C5-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) Differentiate between dimensionality reduction and numerosity reduction techniques for data reduction.
b) How are organizations using the information from data warehouses?
c) What is the purpose of creating data marts?
d) Explain the concept of Multidimensional Data Model with an example.
e) Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning?
f) Compare and contrast Agglomerative and Divisive Hierarchical Clustering methods.
g) What do you understand by visual data mining? Give some examples where we can use visual data mining techniques.

(7x4)

2.

a) There are several typical cube computation methods, such as MultiWay, BUC, and Star-Cubing. Describe any two of these methods and compare their feasibility and performance under the following conditions:
   i) Computing a dense full cube of low dimensionality (e.g., less than eight dimensions).
   ii) Computing an iceberg cube of around 10 dimensions with a highly skewed data distribution.
   iii) Computing a sparse iceberg cube of high dimensionality (e.g., over 100 dimensions).
b) What is a confusion matrix for classifier?

(14+4)

3.

a) What do you understand by Principal Component Portioning Algorithm? Explain the algorithm in detail.
b) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

(9+9)

4.

a) Suppose that a base cuboid has three dimensions A; B; C, with the following number of cells: |A| = 1; 000; 000, |B| = 100, and |C| = 1000. Suppose that each dimension is evenly partitioned into 10 portions for chunking.
   i) Assuming each dimension has only one level, draw the complete lattice of the cube.
   ii) If each cube cell stores one measure with 4 bytes, what is the total size of the computed cube if the cube is dense?
   iii) State the order for computing the chunks in the cube that requires the least amount of space, and compute the total amount of main memory space required for computing the 2-D planes.
b) Briefly describe the following approaches to clustering: partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods. Give examples in each case.

(4+14)

5.

a) The following contingency table summarizes supermarket transaction data, where *hot dogs* refers to the transactions containing hot dogs, $\overline{hotdogs}$ refers to the transactions that do not contain hot dogs, *hamburgers* refers to the transactions containing hamburgers, and $\overline{hamburgers}$ refers to the transactions that do not contain hamburgers.

	hot dogs	$\overline{hotdogs}$	$\sum_{row}$
hamburgers	2,000	500	2,500
$\overline{hamburgers}$	1,000	1,500	2,500
$\sum col$	3,000	2,000	5,000

i) Suppose that the association rule "*hot dogs* → *hamburgers*" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong?

ii) Based on the given data, is the purchase of *hot dogs* independent of the purchase of *hamburgers*? If not, what kind of *correlation* relationship exists between the two?

b) What are multidimensional Association Rules? Explain in brief.

(9+9)

6.

a) Write an algorithm for k-nearest neighbor classification given *k* and *n*, the number of attributes describing each tuple.

a) What is similarity search in time-series analysis? Explain its usefulness in various business functions.

(12+6)

7.

a) What is tilted time frame in stream data analysis? Explain different methods to design titled time frame with example.

b) Explain the following concepts: Data warehouse architecture

(9+9)