8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle-aged	medium	no	excellent	yes
13	middle-aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

5. (a) Discuss Web Mining. Explain the types of Web Mining. (CO5)

   (b) Explain in detail about text mining and its applications. (CO5)

   (c) Write short notes on the following : (CO5)

   (i) Why do spatial data Mining ?

   (ii) Multimedia data mining

H Roll No. .....

TMC–403(5)

M. C. A. (FOURTH SEMESTER) END SEMESTER EXAMINATION, May, 2022

DATA MINING AND WAREHOUSING

Time : Three Hours

Maximum Marks : 100

Note : (i) All questions are compulsory.

(ii) Answer any *two* sub-questions among (a), (b) and (c) in each main question.

(iii) Total marks in each main question are **twenty.**

(iv) Each sub-question carries 10 marks.

1. (a) What is Data Mining ? With the neat diagram, explain the steps involved in KDD. List the different challenges that motivate the development of data mining. (CO1)

P. T. O.

(b) Diagrammatically illustrate and discuss the architecture of a data warehouse.(CO1)

(c) Write short notes on the following : (CO1)

(i) Stars and Snowflake

(ii) OLTP and OLAP

(iii) Features of Data Warehouse

(iv) OLAP operations

2. (a) Suppose that the data for analysis include the attribute the frequency of stop words in documents. The values are given in increasing order : (CO2)

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.

Apply the following methods :

(i) Use smoothing by bins with a depth of 3.

(ii) Use smoothing by bin boundaries.

(iii) Use min-max normalization to transform the value 35 into the range from 0.0 to 1.0.

(iv) Use Z-score to transform the value 35 where the standard deviation of the above frequency is 12.94.

(b) What is Data Pre-processing ? Why is it required ? List and explain different steps involved in data pre-processing. (CO2)

(c) Write shot notes on the following : (CO2)

(i) Pearson's Coefficient of correlation

(ii) Chi-square test

(iii) No quality data, no quality mining results. Justify

(iv) Data Characterization and Data Discrimination

3. (a) Discuss Data Mining System Classification. Explain the integration schemes for integrating a Data Mining System with a DW System. (CO3)

(b) Write short notes on the following : (CO3)

(i) Types of Data

(ii) KDD

(c) Consider the following database has twelve transactions. Let min_support = 25% and find all frequent item set using Apriori algorithm : (CO3)

TID	List of Items
T1	I1, I2, I5
T2	I2, I4, I6
T3	I2, I3
T4	I1, I2, I4
T5	I1, I3
T6	I2, I3
T7	I1, I3
T8	I1, I2, I3, I5
T9	I1, I2, I3
T10	I1, I2, I4, I6
T11	I5, I6
T12	I3, I4, I5

Use 0.3 for the minimum support value.

4. (a) Write short notes on the following : (CO4)

(i) Entropy and Information Gain

(ii) Agglomerative and Divisive

(b) What is clustering ? How is it different than classification ? In what situations clustering can be useful ? List and explain the desired features of cluster analysis. (CO4)

(c) What is Bayes classification theorem ? Classify the tuple X = (age = youth, income = medium, student = yes, credit-rating = fair) using Naïve-Bayesian classification for the following data set. The class label attribute, buys-computer, has two distinct values {yes, no} : (CO4)

ID	Age	Income	Student	Credit-rating	Class (buys-computer)
1	youth	High	no	fair	no
2	youth	High	no	excellent	no
3	middle-aged	High	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle-aged	low	yes	excellent	yes