|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Reg. No.** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

 **SRM Institute of Science and Technology**

SET B

**College of Engineering and Technology**

**School of Computing**

SRM Nagar, Kattankulathur – 603203, Chengalpattu District, Tamil Nadu

**Academic Year: 2022-23 (Odd)**

**Test: CLA-T1** **Date:** 08.09.2022

**Course Code & Title: 18CSE355T - Data Mining and Analytics** **Duration:** 1 Hour

**Year & Sem: III Year / V Sem** **Max. Marks:** 25 Marks

**Course Articulation Matrix: *(to be placed)***

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **S.No.** | **Course Outcome** | **PO1** | **PO2** | **PO3** | **PO4** | **PO5** | **PO6** | **PO7** | **PO8** | **PO9** | **PO10** | **PO11** | **PO12** |
| 1 | CO1 | 2 | 2 | 1 |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Part – A (15 x 1 = 15 Marks)**  **Instructions: 1. Answer all questions.**  **2. The duration for answering the part A is 20 minutes (this sheet will be collected after 20 minutes)** | | | | | | |
| **Q. No** | **Question** | **Mark** | **BL** | **CO** | **PO** | **PI Code** |
| 1. | The output of KDD is \_\_\_\_\_  A. Data  B. Information  C. Query  D. Useful Information  **Answer:**  D. Useful Information | 1 | 1 | 1 | 1 | 1.6.1 |
| 2. | The data objects that do not comply with the general behaviour or model is called \_\_\_\_\_  A. Outliers  B. Prediction  C. Evolution analysis  D. Classification  **Answer:**  A. Outliers | 1 | 1 | 1 | 1 | 1.6.1 |
| 3. | “Handling complex types of data” in data mining algorithms issues comes under the category \_\_\_\_  A. Diverse Data type  B. Descriptive data type  C. Mining Methodology and user interaction issue  D. Performance issue  **Answer:**  A. Diverse Data type | 1 | 1 | 1 | 1 | 1.6.1 |
| 4. | Which of the following analyse class-labelled (training) data sets?  i. Clustering, ii. Classification, iii. regression  A. i only  B. ii only  C. ii and iii  D. i, ii and iii  **Answer:**  C. ii and iii | 1 | 1 | 1 | 1 | 1.6.1 |
| 5. | "Privacy-preserving data mining" - This issue falls under which category of issues in data mining?  A. Mining Methodology  B. Performance Issues  C. Data mining and Society  D. User Interaction  **Answer:**  C. Data mining and Society | 1 | 1 | 1 | 1 | 1.6.1 |
| 6. | The values of a/an \_\_\_\_are measured in fixed and equal units.  A. Ratio-Scaled Attribute  B. Time-Scaled Attribute  C. Interval-Scaled attribute  D. Point-Scaled Attribute  **Answer:**  C. Interval-Scaled attribute | 1 | 1 | 1 | 1 | 1.6.1 |
| 7. | Identify which is the ratio scaled attribute among the given attributes  A. Rank of officers  B. Temperature in Fahrenheit  C. Temperature in Kelvin  D. Grades of students  **Answer:**  C. Temperature in Kelvin | 1 | 1 | 1 | 1 | 1.6.1 |
| 8. | What is the term for the median of the lower half of the data?  A. Q1  B. Q3  C. IQR  D. Maximum  **Answer:**  A. Q1 | 1 | 1 | 1 | 1 | 1.6.1 |
| 9. | The analysis tools pre-compute the summaries of the huge amount of data for what purpose?  A. In order to maintain consistency  B. For authentication  C. For data access  D. To obtain the query response  **Answer:**  D. To obtain the query response | 1 | 1 | 1 | 1 | 1.6.1 |
| 10. | What is a median?  A. Difference between higher half and lower half of the data set  B. Mean of the highest and lowest number in a data sample  C. Value separating higher half from the lower half of a data sample  D. Difference between the highest and lowest number.  **Answer:**  C. Value separating higher half from the lower half of a data sample | 1 | 1 | 1 | 1 | 1.6.1 |
