2022

COMPUTER SCIENCE

M.Sc. Fourth Semester End Examination - 2022

PAPER - CS-403

Data Mining (Elective-III)

Full Marks : 50   *Time : 2 hours*

The figures in the right-hand margin indicate marks.
Candidates are required to give their answers in their own
words as far as practicable.
Illustrate the answers wherever necessary.

1. **Answer any four questions :**   **4×2=8**

a)   How can you choose 'K' in K-means algorithm?

b)   Define subject oriented in Data warehouse.

c)   State a priori principle.

d)   What is the Concept hierarchies in KDD process?

e) Define operational data.

f) What is the role activation function in neural network?

2. **Answer any four questions :** 4×4=16

a) Write down the different OLAP operations with example.

b) What do you mean by metadata in Data Warehousing? Discuss about different categories of metadata used in Data Warehouse.

c) Define clustering. Explain basic features of clustering.

d) What are the utility of schemas in OLAP? Diffentiate Star schema and Snowflake shema.

e) Suppose you want to classify a four problem using SVM. How do you implement four class problem by SVM? What are utility of the karnels in SVM?

f) You are given a data set on cancer detection. You've build a classification model and achieved an accuracy of 95%. Why shouldn't you be happy with your model performance? What can you do about it?

3. **Answer any two questions.** 8×2=16

a) Consider the following datasets which different items are occured in different transaction.

Transaction ID	T100	T200	T300	T400
Items	11,13,14	12, 13, 15	11,12,13,15	12,15

Using Appriori algorithm find the association rules of the item sets for minimum support. Count.

b) What are the steps of data mining task?

Suppose that the data mining task is to cluster the following 8 points [with (x,y) representing location] into two clusters :

X1	2	6
X2	3	4
X3	3	7
X4	4	7
X5	6	3
X6	6	4
X7	7	3
X8	7	4

c) Supposed, among 1000 people you are predicting 200 people to have the cancer. In actual only 150 people to have the cancer and among those 150 people you have diagnosed 120 people correctly. Calculate the precision, recall, accuracy, and F1-scores in terms of confusion matrix.

d) Consider the following hypothetical data concerning student characteristics and whether or not each student should be hired.

ID	Name	GPA	Effort	Attendance	Hireable
1	Sarah	poor	lots	fair	yes
2	Dana	Average	some	excellent	no
3	Alex	average	some	fair	no
4	Annie	average	lots	fair	yes
5	Emily	excellent	lots	fair	yes
6	Pete	excellent	lots	exellent	no
7	John	excellent	lots	excellent	no
8	Kathy	poor	some	fair	no

Use a Naive Bayes classifier to determine whether or not someone with excellent attendance, poor GPA, and lots or effort should be hired.