LOYOLA COLLEGE (AUTONOMOUS), CHENNAI – 600 034

B.Sc. DEGREE EXAMINATION – **STATISTICS**

FOURTH SEMESTER – **APRIL 2024**

UST 4602 – DATA MINING

Date: 15-04-2024 Dept. No. Max. : 100 Marks

Time: 09:00 AM - 12:00 NOON

	SECTION A - K1 (CO1)	
	Answer ALL the Questions -	**(10 x 1 = 10)**
1.	**Fill in the blanks**	
a)	A database management system consists of a collection of interrelated data is known as _____.	
b)	A _____ attribute is a numeric attribute with an inherent zero point.	
c)	_____ to communicate data clearly and effectively through graphical representation.	
d)	Data classification is a two-step process, consisting of a _____ and _____.	
e)	The units in the hidden layers and output layer are sometimes referred to as _____	
2.	**Define the following**	
a)	Data pre-processing	
b)	Equal height binning	
c)	Entropy	
d)	Association measures	
e)	Neural network	
	SECTION A - K2 (CO1)	
	Answer ALL the Questions	**(10 x 1 = 10)**
3.	**Match the following**	
a)	Machine learning - Decision Tree	
b)	Normalization - Spider diagram	
c)	Radar chart - Semi supervised	
d)	Gini coefficient - Random Forest	
e)	Posterior probability - Decimal scaling	
4.	**True or False**	
a)	Multidimensional data mining is also called as exploratory multidimensional data mining.	
b)	Binary attributes are nominal attributes with only three possible states.	
c)	A data set is considered normal if the decimal scaling of all data points is within -1 to 1.	
d)	In classification problem, test data are used to estimate the accuracy of the classification rules.	
e)	The units in the input layer are called input units.	
	SECTION B - K3 (CO2)	
Answer any TWO of the following		**(2 x 10 = 20)**
5.	Explain the various steps involved in knowledge discovery from data in data mining.	
6.	Discuss briefly different methods to identify outliers in the data set.	
7.	Use the following methods to normalize the group of data: 200, 300, 500, 800 and 900. (i) Min-max normalization by setting min=0 and max=1. (ii) Z-score normalization. (iii) Normalization by decimal scaling	

8. Discuss briefly the types of distance measures in data mining.

SECTION C – K4 (CO3)

Answer any TWO of the following **(2 x 10 = 20)**

9. Describe briefly the tabulation and visualization of bivariate data.

10. Explain the following terms:- (i) Network structure (ii) Backpropagation algorithm

11. From the data is given below, calculate Weight of Evidence (WOE) and Information Value (IV) and give the interpretation.

Feature (X)	No. of events	No. of non-events
A	90	2400
B	130	1300
C	80	3500
D	100	1210
E	90	1100

12. Explain the different types of classification techniques in data mining

SECTION D – K5 (CO4)

Answer any ONE of the following **(1 x 20 = 20)**

13. From the following data:-

X	10	5	7	19	11	18
Y	2	3	3	6	7	9

Calculate Mahalanobis distance to identify outliers and give the interpretation.

14. (i) Discuss briefly data mining techniques from various domains and give examples in real life. (12)
(ii) Explain Box plot and Bubble chart for graphical methods. (8)

SECTION E – K6 (CO5)

Answer any ONE of the following **(1 x 20 = 20)**

15. The PC world provided details for ten of the most economical laser printer (PC world, Apr 2020): The following data show the maximum printing speed in pages per minute (ppm) and the price (in US dollar) for each printer.

Names	A	B	C	D	E	F	G	H	I	J
Speed(ppm)	10	9	11	12	6	5	7	6	12	14
Price(in dollar)	95	90	90	105	75	75	80	85	110	115

(i) Estimate the regression line for training data. (14)
(ii) Calculate MAPE for training and testing data and give interpretation. (6)

16. The dataset for classification of a person having cancer disease or not from a particular place with a particular income in accordance to the reference to their gender.

City	Gender	Income	Illness
Chennai	Male	30637	No
Chennai	Female	51524	Yes
Chennai	Male	36373	Yes
Pune	Male	88096	No
Pune	Female	112089	No
Pune	Female	90662	No
Pune	Male	127263	Yes
Chennai	Male	66645	No

The prediction of new test data set is given below

City	Gender	Income	Illness
Chennai	Female	90000	?

Predict the illness of test data using Naïve Bayes classification algorithm.

$$$$$$$$$$$$$$$$$