** SPRING END SEMESTER EXAMINATION-2023**

*Semester: 8th*

*Programme:B.Tech..*

*Branch/Specialization: All Branches*

**8th Semester, B.Tech**

**Data Mining and Data Warehousing (MINOR)**

**IT 3031**

**(For 2019 Admitted Batches)**

**Time: 2 Hours Full Marks: 50**

**KIIT Deemed to be University**

**End Semester Examination(Spring Semester-2023)**

**SECTION-A(Answer All Questions. Each question carries 2 Marks)**

**Time:30 Minutes (7×2=14 Marks)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Question No** | **Question Type (MCQ/SAT)** | **Question** | **CO Mapping** | **Answer Key**  **(For MCQ Questions only)** |
| **Q.No:1** | **MCQ** | Which one of the following can be defined as the data object which does not comply with the general behavior (or the model of available data)?   1. Evaluation Analysis 2. Outliner Analysis 3. Classification 4. Prediction | CO-1 | B |
|  | **MCQ** | Which of the following refers to the steps of the knowledge discovery process, in which the several data sources are combined?   1. Data selection 2. Data cleaning 3. Data transformation 4. Data integration | CO-1 | D |
|  | **MCQ** | Which of the following statements is correct about data mining?   1. It can be referred to as the procedure of mining knowledge from data 2. Data mining can be defined as the procedure of extracting information from a set of the data 3. The procedure of data mining also involves several other processes like data cleaning, data transformation, and data integration 4. All of the above | CO-1 | D |
|  | **MCQ** | To discover the trends of global warming, which types of database is used for mining.   1. Time series database 2. Spatial database 3. Spatiotemporal database 4. Heterogeneous database | CO-1 | C |
| **Q.No:2** | **MCQ** | Which of the following is considered as a metadata for the OLAP cube?   1. Star Schema 2. Snowflake Schema 3. Relational Database 4. None of these | CO-2 | C |
|  | **MCQ** | Which OLAP operation is responsible for computation of all of the data relationship for one or more dimensions?   1. Roll-up 2. Slice 3. Dice 4. Pivot | CO-2 | A |
|  | **MCQ** | Find the correct option of an example of Data Cube.   1. Data transformation 2. Data generalization 3. Data cleaning 4. None of these | CO-2 | A |
|  | **MCQ** | A star schema has what type of relationship between a dimension and fact table?  A. Many-to-many  B. One-to-one  C. One-to-many  D. All of these | CO-2 | C |
| **Q.No:3** | **MCQ** | The third quartile of box plot could be of as:  (A) top 25%  (B) top 50%  (C) bottom 25%  (D) top 75% | CO-2 | A |
|  | **MCQ** | The process of Data Discretization is known as   1. to scale up data 2. to convert continuous attribute to discrete attribute 3. to convert discrete attribute to continuous attribute 4. None | CO-2 | B |
|  | **MCQ** | In which computation requires the distribution of data to be in sorted order?   1. Mode 2. Mean 3. Median 4. Variance | CO-2 | C |
