SLOT – G1

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

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

DIGITAL ASSIGNMENT – I - WINTER SEMESTER 2021-2022

Programme Name & Branch: B.Tech(CSE) Course Name: Data warehouse and Mining

Course Code: CSI3010

Handwritten and should be Uploaded in VTOP Mark split-up (4+3+3=10)

Uploading File Name : Reg.No_QuestionFileName

QUESTION-A

1. Draw a contingency table for each of the following rules using the transactions shown in Table 6.25.

Table 6.25. Example of market basket transactions.

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

Rules: $\{b\} \longrightarrow \{c\}$, $\{a\} \longrightarrow \{d\}$, $\{b\} \longrightarrow \{d\}$, $\{e\} \longrightarrow \{c\}$, $\{c\} \longrightarrow \{a\}$.

2. Narrate the procedure with sample dataset about the preference of filter approach and wrapper approach in feature/variable selection of data pre-processing of datamining.

3. Elaborate with your example about the usage of the Text Mining for Query Likelihood Estimation

QUESTION-B

1. Prove that the Sum of Square Errors (SSE) for a cluster Ci is proportional to the sum of distances between all points in the cluster . More specifically, prove that with the help of sample set of data as discussed in the class.

$$\sum_{x \in C_i} \|x - c_i\|^2 = \frac{1}{2m_i} \sum_{x \in C_i} \sum_{y \in C_i} \|x - y\|^2$$

where: mi is the number of points in the cluster C_i; c_i is the centroid of the cluster ; the clustered points are 2-dimensional real vectors; and ‖ ‖ is the Euclidean distance.

2. How the Performance of the machine learning model is being improvised with the help of various Categories of Encoders about the conversion of categorical variable to

3. Explain briefly about the process of Data Mining techniques to automatically discover and extract information from Web documents and services.

QUESTION-C

1. Let $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots., y_n)$ be two vectors of size n. The sample Pearson correlation coefficient is defined as

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_X)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_Y)^2}}$$

where μ_X, μ_Y are the mean values of vectors X and Y respectively.
Describe how we can use the cosine similarity to compute the sample Pearson correlation coefficient.

2 The goal of this question is to experiment with K-means clustering. You can do your own implementation of the K-means algorithm that we described in class, or use a suitable existing implementation. (K-means is implemented in R, WEKA, and there are several available.

3. Apply any one Frequent Item Set Estimation algorithm for Streaming Data

^^^^^^^^^^^^^^^^^^

Note : Sl.No 1-22 should choose Question –A, Sl.No 23-45 - Question –B , Others Question C.

Sl.No.	Register No.
1	19MID0001
2	19MID0003
3	19MID0005
4	19MID0006

5	19MID0007
6	19MID0008
7	19MID0010
8	19MID0012
9	19MID0013
10	19MID0015
11	19MID0016
12	19MID0017
13	19MID0019
14	19MID0020
15	19MID0023
16	19MID0024
17	19MID0027
18	19MID0028
19	19MID0029
20	19MID0030
21	19MID0031
22	19MID0034
23	19MID0035
24	19MID0038
25	19MID0039
26	19MID0040

27	19MID0041
28	19MID0042
29	19MID0044
30	19MID0046
31	19MID0048
32	19MID0049
33	19MID0051
34	19MID0052
35	19MID0054
36	19MID0056
37	19MID0057
38	19MID0058
39	19MID0059
40	19MID0061
41	19MID0062
42	19MID0067
43	19MID0068
44	19MID0069
45	19MID0075
46	19MID0076
47	19MID0077
48	19MID0087

49	19MID0088
50	19MID0090
51	19MID0093
52	19MID0094
53	19MID0095
54	19MID0097
55	19MID0107
56	19MID0108
57	19MID0110
58	19MID0111
59	19MID0115
60	19MID0116
61	19MID0124
62	19MID0125
63	20MIC0113
64	20MIC0134
65	20MIC0162