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

 **SRM Institute of Science and Technology**

SET A

**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** | | | | | | |
| **Q. No** | **Question** | **Mark** | **BL** | **CO** | **PO** | **PI Code** |
| 1. | The data mining is a process of extracting and discovering patterns in large dataset. And it also known as \_\_\_\_  A. Pattern Recognition  B. Knowledge discovery from data  C. Business Intelligence  D. Process mining  **Answer:**  B. Knowledge discovery from data | 1 | 1 | 1 | 1 | 1.6.1 |
| 2. | The problem of finding abstracted patterns in the unlabeled data is \_\_\_\_\_  A. Classification B. Clustering  C. Outlier analysis D. Regression  **Answer:**  B. Clustering | 1 | 1 | 1 | 1 | 1.6.1 |
| 3. | The outlier analysis finds application in \_\_\_\_\_  A. Prediction of rain fail  B. Grouping of students  C. Fraud detection  D. Finding missing data values  **Answer:**  C. Fraud detection | 1 | 1 | 1 | 1 | 1.6.1 |
| 4. | \_\_\_\_ is/are an objective measure(s) of pattern interestingness  i. Support, ii. Confidence  A. i only B. ii only  C. i and ii D. Neither i, Nor ii  **Answer:**  C. i and ii | 1 | 1 | 1 | 1 | 1.6.1 |
| 5. | "Pattern evaluation" - 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:**  A. Mining Methodology | 1 | 1 | 1 | 1 | 1.6.1 |
| 6. | An attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is known as \_\_\_\_\_  A. Binary attribute  B. Nominal attribute  C. Numeric attribute  D. Ordinal attribute  **Answer:**  D. Ordinal attribute | 1 | 1 | 1 | 1 | 1.6.1 |
| 7. | Identify the type of attributes for the data set:  Color: Black, Brown, White, Green  A. Binary  B. Nominal  C. Ordinal  D. Ratio scaled  **Answer:**  B. Nominal | 1 | 1 | 1 | 1 | 1.6.1 |
| 8. | Which of the following lists all parts of the five number summary?  A. Mean, Median, Mode, Range, and Total  B. Minimum, Quartile 1, Median (Q2) , Quartile 3, and Maximum  C. Smallest, Q1, Q2, Q3, and Q4  D. Minimum, Maximum, Range, Mean, and Median  **Answer:**  B. Minimum, Quartile 1, Median (Q2), Quartile 3, and Maximum | 1 | 1 | 1 | 1 | 1.6.1 |
| 9. | \_\_\_\_\_ provides a multidimensional view of data and allows the pre-computation and fast access of summarized data  A. Data Warehouse  B. Data Mart  C. Data Cube  D. Business Intelligence  **Answer:**  C. Data Cube | 1 | 1 | 1 | 1 | 1.6.1 |
| 10. | The Mean occurs at a value greater than the median implies that we have \_\_\_\_  A. Positively skewed data  B. Negatively skewed data  C. Symmetric data  D. Variance of data  **Answer:**  A. Positively skewed data | 1 | 1 | 1 | 1 | 1.6.1 |
| 11. | Given the minimum and maximum values for attribute income are 10000, 90000, respectively. Using min-max normalization to transform the value 50000, find the new value for it in the new range [0.0, 1.0]  A. 0.45  B. 0.50  C. 0.55  D. 0.60  **Answer:**  B.0.50 | 1 | 2 | 1 | 2 | 2.6.3 |
| 12. | Based on the box and whisker plot below, what is the median of the data?    A. 30 B. 45  C. 20 D.15  **Answer:**  A. 30 | 1 | 2 | 1 | 2 | 2.6.3 |
| 13. | Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order)  20, 21, 22, 22, 25, 25, 30, 33, 33, 35, 35, 35, 35.  What is the median of the data?  A. 25 B. 30  C. 33 D. 35  **Answer:**  B. 30  (20, 21, 22, 22, 25, 25, 30,  33, 33, 35, 35, 35, 35) | 1 | 2 | 1 | 2 | 2.6.3 |
