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

 **SRM Institute of Science and Technology**

SET C

**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. | To retrieve information from large voluminous database about current and update details is called as  A. On-Line Transaction Processing (OLTP)  B. On-Line Analytical Processing (OLAP)  C. On-Line Data Processing (OLDP)  D. On-Line Privacy Processing (OLPP).  **Answer:**  A. On-Line Transaction Processing (OLTP) | 1 | 1 | 1 | 1 | 1.6.1 |
| 2. | “Suppose one wants to predict the number of new-born according to the size of storks' population by performing supervised learning”. This is an example of \_\_\_\_\_  A. Data Characterization  B. Association and Correlation  C. Data Discrimination  D. Regression  **Answer:**  D. Regression | 1 | 1 | 1 | 1 | 1.6.1 |
| 3. | ."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 |
| 4. | \_\_\_\_\_ is a summarization of the general characteristics or features of a target class of data.  A. Data characterization  B. Data discrimination  C. Association  D. Correlation  **Answer:**  A. Data characterization | 1 | 1 | 1 | 1 | 1.6.1 |
| 5. | "Incorporation of background knowledge" - This issue falls under which category of issues in data mining?  A. Mining Methodology  B. Performance Issues  C. Diversity of Data Types  D. User Interaction  **Answer:**  D. User Interaction | 1 | 1 | 1 | 1 | 1.6.1 |
| 6. | The example of symmetric binary attribute is \_\_\_\_\_  A. Pass and Fail  B. Positive and negative  C. Male and female D. True and False  **Answer:**  C. Male and female | 1 | 1 | 1 | 1 | 1.6.1 |
| 7. | Identify the type of attributes for the data set:  Grade: O, A, B+, B, B-  A. Binary  B. Nominal  C. Ordinal  D. Ratio scaled  **Answer:**  C. Ordinal | 1 | 1 | 1 | 1 | 1.6.1 |
| 8. | What is the term for the median of the upper half of the data?  A. Q1  B. Q3  C. IQR  D. Maximum  **Answer:**  B. Q3 | 1 | 1 | 1 | 1 | 1.6.1 |
| 9. | Match the data pre-processing techniques:  i. Data cleaning - a. Discretization  ii. Data Integration - b. Regression  iii. Data Reduction- c. Data value conflict detection and resolution  iv. Data Transformation- d. Principal Component analysis  A. i – c, ii – d, iii – a, iv-b  B. i – d,ii – b,iii – a,iv – c  C. i – b, ii – c, iii – d, iv – a  D. i – b , ii – d , iii – a , iv – c  **Answer:**  C. i – b, ii – c, iii – d, iv – a | 1 | 1 | 1 | 1 | 1.6.1 |
| 10. | The Mean occurs at a value less than the median implies that we have \_\_\_\_  A. Positively skewed data  B. Negatively skewed data  C. Symmetric data  D. Variance of data  **Answer:**  B. Negatively skewed data | 1 | 1 | 1 | 1 | 1.6.1 |
| 11. | Given the mean of the dataset is 21 and the standard deviation is 27. Using z-score normalization to transform the value 30, give a new value for it.  A. 0.33  B. 0.45  C. 0.55  D. 0.65  **Answer:**  A. 0.33  30-27/27 | 1 | 2 | 1 | 2 | 2.6.3 |
| 12. | What is the 5 number summary of the following data?  100, 150, 50, 30, 90, 50  A. 30, 50, 50, 100, 150  B. 30, 50, 90, 100, 150  C. 30, 50, 75, 100, 150  D. 30, 50, 100, 100, 150  **Answer:**  A, B, C, D  (Correct option of 30, 50, 70, 100, 150 was not there in the given options) | 1 | 2 | 1 | 2 | 2.6.3 |
| 13. | What is the Median of the following data sample?  3, 7, 4, 8, 9, 6, 10, 12, 13, 15  A. 7.5  B. 9  C. 8.5  D. 10  **Answer:**  C. 8.5 | 1 | 2 | 1 | 2 | 2.6.3 |
| 14. | Suppose a group of 12 sales price records has been sorted and partition for ‘3’ bins:  5,10,11,14,15,35,50,60,72,92,204,215.  Find smoothing by mean for 2nd bin using equal-frequency bins  A. 35, 35, 35, 35  B. 41, 41, 41, 41  C. 40, 40, 40, 40  D. 42, 42, 42, 42  **Answer:**  C. 40, 40, 40, 40  (Already sorted; 2nd bin is 15,35,50,60 – Total=160; Mean=40) | 1 | 2 | 1 | 2 | 2.6.3 |
| 15. | Dr.Mani is a dentist. He needs to report on the mean of the number of cavities that his patients have. The number of cavities reported from patients is: 1,0,1,5,6,3,4 (Corrected up to two decimal digits)  A. 3.33  B. 2.85  C. 3.00  D. 3.50  **Answer:**  B.2.85  Total=20; Mean=20/7=2.85 | 1 | 2 | 1 | 2 | 2.6.3 |
|  | **PART B (1 X 10 = 10 Marks)**  **Instruction: Answer either A or B** |  |  |  |  |  |
| 16 A. | i. Discuss about issues in data mining. Among these issues, which is the most important issue in your view? Justify?  **Key:**  Five groups:  Mining methodology,  User interaction,  Efficiency and scalability,  Diversity of data types, and  Data mining and society.  5 X 1 =5 marks  Justification = 1 mark | 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, 22, 22, 35, 35, 36, 60, 72, 80. Find out the Five number summary for the above data.  **Answer:**  13, 15, 16, 16, 19, 20, 20, 21,  22, 22, 35, 35, 36, 60, 72, 80  Min: 13  Max: 80  Median: 21.5  Q1: 17.5  Q2: 35.5  **Key:**  Min, Max – 1 mark  Median – 1 mark  Q1 – 1 mark  Q3 – 1 mark | 4 | 3 | 1 | 2 | 2.6.3 |
| 16 B. | i. Explain Data transformation by normalization.  **Key:**   * min-max normalization * z-score normalization * normalization by decimal scaling.   Explanation of 3 methods = 3 \* 2 = 6 marks | 6 | 2 | 1 | 1 | 1.6.1 |
|  | ii. Find the central tendency measures mean, mode and median for the following data set.  72,78,79,62,85,40,64,90,130,78,46,76.  **Answer:**  72,78,79,62,85,40,64,90,130,78,46,76.  Total: 900/12=75 Mean  Mode: 78  Median: Sorting we get  40, 46, 62, 64, 72, 76,  78, 78, 79, 85, 90, 130  Median: 77  **Key:**  Sorting – 1 mark  Mean – 1 mark  Mode – 1 mark  Median – 1 mark | 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**