BE6-R4: DATA WAREHOUSING AND DATA MINING

NOTE:

1. **Answer question 1 and any FOUR from questions 2 to 7.**
2. **Parts of the same question should be answered together and in the same sequence.**

Time: 3 Hours **Total Marks: 100**

1.

a) What is data warehouse and how is it different from a database system?

b) Differentiate between Discrete attribute and continuous attributes.

c) Briefly explain any two methods for handling missing values while data cleaning.

d) Explain LIFT, a correlation measure, used to measure rule interestingness.

e) Briefly explain 2-steps procedure of classification.

f) Distinguish Clustering and Classification.

g) Explain Slice and Dice operation on a data cube.

(7x4)

2.

a) Define following terms:

i) Data characterization

ii) Data discrimination

iii) Data Cube

iv) OLAP

b) "Efficiency and scalability are two important challenges to data mining algorithm". Briefly discuss.

([3x4]+6)

3.

a) Write down suitable attribute type for the given cases and give reasoning for its selection.

i) The occupation attribute needs to be maintained which can take values *teacher, dentist, programmer, farmer* only. There is no order among values.

ii) The medical-test of a patient with two outcomes (*Positive or negative*) needs to be recorded.

iii) Daily temperature is measured (in Celsius) for the month February and can be used in quantification.

iv) The grade for students are stored where grade can be one of the four values only (*O, A, B, C*) where O is the highest grade and C is the least grade.

b) Write an algorithms for K-Nearest Neighbour Classification?

c) "It is said that snowflake schema may perform poorly than star schema if resulting dimensions are large". Explain.

([2.5x4]+4+4)

4.

a) Find all frequent item sets in following transactional database using Apriori (minimum support is 40%). Also, write down steps used in each pass.

TID	A	B	C	D	E
T_1	1	1	1	0	0
T_2	1	1	1	1	1
T_3	1	0	1	1	0
T_4	1	0	1	1	1
T_5	1	1	1	1	0

b) Define support and confidence of a rule. Also, compute support and confidence for the rule B→C.

(12+6)

5.

a) For the given six 2-dimensional data points (2,2), (2,3), (3,3), (1,2), (10,5),(10,8), find two clusters using k-means clustering algorithm assuming initial cluster centers are (2,3) and (10,5). Show cluster centers after each iteration and iterate the k-means algorithm for two times only.

b) What is Naive Bayes classifies? What is the weakness of the assumption in the method?

(12+6)

6.

a) What is Neural Network? Briefly describe multilayer-feed forward neural network.

b) What is Normalization? Given the following set of numbers, normalize using MIN-MAX Normalization:

23, 3, 67, 10, 38, 10, 45, 92, 56

c) For the given confusion matrix, compute precision and sensitivity of a classifier.

Classes	**Yes**	**No**	**Total**
Yes	90	210	300
No	140	400	540
Total	230	610	-

(6+7+5)

7. Write short notes on **any three** of the following:

a) Web usage mining
b) Genetic algorithms
c) Outlier Analysis
d) Hypo Thesis Testing
e) Graph pattern mining

(6x3)