VIT
Vellore Institute of Technology

Final Assessment Test – June 2023

Course: ITA5007 - Data Mining and Business Intelligence

Class NBR(s): 0296 / 0528 / 0530 Slot: C2+TC2

Time: Three Hours Max. Marks: 100

Faculty Name : Prof. EPHZIBAH E.P/ Prof. HARSHITA PATEL/ Prof. JAGADEESAN S

KEEPING MOBILE PHONE/SMART WATCH, EVEN IN "OFF" POSITION IS TREATED AS EXAM MALPRACTICE

Answer ALL Questions

(10 X 10 = 100 Marks)

1. We have studied that data mining is the result of the evolution of database technology. Do you think that data mining is also the result of the evolution of machine learning research? Can you present such views based on the historical progress of this discipline? Address the same for the fields of statistics and pattern recognition.

2. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

3. Suppose we have the following dataset that represents the number of hours studied and the corresponding test scores for a group of students. You have to build a **linear regression model** to predict the test score based on the number of hours studied.

Hours Studied	Test Score
1	60
2	70
3	80
4	90
5	100
6	110
7	120
8	130
9	140
10	150

4. Outline the major steps of decision tree classification.

5. A database has 5 transactions. Let min support = 60% and min confidence = 80%.

TID	ITEM IDs
1	{M, O, N, K, E, Y}
2	{D, O, N, K, E, Y}
3	{M, A, K, E}
4	{M, U, C, K, Y}
5	{C, O, O, K, I ,E}

Find all frequent itemsets using FP-growth algorithm.

6. Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are:

A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9).

The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm for three clusters and show all the steps.

7. Apply complete-link agglomerative clustering to cluster the following data points and draw the dendrogram.

A1 = (1, 2), A2 = (2, 1), A3 = (2, 3), A4 = (3, 2), A5 = (8, 9), A6 = (9, 8), A7 = (9, 10)

8. Forecasting is a technique that uses historical data as inputs to make informed estimates that are predictive in determining the direction of future trends and help businesses to plan their strategies. Explain the methods of business forecasting in detail.

9. Differentiate between Explanatory versus Predictive modelling with appropriate examples.

10. Consider the given data:

Brightness	Saturation	Class
40	20	Red
50	50	Blue
60	90	Blue
10	25	Red
70	70	Blue
60	10	Red
25	80	Blue

Find out the class labels for following data using K nearest neighbor classifier for K=3 and K=5.

Brightness	Saturation	Class
20	35	?

⇔⇔⇔