**SRM Institute of Science and Technology**

**Mode of Exam**

**OFFLINE**

**SET D**

**College of Engineering and Technology**

**School of Computing**

SR SRM Nagar, Kattankulathur – 603203, Chengalpattu District, Tamilnadu

**Academic Year: 2023-24 (ODD SMESTER)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **S.No.** | **Course Outcome** | **PO1** | **PO2** | **PO3** | **PO4** | **PO5** | **PO6** | **PO7** | **PO8** | **PO9** | **PO10** | **PO11** | **PO12** |
| 1 | **CO1** | L | H |  | H | L |  |  |  | L | L |  | H |
| 2 | **CO2** | M | H |  | H | L |  |  |  | M | L |  | H |
| 3 | **CO3** | M | H |  | H | L |  |  |  | M | L |  | H |
| 4 | **CO4** | M | H |  | H | L |  |  |  | M | L |  | H |
| 5 | **CO5** | H | H |  | H | L |  |  |  | M | L |  | H |

**Test: CLAT-1 Date: 09.08.2023**

**Course Code & Title:** **18CSE355T & DATA MINING AND ANALYTICS Duration: 1 Period**

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

**PART A**

**(10 x 1 = 10 Marks)**

**Instructions: Answer all Questions:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | Infer & choose the notation of values more than numeric attribute x on k th q-quantile for a given data distribution  Note: k is an integer such that 0<k<qa. k / q b. **(q - k) / q** c. 0 d. (k – q) / q | **1** | **B2** | **2** | **3** | **3.5.4** |
| 2 | Infer & choose the method / rules from the following one, which is used in classification techniques as for representing a derived model.  a. if-then Statement  b. if-then-else Statement  c. switch Statement  d. for Statement | **1** | **B2** | **2** | **3** | **3.5.4** |
| 3 | State the Graphic Display type from the option, which compare data across categories.a. Boxplot b. **Bar Chart**  c. Histogram d. Scatter Plot | **1** | **B1** | **2** | **1** | **1.7.1** |
| 4 | **Select the percentile in Q3 Quartile from distribution of equal-sized consecutive subsets of data.**   1. 25% 2. 50% 3. **75%** 4. 100% | **1** | **B2** | **2** | **2** | **2.6.2** |
| 5 | **Record the type of Quadrant as present of negative values in x- axis and y- axis on the plotting graph.**   1. Quadrant I 2. Quadrant II 3. **Quadrant III** 4. Quadrant IV | **1** | **B1** | **2** | **1** | **1.7.1** |
| 6 | **Make use of which attribute type does miles per hour (Speed = 20) are categorized as Fast / Slow?**   1. Ordinal 2. Nominal 3. **Binary** 4. Interval | **1** | **B2** | **3** | **2** | **1.7.1** |
| 7 | Specify the correct option from the following   1. Noise data unnecessarily decreases the storage capacity 2. Outliers are not present in the data 3. Outliers do belong to the range of the majority datapoint in the dataset 4. cause barrier in the result obtained | **1** | **B1** | **2** | **1** | **1.2.2** |
| 8 | Show the median value from the given dataset observations for 7 participants on a computer task.    a. **Medium** b. Slow c. Fast d. 0 | **1** | **B1** | **2** | **1** | **1.7.1** |
| 9 | Match the correct option from the observations of data as,  If a data set had values of 2, 4 and 6, the normalized value of the first data point is  a. 0  b. 0.5  c. 1  d. 1.5 | **1** | **B1** | **2** | **1** | **1.7.1** |
| 10 | Name the type of distribution in the given plotted graph from the dataset shown below:  a. Normal Distribution b. **Positively Skewed Distribution** c. Negatively Skewed Distribution d. Zero Skewed Distribution | **1** | **B2** | **3** | **1** | **2.7.1** |

