SOMAIYA VIDYAVIHAR UNIVERSITY

Semester: January 2023 – May 2023
Examination: In-Semester Examination

Maximum Marks: 30 **Duration : 1 hour & 15 mins**

Programme code: 1 **Programme: B.Tech Computer Engineering**	**Class:** TY	**Semester:** VI **(SVU 2020)**
Name of the Constituent College: **K. J. Somaiya College of Engineering**	**Name of the department:** COMP - Honours (DSA)	
Course Code: 116h54C601	**Name of the Course: Advanced Data Mining**	

Question No.		Max. Marks	CO Mapped	BT Level
Q1	Big Basket saw the following transactions from its customers and based on it they wish to identify the possible cases of bundle pricing: **Transaction ID** / **Itemset** 1 — Apple, Banana, Basil, Kiwi, Watermelon, Orange 2 — Grapes, Banana, Basil, Kiwi, Watermelon, Orange 3 — Apple, Jackfruit, Orange, Kiwi 4 — Apple, Tiramisu, Pears, Orange, Watermelon 5 — Pears, Banana, Orange, Kiwi Use Apriori algorithm Given the minimum support of 3, apply Apriori algorithm for generating all frequent itemsets. Identify 2 valid association rules given minimum confidence = 90%	10	CO1 & CO2	Applying & analyzing
Q2	Employ the DGIM algorithm. Shown below is a data stream with N=24 and current bucket configuration. New elements enter the window at the right. Thus the oldest bit of the window is the left-most bit shown **101011000101110110010110** a) Show how the initial stream will be divided into buckets b) What is the estimate of number of 1's in the latest k=14 bits of this window c) The following bits enter the window, one at a time: 10101011. What is the bucket configuration in the window after this sequence of bits has been processed by DGIM?	10	CO2	Applying, analyzing & understanding

<table>
<tr><td>Q2</td><td>OR

What is Frequent pattern mining over data streams? Explain Compact Pattern Stream tree Algorithm with example.</td><td>10</td><td></td><td></td></tr>
<tr><td>Q3</td><td>Answer any Two:

a) Term frequency matrix for 3 articles titled SAS, PAP and WH are given below. Q =jealous gossip. The unit vector for the query is (0, 0.77, 0.707). Using tf-idf score find the similarity between query and article. Identify top scoring article for query.

<table><tr><th>Term</th><th>SaS</th><th>PaP</th><th>WH</th></tr><tr><td>Affection</td><td>115</td><td>58</td><td>20</td></tr><tr><td>Jealous</td><td>10</td><td>7</td><td>11</td></tr><tr><td>Gossip</td><td>2</td><td>0</td><td>6</td></tr></table>
b) Give any 5 examples of data mining task for a real world dataset.
c) Explain in short Naïve Bayes Text Classification method.</td><td>(05+05)</td><td>CO1 & CO4</td><td>Applying & understanding</td></tr>
</table>