Roll No. | | | | | | | | | | |

ANNA UNIVERSITY:: CHENNAI - 25

B.E(FT) END SEMESTER EXAMINATIONS – NOV/DEC 2023

Computer Science and Engineering

Seventh Semester

CS6001 & DATA MINING

(Regulation 2018 - RUSA)

Time:3 Hours | Max.Marks: 100

CO1	Demonstrate the knowledge of the ethical considerations involved in Data Mining.
CO2	Examine data and select suitable methods for data analysis.
CO3	Integrate various Classification, Clustering, Association rule mining techniques on real world data.
CO4	Synthesize the different algorithms and analyze it with the support of tools.
CO5	Interpret the concept of Spatial, Multimedia and Distributed, text and web mining and able to retrieve the data, analyze and make decision.

Blooms Level: L1- Remember L2- Understanding L3- Apply L4- Analyze L5- Evaluate L6- Create

	Answer All Questions **PART - A (10 x 2 = 20 Marks)**	**CO**	**BL**
1.	"Data mining turns a large collection of data into knowledge" – Comment this saying with an example.	CO1	L1
2.	Say at what step of KD process, Visualization is needed? Why?	CO1	L1
3.	Use min-max normalization by setting *min* D 0 and *max* D 1, to *normalize* the data 200, 300, 400, 600,1000	CO2	L2
4.	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. Use *smoothing by bin means* to smooth these data, using a bin depth of 3	CO2	L2
5.	Imagine that you are a sales manager at *AllElectronics*, and you are talking to a customer who recently bought a PC and a digital camera from the store. What should you recommend to her next?	CO3	L2
6.	A medical researcher wants to analyze breast cancer data to predict which one of three specific treatments a patient should receive. What kind of data mining task, he has to do? Why?	CO3	L2
7.	Distinguish the Data Mining situation where one can apply regression and correlation.	CO3	L1
8.	Suppose we have a data set of handwritten digits, where each digit is labeled as either 1, 2, 3, and so on. Take the number 2, for example. Some people may write it with a small circle at the left top part, while some others may not. What technique can be applied to determine subclasses for "2,"? Why?	CO3	L2

9.	How is the ensemble of classifiers used to predict the class label of a tuple, X after boosting?	CO3	L1
10.	What is a *recommender system*? In what ways does it differ from a customer or product based clustering system?	CO3	L1

PART – B (8 x 8 = 64 marks)

(Answer any 8 questions)

11.	Justify the view that data mining is either the result of the evolution of database technology or the result of the evolution of machine learning research. Write such views using the historical progress of this discipline. Also discuss the similar fields of data mining like statistics and pattern recognition.	8	CO1	L2
12.	*Outliers* are often discarded as noise. However, one person's garbage could be another's treasure. For example, exceptions in credit card transactions can help us detect the fraudulent use of credit cards. Using fraudulence detection as an example, propose two methods that can be used to detect outliers and discuss which one is more reliable.	8	CO2	L3
13.	Suppose a group of 12 sales price records has been sorted as follows: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215. Partition them into three bins by each of the following methods: equal-frequency (equal-depth) partitioning, equal-width partitioning and clustering.	8	CO2	L3
14.	Robust data loading poses a challenge in database systems because the input data are often dirty. In many cases, an input record may miss multiple values; some records could be *contaminated*, with some data values out of range or of a different data type than expected. Work out an automated *data cleaning and loading* algorithm so that the erroneous data will be marked and contaminated data will not be mistakenly inserted into the database during data loading.	8	CO2	L3
15.	A database has five transactions. Let min sup D 60% and min conf D 80%. TID | Items_bought T100 | {M,O,N,K,E,Y} T200 | {D,O,N,K,E,Y} T300 | {M, A, K, E} T400 | {M, U,C, K, Y} T500 | {C,O,O,K,I,E} Find all frequent itemsets using Apriori and illustrate the procedure.	8	CO3	L4
16.	Given an INPUT "Data partition, *D*,; *attribute list*, and *Attribute selection method*," Produce OUTPUT " Decision Tree". Write the algorithm for inducing a decision tree in a standard format	8	CO3	L2
17.	Compare the advantages and disadvantages of *eager* classification (e.g., decision tree, Bayesian, neural network) versus *lazy* classification (e.g., *k*-nearest neighbor, case-based reasoning).	8	CO3	L2
18.	Suppose that the data mining task is to cluster points (with .(*x*, *y*) representing location) into three clusters, where the points are *A*1.(2,10),*A*2.(2,5),*A*3.(8,4),*B*1.(5,8),*B*2.(7,5),*B*3.(6,4),*C*1.(1,2),*C*2.(4,9), The distance function is Euclidean distance. Suppose initially we assign	8	CO4	L5