**PART B** **(1 x 5 = 5 Marks)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **11** | Apply the concept of Inter Quartile Range (IQR) for the given scenario as:  Suppose the distribution of math scores in a class of 19 students in ascending order is:  59, 60, 65, 65, 68, 69, 70, 72, 75, 75, 76, 77, 81, 82, 84, 87, 90, 95, 98  Calculate median (Q2) from the observations of data in the dataset.  **Solution:**  First, mark down the median, Q2, which in this case is the 10th value: 75.  Q1 is the central point between the smallest score and the median.  In this case, Q1 falls between the first and fifth score: 68(5th Observation)  Note: that the median can also be included when calculating Q1 or Q3 for an odd set of values.  If We were to include the median on either side of the middle point, then Q1 will be the middle value between the first and 10th score, which is the average of the fifth and sixth score— (fifth + sixth)/2 = (68 + 69)/2 = 68.5).  Q3 is the middle value between Q2 and the highest score: 84(15th Observation)  (Or if We include the median, Q3 = (82 + 84)/2 = 83).  IQR = Q3 - Q2 = 84 -68 = 16  **Quartiles:**  Interpret their numbers.  A score of 68 (Q1) represents the first quartile and is the **25th percentile**.  Sixty-eight is the median of the lower half of the score set in the available data—that is, the median of the scores from 59 to 75.  Q1 tells us that 25% of the scores are less than 68 and 75% of the class scores are greater.  Q2 (the median) is the **50th percentile** and shows that 50% of the scores are less than 75, and 50% of the scores are **above** 75. Finally, Q3, the **75th percentile**, reveals that 25% of the scores are greater and 75% are less than 84. | 5 | B2 | 2 | 2 | 2.6.2 |
| **(OR)** | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **12** | **Analyze the sequence of operations on calculating the following for the given dataset:**  **x= 2, 7, 3, 12, 9**  **Methods:**   1. **Mean** 2. **Standard Deviation**   **Solution:**  Step 1: Calculate the mean 2 Marks  = 2 + 7 + 3 + 12 + 9 / 5  = 33 / 5  = 6.6    Step 2: Take each value in the data set, subtract the mean and square the difference.  Formula: - 1 Mark      **Variance:**  Computation – 2 Marks  **For instances, 5 values:**  (2 - 6.6)2 = 21.16  (7 – 6.6)2 = 0.16  (3 – 6.6)2 = 12.96  (12 – 6.6)2 = 29.16  (9 – 6.6)2 = 5.76  Squared Differences for all values are added:  21.16 + 0.16 + 12.96 + 29.16 + 5.76 = 69.20  Sum is then divided by number of data points:  69.20 ÷5 = 13.84  **Variance = 13.84**  **Standard Deviation**  Square Root of the variance  **Standard Deviation = 3.72** | **5** | **B2** | **2** | **2** | **2.6.5** |

