td>6</td><td>senior</td><td>low</td><td>yes</td><td>excellent</td><td>no</td></tr><tr><td>7</td><td>middle_aged</td><td>low</td><td>yes</td><td>excellent</td><td>yes</td></tr><tr><td>8</td><td>youth</td><td>medium</td><td>no</td><td>fair</td><td>no</td></tr><tr><td>9</td><td>youth</td><td>low</td><td>yes</td><td>fair</td><td>yes</td></tr><tr><td>10</td><td>senior</td><td>medium</td><td>yes</td><td>fair</td><td>yes</td></tr><tr><td>11</td><td>youth</td><td>medium</td><td>yes</td><td>excellent</td><td>yes</td></tr><tr><td>12</td><td>middle_aged</td><td>medium</td><td>no</td><td>excellent</td><td>yes</td></tr><tr><td>13</td><td>middle_aged</td><td>high</td><td>yes</td><td>fair</td><td>yes</td></tr><tr><td>14</td><td>senior</td><td>medium</td><td>no</td><td>excellent</td><td>no</td></tr></table>	**5**
4B.	Outline the working of the .632 Bootstrap algorithm. Justify the presence of the figure 63.2% in this technique	**3**
4C.	Using sufficient illustration and detailed explanation elaborate on the working of K Nearest Neighbor classifier.	**2**
5A.	Present a detailed overview of the basic clustering methods available in the field of Data Mining	**4**
5B.	With the help of a diagram and proper explanation illustrate how BIRCH performs hierarchical clustering of data.	**4**
5C.	Using proper illustration summarize the working of the k-medoids clustering method	**2**