	*A*1, *B*1, and *C*1 Use the *k-means* algorithm to show the final three clusters.			
19.	A biologist assumes that there is a linear relationship between the amount of fertilizer supplied tomato plants and the subsequent yield of tomatoes obtained. Eight tomato plants, of the same variety, were selected at random and treated, weekly, with a solution in which x grams of fertilizer was dissolved in a fixed quantity of water. The yield, y kilograms, of tomatoes was recorded. Calculate the equation for the linear regression of y on x. Estimate the yield of a plant treated, weekly, with 3.2 grams of fertilizer	8	CO4	L5

Plant	A	B	C	D	E	F	G	H
X	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
Y	3.9	4.4	5.8	6.6	7.0	7.1	7.3	7.7

20.	The authors of scientific publications form a social network, where the authors are vertices and two authors are connected by an edge if they coauthored a publication. How one can measure the similarity or distance between two authors in the network?	8	CO5	L2
21.	Propose a few implementation methods for audio data mining. Can we integrate audio and visual data mining to bring fun and power to data mining? Is it possible to develop some video data mining methods? State some scenarios and your solutions to make such integrated audiovisual mining effective.	8	CO5	L4
22.	Suppose that your local bank has a data mining system. The bank has been studying your debit card usage patterns. Noticing that you make many transactions at home renovation stores, the bank decides to contact you, offering information regarding their special loans for home improvements. Describe a *privacy-preserving data mining* method that may allow the bank to perform customer pattern analysis without infringing on its customers' right to privacy.	8	CO5	L2

PART–C(2x8=16marks)

Answer all the Questions

23.	The following table consists of training data from an employee database. The data havebeen generalized. For example, "31 : : : 35" for age represents the age range of 31 to 35.For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row. Let status be the class-label attribute.			

Department	Status	Age	Salary	Count
Sales	Senior	31…35	46K….50K	30
Sales	Junior	26…30	26K…30K	40
Sales	Junior	31…35	31K…35K	40
Systems	Junior	21…25	46K…50K	20
Systems	Senior	31…35	66K…70K	5
Systems	Junior	26…30	46K…50K	3
Systems	Senior	41…45	66K…70K	3

		Marketing	Senior	36...40	46K...50K	10	8	CO2 CO3	L4
		Marketing	Junior	31...35	41K...45K	4			
		Secretary	Senior	46...50	36K...40K	4			
		Secretary	Junior	26...30	26K...30K	6			
	Design a multilayer feed-forward neural network for the given data. Label the nodes in the input and output layers. Show the weight values after one iteration of the backpropagation algorithm, given the training instance "*(sales, senior, 31 . . . 35, 46K . . . 50K)*". Indicate your initial weight values and biases and the learning rate used.								
24.	Why is naive Bayesian classification called "naive"? For the data in Question No.23, given a data tuple having the values "systems," "26 . . . 30," and "46–50K" for the attributes department, age, and salary, respectively, what would a naive Bayesian classification of the status for the tuple be?						8	CO2 CO3	L4