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

Answer ALL the questions **(10 x 2 = 20)**

1. Define Attributes and Tuples.
2. Distinguish between supervised and unsupervised learning.
3. What is meant by Z-score?
4. The following data for the midterm marks: 83, 63, 77, 78, 90, 75. Perform Min-max normalization.
5. Write any two uses of scatter plot.
6. Write any two advantages of classification.
7. Define Gini index.
8. State any two demerits of KNN classification.
9. What is Naïve Bayes classifier?
10. Differentiate between human brain and artificial neural network.

SECTION – B

Answer any FIVE questions **(5 x 8 = 40)**

11. Explain briefly knowledge discovery from data in data mining.
12. Use these methods to normalize the following group of data: 500, 600, 700, 800, 1400
    (a) Min-max normalization by setting min=0 and max=1.
    (b) Z-score normalization.
    (c) Normalization by decimal scaling.
13. Explain different methods to identify outliers in the data.
14. Discuss briefly scatter chart and Spider chart for graphical methods.
15. 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:-

    $\{Bread \Rightarrow Milk\}, \{Butter \Rightarrow Milk, Peanut\}, \{Bread, Butter \Rightarrow Milk\}$.

    Calculate Support, Confidence and Lift and give the interpretation.
16. Explain the following terms:- (i) Distance measures (ii) Activation function.
17. Write the algorithm for KNN classification given k, the nearest number of neighbours, and n, the number of attributes describing each tuple.
18. Briefly discuss back propagation algorithm using ANN.

SECTION – C

Answer any TWO questions **(2 x 20 = 40)**

19. Explain data mining techniques from various domains and give example in real life.

20. The explore student data set with exam score and no. of hours they spent studying as given below:-

S. No	1	2	3	4	5
Scores	40	53	96	77	88
No. of hours	8	9	5	6	4

Calculate Mahalanobis distance to identify outliers and give the interpretation.

21. (i) Discuss briefly classification processes in data mining. (12)
(ii) Compare the advantages and disadvantages of classifier (a) Decision tree (b) Naive bayes classifier. (8)

22. 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**
Bangalore	Male	40629	No
Delhi	Female	43524	Yes
Delhi	Male	56373	Yes
Delhi	Male	88096	No
Bangalore	Female	92089	No
Bangalore	Female	110662	No
Delhi	Male	137263	Yes
Delhi	Male	76645	No

For the new test data set is given below apply Naïve Bayes classification algorithm and predict the patient illness.

City	**Gender**	**Income**	**Illness**
Delhi	Female	150000	?

$$$$$$$