|  | **MCQ** | If the scatter plot is a straight line from upper left to bottom right corner, what kind of correlation is represented?   1. Zero 2. Highly positive 3. Perfectly positive 4. Perfectly negative | CO-2 | D |
| **Q.No:4** | **SAT** | Compute the Jaccard similarity between the following two binary vectors  {x = 1101010001, y = 1100011000} | CO-2 |  |
|  | **SAT** | Compute the Jaccard similarity between the following two binary vectors  {x = 0101010011, y = 1100011010} | CO-2 |  |
|  | **SAT** | Compute the Jaccard similarity between the following two binary vectors  {x = 0101010011, y = 1101011010} | CO-2 |  |
|  | **SAT** | Compute the Jaccard similarity between the following two binary vectors  {x = 1101010111, y = 0100111010} | CO-2 |  |
| **Q.No:5** | **MCQ** | How many frequent 3-itemsets are there for the following transaction data items with minimum\_support = 60%?   |  |  | | --- | --- | | TID | Items | | T1 | B, N, P | | T2 | B,C, P, N | | T3 | B, P, E, N | | T4 | B, N, E, M | | T5 | N, C, P, E, M |  1. 1 2. 2 3. 0 4. 3 | C03 | A |
|  |  | How many frequent 2-itemsets are there for the following transaction data items with minimum\_support = 80%?   |  |  | | --- | --- | | TID | Items | | T1 | B, N, P | | T2 | B,C, P, N | | T3 | B, P, E, N | | T4 | B, N, E, M | | T5 | B,N, C, P, E, M |  1. 1 2. 2 3. 0 4. 3 | Co3 | D |
|  | **MCQ** | Choose the correct Confidence for (Bread→Milk), if support of Bread is 120, support of Milk is 50, and support of together purchased is 60?  A. 50 %  B. 100%  C. 70%  D. 80% | C03 | A |
|  | **MCQ** | Suppose the support of itemset {x, y} is 10, which of the following numbers are the possible supports of itemset {x , y, z}?  Options:   1. 9 , 10 2. 10, 11 3. 11, 12 4. Unable to predict | C03 | A |
| **Q.No:6** | **MCQ** | If a classifier correctly classify 90 tuples out of 120 tuples of test data what will be error rate of the classifier?   1. 45% 2. 25% 3. 75% 4. 65% | CO-4 | B |
|  | **MCQ** | Consider the given confusion matrix for a binary classification problem.   |  |  |  | | --- | --- | --- | |  | P(+) | P(-) | | A(+) | 340 | 40 | | A(-) | 20 | 90 |   Select the correct Accuracy%.   1. 96 % 2. 69.38% 3. 86.66% 4. 87.76% | CO-4 | D |
|  | **MCQ** | Consider the given confusion matrix for a binary classification problem.   |  |  |  | | --- | --- | --- | |  | P(+) | P(-) | | A(+) | 260 | 40 | | A(-) | 60 | 40 |   Select the correct True Positive Rate(TPR).   1. 96 % 2. 86.4% 3. 86.66% 4. 87% | CO-4 | C |
|  | **MCQ** | What will be the correct Mean Absolute Error rate for the test set?   |  |  | | --- | --- | | Y | Y’ | | 5 | 5 | | 6 | 4 | | 7 | 7 | | 8 | 9 | | 3 | 2 |  1. 90% 2. 88% 3. 80% 4. 98% | CO-4 | C |
| **Q.No:7** | **SAT** | What is web mining? List out some applications of it. | CO-6 |  |
|  | **SAT** | Explain spatial mining and list out some applications of it. | CO-6 |  |
|  | **SAT** | What is text mining? Write some applications of it. | CO-6 |  |
|  | **SAT** | Explain multimedia mining and list out some applications of it. | CO-6 |  |