| 11. | Given the mean of the dataset is 13 and the standard deviation is 5. Using z-score normalization to transform the value 8, gives a new value of  A. 0  B. 1  C. -1  D. -1.5  **Answer:**  C. -1 (new value = x- mew / Sigma) | 1 | 2 | 1 | 2 | 2.6.3 |
| 12. | Based on the box and whisker plot below, what is the IQR of the data?    A. 30  B. 25  C. 20  D.15  **Answer:**  B. 25 | 1 | 2 | 1 | 2 | 2.6.3 |
| 13. | What is the Median of the following data sample?  2, 7, 4, 8, 9, 10, 6, 12, 13  A. 8 B. 11  C. 9 D. 10  **Answer:**  A. 8  Sorting we get 2,4,6,7,8,9,10,12,13 | 1 | 2 | 1 | 2 | 2.6.3 |
| 14. | Use these method to normalize the following group of data: 200, 300, 400, 600 and 1000.  Use min-max normalization by setting min=0, and max=1 for the value 600.  A. 0.4 B. 0.5  C. 0.6 D. 0.7  **Answer:**  B. 0.5 | 1 | 2 | 1 | 2 | 2.6.3 |
| 15. | A survey was given to a random sample of 20 college students. They were asked, “how many textbooks do you own?”.  Their responses, were:  0, 0, 2, 5, 8, 8, 8, 9, 9, 10, 10, 10, 11, 12, 12, 12, 14, 15, 20, and 25.  Which among the following is TRUE?  A. Q1=20 B. Q3=0  C. Median=8 D. The value 20 is an outlier  **Answer:**  D. The value 20 is an outlier | 1 | 2 | 1 | 2 | 2.6.3 |
|  | **PART B (1 X 10 = 10 Marks)**  **Instruction: Answer either A or B** |  |  |  |  |  |
| 16 A. | i. Can all kinds of data be mined? Discuss about what kinds of data can be mined.  **Answer:**   * Database-oriented data sets and applications   + Relational database, data warehouse, transactional database * Advanced data sets and advanced applications   + Object-Relational Databases   + Temporal Databases, Sequence Databases, Time-Series databases   + Spatial Databases and Spatiotemporal Databases   + Text databases and Multimedia databases   + Heterogeneous Databases and Legacy Databases   + Data Streams   + The World-Wide Web   **Key:**  Any 8 types - brief explanation; 8 X 0.75 = 6 marks | 6 | 2 | 1 | 1 | 1.6.1 |
|  | ii. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order)  13, 15, 16, 16, 19, 20, 20, 21, 36, 40, 52, 70. Find out the Five number summary for the above data.  **Answer:**  Min: 13  Max: 70  Median: 13, 15, 16, 16, 19, 20,  20, 21, 36, 40, 52, 70  20  Q1: 16  Q2: 38  **Key:**  Min, Max – 1 mark  Median – 1 mark  Q1 – 1 mark  Q3 – 1 mark | 4 | 3 | 1 | 2 | 2.6.3 |
| 16 B. | i. Justify how attribute subset selection helps to make patterns easier to understand.  **Answer:**  Attribute subset selection: where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.  Attribute subset selection reduces the data set size by removing irrelevant or redundant attributes (or dimensions). The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes.  Mining on a reduced set of attributes has an additional benefit: It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand.  Methods:  Forward selection  Backward elimination  Combination of forward selection and backward elimination  Decision tree induction  **Key:**  Justification – 2 marks  4 methods – 4 \* 1 = 4 marks | 6 | 2 | 1 | 1 | 1.6.1 |
|  | ii. Normalize the following group of data:  1000,2000,3000,9000 using min-max normalization by setting min:0 and max:1  **Answer:**  Min: 1000  Max: 9000   |  |  |  | | --- | --- | --- | | 1000 | (1000-1000)/(9000-1000)\*(1-0) | 0 | | 2000 | (2000-1000)/(9000-1000)\*(1-0) | 0.125 | | 3000 | (3000-1000)/(9000-1000)\*(1-0) | 0.25 | | 9000 | (9000-1000)/(9000-1000)\*(1-0) | 1 |   **Key:**  For each new value 1 mark; 4 \* 1 = 4 marks | 4 | 3 | 1 | 2 | 2.6.3 |

\*Program Indicators are available separately for Computer Science and Engineering in AICTE examination reforms policy.

**Course Outcome (CO) and Bloom’s level (BL) Coverage in Questions**

**Approved by the Audit Professor/Course Coordinator**