**PART C (1 x 10 = 10 Marks)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **13** | **Draw a sketch & arrange the 9 Steps of Main Workflow in Knowledge Discovery of Databases. Discuss briefly about it.**  **Answer:**  Schematic Sketch / Diagram – 3 Marks  **KDD workflow with the main steps**  **(Or)**    **Explanations:** 7 Marks  1. Developing an understanding of the application domain - preparatory step for understanding what should be done with many decisions  if set wrong, can lead to false interpretations and negative impacts on the end-user  2.Selecting and creating a data set: this includes finding out what data is available and select a subset on which discovery will be performed  3.Pre-processing and cleaning: in this stage data reliability is enhanced, it includes data cleaning such as handling missing values and removal of noise or outliers, redundant and low-quality data from the data set in order to improve the reliability of the data  4.Data transformation: in this stage the generation of better data for the data mining is prepared and developed  5.Choosing the appropriate Data Mining task: we are ready to decide on which type of data mining to use, for example,  classification,  regression or  clustering  6.Choosing the data mining algorithm: this stage includes selecting the specific method and so algorithm to be used for searching patterns in the data.  7.Employing the data mining algorithm: finally, the implementation of the data mining algorithm is reached, and algorithms are applied in order to extract data patterns.  8.Evaluation of mined patterns: in this stage we evaluate and interpret the mined patterns with respect to the goals defined in the first step  9.Using the discovered knowledge: we are now ready to incorporate the knowledge into another system for further action | **10** | **B1** | **2** | **1** | **1.2.2** |
| **(OR)** | | | | | | |
| **14** | **Interpret the methods & elaborate the working of normalization in data mining with each following example:**  **Method:**  **i.) Min-Max Normalization**  **Dataset:**    **Solution:**  Minimum value - **8**  Maximum value - **20**  As this formula scales the data between 0 and 1,  The **new min** is 0  The **new max** is 1  V - respective value of the attribute, i.e., 8, 10, 15, 20  Formula : 1 Mark    Computation : 4 Mark  For 8 years of experience: v’= (8 – 8 / 20 – 8) ( 1 – 0) + 0 = 0  For 10 years of experience: v’ = (10 – 8 / 20 – 8) ( 1 – 0) + 0  = 0.16  For 15 years of experience: v’ = (15 – 8 / 20 – 8) ( 1 – 0) + 0  = 0.58  For 20 years of experience: v’ = (15 – 8 / 20 – 8) ( 1 – 0) + 0  = 1  **Method:**  **ii.) Z-Score Normalization**  **Dataset: 3,5,5,8,9,12,12,13,15,16,17,19,22,24,25,134**  **Solution:**  **Step 1: Find out Mean**  = 3 + 5 + 5 + 8 + 9 + 12 + 12 + 13 + 15 + 16 +17 + 19 + 22 +24 + 25 + 134 / 16  = 339 / 16  = 21.18 ~ 21.2  **Mean = 21.2**      **Step 2:**  Calculation on Standard Deviation – 2 Marks  Variance:  For instances, 16 values  (3 – 21.2)2 = 331.24  (5 – 21.2)2 = 262.44  (5 – 21.2)2 = 262.44  (8 – 21.2)2 = 29.16  (9 – 21.2)2 = 174.24  (12 - 21.2)2 = 84.64  (12 – 21.2)2 = 84.64  (13 – 21.2)2 = 67.24  (15 – 21.2)2 = 38.44  (16 – 21.2)2 = 27.04  (17 - 21.2)2 = 17.64  (19 – 21.2)2 = 4.84  (22 – 21.2)2 = 0.64  (24 – 21.2)2 = 7.84  (25 – 21.2)2 = 14.44  (134 – 21.2)2 = 12723.84  **The squared differences for all values are added**  Variance = 14130.76 / 16 = **883.17**  Standard Deviation:  Square Root of 883.17 = 29.718 ~ **29.8**  **Step 3:**  **Z-Score Normalization:**  **Formula: -** 1 Mark      New value = (3 - 21.2) / 29.8  ∴ New value = -0.61  Similarly for other Data,  Computation – 2 Marks    **Z- Score Normalized Value are**  (5 - 21.2) / 29.8 = -0.54  (5 - 21.2) / 29.8 = -0.54  (8 - 21.2) / 29.8 = -0.44  (9 - 21.2) / 29.8 = -0.41  (12 - 21.2) / 29.8 = -0.31  (12 - 21.2) / 29.8 = -0.31  (13 - 21.2) / 29.8 = -0.28  (15 - 21.2) / 29.8 = -0.21  (16 - 21.2) / 29.8 = -0.17  (17 - 21.2) / 29.8 = -0.14  (19 - 21.2) / 29.8 = -0.07  (22 - 21.2) / 29.8 = 0.03  (24 - 21.2) / 29.8 = 0.09  (25 - 21.2) / 29.8 = 0.13  (134 - 21.2) / 29.8 = 3.79 | **10** | **B3** | **3** | **4** | **4.5.1** |

**\*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**