**SECTION-B(Answer Any Three Questions. Each Question carries 12 Marks)**

**Time: 1 Hour and 30 Minutes** **(3×12=36 Marks)**

|  |  |  |
| --- | --- | --- |
| **Question No** | **Question** | **CO Mapping**  **(Each question should be from the same CO(s))** |
| **Q.No:8** | (i)The distribution below shows the scores on a player's test for 21 applicants. Identify the range of both low and high outliers in the given scatter plot. Based on Q1, Q2, Q3 and IQR Draw a box-plot for the above data with proper labelling.  pic2  (ii) Draw the architecture of a data mining system and describe each of its components briefly. | CO-1, CO-2 |
| 1. What is data warehouse? Give the steps for design and construction of data warehouses and explain it with three-tier architecture diagram.   b. Suppose two attributes of 5 products in a shop is given in the below table.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Price | 10.4 | 57.3 | 60 | 130 | 220 | | Weight | 0.7 | 2.6 | 1.1 | 4.7 | 9.4 |  1. Normalize the weight variables based on the min-max normalization (min = 10, max = 20)   ii. Normalize the price variables with z-score normalization  iii. Calculate the Pearson correlation coefficient. Are these two variables positively or negatively correlated? |
| a. Consider categorical data of the students of 4 different departments of a university based on their occupations. Calculate the chi-square value of the below contingency table.    b. Draw the flow chart of genetic algorithm and explain it’s steps briefly. |
| **Q.No:9** | Consider a transaction data shown in the table below from a fast food restaurant. Find out Maximal frequent item sets, closed itemsets and strong association rules with minimum support 22% and minimum confidence 75%. | CO-3 |
| A transaction table shown below, and assuming a minimum level of support mini\_sup = 60% and a minimum level of confidence min\_conf = 80%.  (a) Find all frequent itemsets using the Apriori algorithm.  (b) List all of the strong association rules, along with their support and confidence value.   |  |  | | --- | --- | | TID | Item Bought | | T100 | { B,A,D} | | T200 | { C,A,B,E,K} | | T300 | { D,A,C,E,B} | | T400 | { K,A,D,B} | | T500 | {K,B,D,E} | |
| Consider the following transaction database:     |  |  | | --- | --- | | TID | List of Items | | T1 | I1 , I2 , I3 , I4 , I5 , I6 | | T2 | I7 , I2 , I3 , I4 , I5 , I6 | | T3 | I1 , I8 , I4 , I5 | | T4 | I1 , I9 , I10 , I4 , I6 | | T5 | I10 , I2 , I4 , I11 , I5 |   List the frequent k-itemset for the largest k, and all the strong association rules with their support 60% and confidence 50% containing the frequent k-itemset for the largest k. |
| **Q.No:10** | a. Suppose the task is to cluster the following 8 points (with (x, y) representing location) using hierarchical clustering (single link, complete link and average link): A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9). The distance function is Euclidean distance.  b. Discuss the types of hierarchical clustering as well as the advantages and disadvantages of it. | CO-5 |
| a. Consider a database D, consisting of 7 transactions. Use this table to show the steps of k-means algorithm with Euclidean distance function. Use K=2 and suppose A and C are selected as the initial means. Show the updated cluster centers after first round of execution.   |  |  |  | | --- | --- | --- | | Samples | X1 | X2 | | A | 1 | 4 | | B | 9 | 5 | | C | 0 | 2 | | D | 2 | 4 | | E | 3 | 5 | | F | 6 | 9 | | G | 8 | 5 |   b. Discuss the limitations of K-means algorithm. |
| 1. Draw simple link, average link and complete link dendogram to represent the cluster for the given distance matrix.  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | P1 | P2 | P3 | P4 | P5 | | P1 | 0 |  |  |  |  | | P2 | 15 | 0 |  |  |  | | P3 | 8 | 7 | 0 |  |  | | P4 | 10 | 9 | 13 | 0 |  | | P5 | 16 | 12 | 5 | 11 | 0 |   b. Discuss, how clustering is different from classification and list out three popular distance measures. |
| **Q.No:11** | Refer to the Table, which illustrates the fish data-set. With this data-set, establish a simple linear regression model for each species by estimating the weight of the fish from its length. By demonstrating a step-by-step procedure, the regression model has to capture the values of the slope and intercept for each species. In the data-set, the weight of a fish is expressed in grams, and its length and height are in centimeters.   |  |  |  |  | | --- | --- | --- | --- | | Species | Weight | Length | Height | | Bream | 242 | 25.4 | 11.52 | | Bream | 290 | 26.3 | 12.48 | | Bream | 340 | 26.5 | 12.73 | | Bream | 363 | 29 | 12.75 | | Bream | 500 | 29.7 | 13.65 | | Bream | 1000 | 37 | 18.9 | | Roach | 200 | 23.5 | 7.3 | | Roach | 180 | 25.2 | 7.10 | | Roach | 290 | 26 | 8.88 | | Roach | 390 | 31.7 | 9.5 | | Roach | 160 | 22.5 | 6.5 | | Roach | 140 | 20.8 | 6.4 | | Roach | 40 | 14.5 | 4.15 | | CO-4 |
| Consider the given set of observations of the patients in the below table with symptoms and diagnosis.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Cold | Runny nose | Headache | Fever | Flu ? | | Y | N | Mild | Y | N | | Y | Y | No | N | Y | | Y | N | Strong | Y | Y | | N | Y | Mild | Y | Y | | N | N | No | N | N | | N | Y | Strong | Y | Y | | N | Y | Strong | N | N | | Y | Y | Mild | Y | Y |   Predict whether the patients has flu or not using Bayesian classification technique for the below test sample of a new patient.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Cold | Runny nose | Headache | Fever | Flu? | | N | N | Mild | Y | ? | |
| Consider the given set of observations of the patients in the below table with symptoms and diagnosis.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Cold | Runny nose | Headache | Fever | Flu ? | | Y | N | Mild | Y | N | | Y | Y | No | N | Y | | Y | N | Strong | Y | Y | | N | Y | Mild | Y | Y | | N | N | No | N | N | | N | Y | Strong | Y | Y | | N | Y | Strong | N | N | | Y | Y | Mild | Y | Y |   Predict whether the patients has flu or not using decision tree classification technique for the below test sample of a new patient.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Cold | Runny nose | Headache | Fever | Flu ? | | Y | Y | Strong | N | ? | |