****

**Sample Question Format**

**KIIT Deemed to be University**

**Online Mid Semester Examination (Spring Semester-2024)**

**Subject Name & Code:** **Data Mining and Data Warehousing (IT 3031**) **Applicable to Courses:B.Tech**

**Full Marks=20** **Time: 60 minutes**

**SECTION-A(Answer All Questions. All questions carry 2 Marks)**

**(5×2=10 Marks)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Question No** | **Question Type(MCQ/SAT)** | **Question** | **Answer Key(if MCQ)** | **CO Mapping** |
| **Q.No:1(a)** | **SAT** | Give four Example of descriptive data mining task. |  | CO1 |
|  | **SAT** | What are the major issues in data mining? |  | CO1 |
|  | **SAT** | Distinguish between MCAR and MAR. |  | CO1 |
|  | **SAT** | Write the seven steps of KDD. |  | CO1 |
| **Q.No:1(b)** | **SAT** | *If a* and *b*, are the binary attributes find SMC and Jaccard coefficient.  *a* = 1 0 1 0 0 1 0 0 0 1  *b* = 1 1 1 0 0 0 1 0 0 1 |  | CO1 |
|  | **SAT** | Calculate the trimmed 10 % Mean of the following data  [12, 8, 6, 5, 10, 9,11, 14, 2,3,5,7,6,13,15,18,10,14,2,1]. |  | CO1 |
|  | **SAT** | Calculate the median and mode of the following data  [5,9,9,8,7,6,11,12,10,8,6,5]. |  | CO1 |
|  | **SAT** | Given two objects represented by the tuples (24, 12, 40, 16) and (22, 9, 37, 8). Compute the Manhattan distance between the two objects. |  | CO1 |
| **Q.No:1(c)** | **SAT** | Differentiate between OLAP and OLTP. |  | CO2 |
|  | **SAT** | Differentiate between virtual data warehouses and data mart. |  | CO2 |
|  | **SAT** | Differentiate between Star schema and snowflake schema. |  | CO2 |
|  | **SAT** | Differentiate between data warehouses and data mart. |  | CO2 |
| **Q.No:1(d)** | **SAT** | 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, 23, 29,35, 41, 44, 53, 62, 69, 72.  Use Decimal Scaling to normalize the data. |  | CO2 |
|  | **SAT** | Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order):  25, 30, 35, 36, 40.  Use zero-mean Normalization to normalize the age. |  | CO2 |
|  | **SAT** | 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  Use min-max normalization to normalize the age onto the range [0,1] |  | CO2 |
|  | **SAT** | Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order):  20, 22, 25, 28, 30.  Use Z-score normalization to normalize the age. |  | CO2 |
| **Q.No:1(e)** | **SAT** | Differentiate between Probability Sampling and Non- Probability Sampling |  | CO2 |
|  | **SAT** | Distinguish between Systematic Sampling and Stratified Sampling |  | CO2 |
|  | **SAT** | Differentiate between judgemental Sampling and convenience Sampling. |  | CO2 |
|  | **SAT** | State about five common type of sampling error. |  | CO2 |

**SECTION-B(Answer Any One Question. Each Question carries 10 Marks)**

**(1×10=10 Marks)**

|  |  |  |
| --- | --- | --- |
| **Question No** | **Question** | **CO Mapping** |
| **Q.No:2** | Explain Pearson’s correlation coefficient. Suppose a survey was conducted in your city. Given is the following sample data containing a person's age and their corresponding income. Find out whether the increase in age has an effect on income using the correlation coefficient formula.  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Age | 26 | 31 | 37 | 44 | 50 | | Income | 25000 | 30000 | 40000 | 50000 | 60000 | | CO2 |
| **Q.No:3** | Is gender independent of education level? A random sample of 395 people were surveyed and each person was asked to report the highest education level they obtained. The data that resulted from the survey is summarized in the following table:   |  | High School | Bachelors | Masters | Ph.d. | Total | | --- | --- | --- | --- | --- | --- | | Female | 60 | 54 | 46 | 41 | 201 | | Male | 40 | 44 | 53 | 57 | 194 | | Total | 100 | 98 | 99 | 98 | 395 |   Question: Perform a chi square to find are gender and education level dependent at 5% level of significance? In other words, given the data collected above, is there a relationship between the gender of an individual and the level of education that they have obtained? (For the above case the critical value of chi square with 3 degree of freedom is 7.815) | CO2 |
| **Q.No:4** | For the following data set find the five-number summary, IQR, Tukey fence, and outlier (if any).Draw the box plot to describe the distribution of data in the data set  19, 33, 76, 29, 15, 41, 46, 26, 52, 38, 20, 32, 42, 24 | CO1 |
| **Q.No:5** | In a database, there are 1500 patient record out of which 500 are covid positive. A disease predictor application accesses the database and predicts 200 covid positive patients out of 250 positive case and 400 covid negative cases, find the following parameters, a) Sensitivity b) Specificity c) Misclassification Rate | CO1 |
| **Q.No:6** | Consider a Big Bazar scenario where the product set is P = {Rice, Pulse, Oil, Milk, Apple}. The database comprises six transactions where 1 represents the presence of the product and 0 represents the absence of the product.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Transaction ID | Rice | Pulse | Oil Milk | Apple | | t1 | 1 | 1 | 1 | 0 | | t2 | 0 | 1 | 1 | 1 | | t3 | 0 | 0 | 0 | 1 | | t4 | 1 | 1 | 0 | 1 | | t5 | 1 | 1 | 1 | 0 | | t6 | 1 | 1 | 1 | 1 |   Using apriori algorithm find all frequent item sets and association rules, if the minimum support is more than 50% and the minimum confidence is 80%. | CO3 |
| **Q.No:7** | With neat diagram briefly explain about different type of schema. | CO3 |

**Controller of Examinations**