| 14. | 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, **25, 25, 25, 25**, 30,  33, 33, **35, 35, 35, 35,** 36, 40, 45, 46, 52, 70. What is the mode of the data?  A. 1 B. 2  C. 3 D. 4  **Answer:**  A, B, C, D.  (25, 35 occur 4 times. Options do not contain 25, 35. Hence A, B,C,D Can be considered as correct answer) | 1 | 2 | 1 | 2 | 2.6.3 |
| 15. | Compute the median value for the following data.  200,450,300,1500,700,44  A. 525  B. 515  C. 375  D. 500  **Answer:**  C. 375  Sorted List: 44 200 300 450 700 1500 | 1 | 2 | 1 | 2 | 2.6.3 |
|  | **PART B (1 X 10 = 10 Marks)**  **Instruction: Answer either A or B** |  |  |  |  |  |
| 16 A. | i. “The term Data mining is a misnomer”-Justify. Illustrate the knowledge discovery process. | 6 | 2 | 1 | 1 | 1.6.1 |
|  | **Answer:**  Justification  In data mining, data is not the output of mining process. Knowledge is mined out of data. Hence, it is a misnomer.  Knowledge Discovery (KDD) Process   * Learning the application domain   + relevant prior knowledge and goals of application * Identifying a target data set: data selection * Data processing   + **Data cleaning** (remove noise and inconsistent data)   + **Data integration** (multiple data sources maybe combined)   + **Data selection** (data relevant to the analysis task are retrieved from database)   + **Data transformation** (data transformed or consolidated into forms appropriate for mining)   (Done with data preprocessing)   * + **Data mining** (an essential process where intelligent methods are applied to extract   data patterns)   * + **Pattern evaluation** (identify the truly interesting patterns)   + **Knowledge presentation** (mined knowledge is presented to the user with   visualization or representation techniques)   * Use of discovered knowledge     **Key:**  Justification – 1 mark  Discovery process explanation -3 marks  KDD process diagram – 2 marks |  |  |  |  |  |
|  | ii. Draw a box and whisker plot for the following data: 72,78,79,62,85,41,64,90,130,70,46,76.  **Answer:**  After Sorting:  41, 46, 62, 64, 70, 72,  76, 78, 79, 85, 90, 130  Min: 41; Max:130; Median: (72+76)/2=74  Q1: 63; Q3: 82  IQR=19; IQR\*1.5=28.5  Q1-1.5\*IQR=34.5; Q3+1.5\*IQR=110.5  (130 can be plotted as outlier)  Box Plot:    **Key:**  Five number summary – 3 marks  Box plot – 1 mark | 4 | 3 | 1 | 2 | 2.6.3 |
| 16 B. | i. With an example discuss about the central tendency measures: Mean, Median and Mode.  **Key:**  Central tendency measures – overview – 1 mark  Mean – 1 mark  Median – 2 marks  Mode - 2 marks | 6 | 2 | 1 | 1 | 1.6.1 |
|  | ii. Apply the data smoothing methods - binning by mean, median and boundary for the following dataset: (Bin size:4)  8, 16, 9, 15, 21, 21, 24, 30, 26, 27, 30, 36  **Answer:**  After Sorting:  8, 9, 15, 16, (Total=48; Mean=12)  21, 21, 24, 26, (Total=92; Mean=23)  27, 30, 30, 36 (Total=123; Mean=30.75)  By mean:  12, 12, 12, 12  23,23,23,23  30.75,30.75,30.75,30.75  By median:  12, 12, 12, 12  22.5,22.5,22.5,22.5  30,30,30,30  By Boundary:  8, 8, 16, 16,  21, 21, 26, 26,  27, 27, 27, 36  **Key:**  Sorting & binning (3 bins of size 4) – 1 mark  By Mean – 1 mark  By median – 1 mark  By boundary – 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**