VIT

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology and Engineering

Winter Semester 2022-2023 - Fresher

Continuous Assessment Test – I

Programme Name & Branch: MCA

Course Name & code: Data Mining and Business Intelligence (ITA5007)

Class Number (s): 0528, 0296, 0530

Slot: C2+TC2

Faculty Name (s): Harshita PateL, Dr. Ephzibah E.P., Jagadeesan S.

Exam Duration: 90 Min. **Maximum Marks: 50**

Q.No.	Question	Max Marks
1.	There is a strong linkage between statistical data analysis and data mining. Some people think of data mining as an automated and scalable method for statistical data analysis. Do you agree or disagree with this perception? Present one statistical analysis method that can be automated and/or scaled up nicely by integration with the present data mining methodology.	10
2.	Briefly outline how to compute the dissimilarity between objects described by the following: (a) Nominal attributes (b) Asymmetric binary attributes (c) Numeric attributes (d) Term-frequency vectors	10
3.	Use these methods to normalize the following group of data: 200, 300, 400, 600,1000 (a) min-max normalization by setting min = 0 and max = 1 (b) z-score normalization (c) normalization by decimal scaling	10
4.	Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results: (table below) (a) Calculate the mean, median, and standard deviation of age and %fat. (b) Draw the boxplots for age and %fat. (c) Draw a scatter plot and a q-q plot based on these two variables.	10
5.	Consider the following data (in increasing order) for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22,|22, 25, 25, 25, 25, 30, 33, 33, 35,|35, 35, 35, 36, 40, 45, 46, 52, 70. (a) Use smoothing by bin means to smooth these data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data. (b) How might you determine outliers in the data? (c) What other methods are there for data smoothing?	10

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

VIT

Vellore Institute of Technology

School of Information Technology and Engineering

Winter Semester 2022-2023

Continuous Assessment Test – II

Programme Name: MCA

Course Name & code: Data Mining and Business Intelligence - ITA5007

Class Number (s): 0528, 0296, 0530 **Slot: C2+TC2**

Faculty Name (s) (Dr. E.P.Ephzibah, Dr.Harshita Patel and Dr.S.Jagadeesan)

Exam Duration: 90 Min. **Maximum Marks: 50**

Answers all the Questions (5*10=50 Marks)

1. Use ID3 algorithm to construct a decision tree from the given data. Age, Competition, and Type are the input attributes. The Class is the output attribute with class labels Up and Down. Draw the generated decision tree with appropriate labels and node information.**(10)**

S.No	Age	Competition	Type	Class(Profit)
1	Old	Yes	Software	Down
2	Old	No	Software	Down
3	Old	No	Hardware	Down
4	Mid	Yes	Software	Down
5	Mid	Yes	Hardware	Down
6	Mid	No	Hardware	Up
7	Mid	No	Software	Up
8	New	Yes	Software	Up
9	New	No	Hardware	Up
10	New	No	Software	Up

2. The table given below has five weeks' sales data in Rupees (thousands) that deals with one dependent (y) and one independent variable (x). Implement a linear regression model on the data to find the line of regression. Predict the 7th and 12th week sales. **(10)**

S.No	Week(x)	Sales in Rupees (thousands) (y)
1	1	1.2
2	2	1.8
3	3	2.6
4	4	3.2
5	5	3.8

3. The table given below lists the training instances. Each training instance has two input attributes x1, x2, and one output attribute with class labels 1 and 0. Classify the new incoming instance t1= (3, 7), with k=3 using K-Nearest Neighbour algorithm. Give your observations for assigning an even number to the parameter k. **(10)**

Training Instance	X1	X2	Output
I1	7	7	0
I2	7	4	0
I3	3	4	1
I4	1	4	1

4. Consider the training samples in the dataset given below. Let the test instance be X= (Slow, Rarely, No). Find the most appropriate class label for the given record, X using the Naive Bayes classifier. **(10)**

Swim	Fly	Crawl	Class label
Fast	No	No	Fish
Fast	No	Yes	Animal
Slow	No	No	Animal
Fast	No	No	Animal
No	Short	No	Bird
No	Short	No	Bird
No	Rarely	No	Animal
Slow	No	Yes	Animal
Slow	No	Yes	Fish
Slow	No	Yes	Fish
No	Long	No	Bird
Fast	No	No	Bird

5. The following table shows six transactions by customers at grocery store. Let minimum support is 33.34% and minimum confidence is 60%. Find all frequent item sets and the generated strong rules using the Apriori algorithm.**(10)**

Transaction Id	Items
T1	HotDogs,Buns,Ketchup
T2	Hotdogs, Buns
T3	HotDogs,Coke,Chips
T4	Chips,Coke
T5	Chips,Ketchup
T6	HotDogs,Coke,Chips

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

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	?

⇔⇔⇔