VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Continuous Assessment Test-2 Seniors

Winter 2021-22 Semester, (March-2022)

Course Name & code :Data Warehousing and Data Mining - CSI3010

Programme Name: M.Tech. (Integrated) - Computer Science and Engineering with Specialization in Data Science
M.Tech. (Integrated) - Computer Science and Engineering

Slot: G1+TG1

ANSWER ALL QUESTIONS 5*10=50 Marks

<table>
<tr><th>Q.No.</th><th>Question</th><th>Max Marks</th><th>CO</th><th>BL</th></tr>
<tr><td>1.</td><td>With detailed step-by-step visualize how to apply agglomerative clustering to the following data points. Draw the Dendrogram of the HAC Cluster. Let's consider we have 5 different CAR Models with 3 different continuous features and we want to see how we could cluster these CARS. First thing first, use Manhattan distance as our clustering distance measurement.
<table>
<tr><th>CAR Model</th><th>CAR Age</th><th>Cost in Lakh</th><th>Mileage in KM</th></tr>
<tr><td>Tata</td><td>3</td><td>10</td><td>17</td></tr>
<tr><td>Maruthi</td><td>2</td><td>07</td><td>21</td></tr>
<tr><td>Toyota</td><td>4</td><td>16</td><td>15</td></tr>
<tr><td>Honda</td><td>2</td><td>12</td><td>19</td></tr>
<tr><td>Mahindra</td><td>5</td><td>20</td><td>21</td></tr>
</table></td><td>10</td><td>CO4</td><td>BL4</td></tr>
<tr><td>2.</td><td>Demonstrate how well the silhouette coefficient is determined for the following clustering of data point. Assume that k-means clustering was already applied on the 1 dimensional data set for K=2.
Whereas the d1 is the distance between the data point to the centroid of 1 and d2 is the distance between the data points to the second centroid.</td><td>10</td><td>CO4</td><td>BL3</td></tr>
</table>

Datapoint	D1	D2	Cluster
2	2	9	C1
4	0	7	C1
10	6	1	C2
12	8	1	C2
3	1	8	C1
20	16	9	C2
30	26	19	C2
11	7	0	C2
25	21	14	C2

3.	Track the sequence of workflow in applying the DBScan Clustering algorithm by considering a typical data points.	10	CO4	BL3
4	Use the Apriori technique to determine the association rules in the Symptoms dataset. Choose Headache, Fever, Cold, Smell Loss, Diarrhea and Body pain as the parameters of the given dataset for analysis.	10	CO2	BL4

PatientID	Symptoms
1	Headache, Fever, Cold
2	Smell Loss, Fever, diarrhea
3	Headache, Smell Loss, Fever , diarrhea
4	Smell Loss, diarrhea
5	Body Pain

Assume min-support as 2 and min confidence as 60%

5.	Given the collections of symptoms in the question 4, apply the Frequent Pattern Growth approach with a minimum support as 2. Construct the FP-Tree corresponding to the set of symptoms given	10	CO2	BL4

**********ALL THE BEST**********