LOYOLA COLLEGE (AUTONOMOUS), CHENNAI – 600 034

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

FOURTH SEMESTER – **APRIL 2023**

UST 4602 – DATA MINING

Date: 06-05-2023 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.	**Define the following**
a)	Attributes and Tuples.
b)	Equal width binning.
c)	Scatter plot.
d)	Association measures.
e)	Neural network brain function.
2.	**Fill in the blanks**
a)	_____ is the process of finding a model that describes and distinguishes data classes or concepts.
b)	An_____ is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known.
c)	_____ method calculates the distance of a point from the mean of a dataset taking into account the covariance of the data.
d)	In a given set of data, when a group of data points deviates from the rest of the data set, it is called _____.
e)	_____ allow class conditional independencies to be defined between subsets of variables.
SECTION A – K2 (CO1)	
Answer ALL the Questions	**(10 x 1 = 10)**
3.	**Match the following**
a)	Binary attribute — Bubble chart
b)	Data cleaning — Internal and leaf node
c)	Back propagation — Boolean
d)	To compare three variables — Neural network
e)	Decision tree — To remove noise and inconsistent data
4.	**TRUE or FALSE**
a)	Data characterization is a summarization of the general characteristics or features of a target class of data.
b)	The values of an interval scale attribute are measured in not fixed and unequal units.
c)	A data set is considered normal if the Z-scores of all data points are within 0 to ∞.
d)	Market basket analysis is one of the key techniques used by large relations to show associations between items.
e)	Activation function is used in the hidden layer as well as at the output layer of the network.

SECTION B- K3 (CO2)

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

5. What is meant by CRISP and describe each phases of CRISP in data mining.

6. The following data for the midterm marks: 83, 63, 77, 78, 90, 75, 49, 79, 52, 74. Calculate
   (a) Min-max normalization. (3)
   (b) Z-score normalization. (4)
   (c) Normalization by decimal scaling. (3)

7. Explain briefly Artificial Neural Network.

8. Explain the following terms:- (i) Any two distance measures (4)
   (ii) Back propagation algorithm. (6)

SECTION C – K4 (CO3)

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

9. From the data given below:

TID	Items
1	Bread, Butter, Peanut
2	Bread, Butter, Milk
3	Butter, Peanut
4	Bread, Peanut
5	Butter, Peanut, Milk

The association rule between $X \Rightarrow Y$ as follows:-
$\{Bread \Rightarrow Butter\}, \{Butter \Rightarrow Peanut\}, \{Bread, Butter \Rightarrow Peanut\}$. Calculate Support, Confidence and Lift and give the interpretation.

10. Describe the steps involved in data mining when viewed as a process of knowledge discovery from data.

11. Discuss briefly classification processes in data mining.

12. The following table consists of customer name, age, loan and default status as given below:-

Customer Name	Age	Loan	Default
John	25	40000	N
Smith	35	60000	N
Alex	45	80000	N
Jade	20	20000	Y
Kate	35	120000	Y
Mark	52	18000	Y
Anil	23	95000	N
Pat	40	62000	N
George	60	100000	N
Jim	48	220000	Y
Jack	33	150000	Y
Andrew	48	142000	?

By applying K^{th} nearest neighbour algorithm classification and to predict the Andrew default status (Yes or No).

SECTION D – K5 (CO4)

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

13. Suppose that a hospital tested the age and body fat data for 9 randomly selected adults with the following results: (Assuming S.No: 1 to 7 for training data and remaining testing data)

S. No	1	2	3	4	5	6	7	8	9
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

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

14. Consider a fictional dataset that describes the weather conditions for playing a game of golf. Given the weather conditions, each tuple classifies the conditions as fit ("Yes") or unfit ("No") for playing golf.

S. No	Outlook	Temperature	Humidity	Windy	Play Golf
1	Rainy	Hot	High	Weak	No
2	Rainy	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Sunny	Mild	High	Weak	Yes
5	Sunny	Cool	Normal	Weak	Yes
6	Sunny	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Rainy	Mind	High	Weak	No
9	Rainy	Cool	Normal	Weak	Yes
10	Sunny	Mild	Normal	Weak	Yes
11	Rainy	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Sunny	Mild	High	Strong	No

By using decision tree classification algorithm with 4 terminal nodes and to decide whether to play Golf or not.

SECTION E – K6 (CO5)

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

15. (i) Discuss briefly data mining techniques from various domains and give examples in real life. (12)

(ii) Discuss briefly Dual Axis chart and Spider chart for graphical methods. (8)

16. The dataset for classification of a person having heart 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	50638	No
Chennai	Female	41524	Yes
Chennai	Male	66373	Yes
Pune	Male	98096	No
Pune	Female	112088	No
Pune	Female	100662	No
Pune	Male	127263	Yes
Chennai	Male	56645	No

The prediction of new test data set is given below

City	Gender	Income	Illness
Chennai	Female	90000	?

By applying Naive Bayes classification algorithm and to predict the patient will have heart disease or not.

$$$$$$$