Second Semester M.Tech. Degree Examination, Dec.08/Jan.09
Data Warehousing and Mining

Time: 3 hrs.　　　　　　　　　　　　　　　　　　　　　　　　　Max. Marks:100

Note: *Answer any FIVE full questions.*

1　a.　What is the need for Data Warehousing?　　　　　　　　　　**(04 Marks)**
　　b.　Write six different definitions of data warehouse.　　　　　　　**(06 Marks)**
　　c.　Differentiate between operational data and informational data.　　**(04 Marks)**
　　d.　Explain 2-tier data warehouse architecture.　　　　　　　　　**(06 Marks)**

2　a.　Differentiate between the two types of metadata.　　　　　　　**(04 Marks)**
　　b.　List the 5 groups of access tools and explain any one of them.　　**(06 Marks)**
　　c.　Write the overall architecture of a data warehouse.　　　　　　**(04 Marks)**
　　d.　What are the management issues for data warehouse administration?　**(06 Marks)**

3　a.　Describe ROLAP architecture.　　　　　　　　　　　　　　**(04 Marks)**
　　b.　List the 12 OLAP guidelines.　　　　　　　　　　　　　　　**(06 Marks)**
　　c.　Describe the link between internet and data warehouse along with the web-processing
　　　　model.　　　　　　　　　　　　　　　　　　　　　　　　**(10 Marks)**

4　a.　Explain Chi-square test with an example.　　　　　　　　　　**(04 Marks)**
　　b.　Define (i) Baye's theorem (ii) Causality and (iii) Linear regression.　**(06 Marks)**
　　c.　Describe Hypothesis testing with an example.　　　　　　　　**(10 Marks)**

5　a.　What is data mining? What are the factors used for measuring data mining effectiveness?
　　　　　　　　　　　　　　　　　　　　　　　　　　　　　(04 Marks)
　　b.　With an example, describe the application score card for a decision tree.　**(06 Marks)**
　　c.　Describe the use of CART for predicting wireless communication churn.　**(10 Marks)**

6　a.　Explain the business scorecard for nearest neighbor classifier technique.　**(04 Marks)**
　　b.　A data set has 3 distinct classes C_1, C_2 & C_3. There are 50, 40 and 40 elements/patterns in
　　　　classes C_1, C_2 & C_3 respectively. Determine the entropy value of the whole data set. If the
　　　　whole data set is split based on some attribute 'X' into two subsets S_1 and S_2 with $n_1 = 70$
　　　　and $n_2 = 60$ and the splitting is as shown below, determine the entropy index value of the
　　　　data set after the segmentation.　　　　　　　　　　　　　　**(06 Marks)**

S_1

n_1	C_1	C_2	C_3
70	40	20	10

S_2

n_2	C_1	C_2	C_3
60	10	20	30

　　c.　Describe (i) Image recognition for human handwriting and (ii) K-nearest neighbor
　　　　classification technique.　　　　　　　　　　　　　　　　**(10 Marks)**

7　a.　Differentiate between the two main types of hierarchical clustering.　**(04 Marks)**
　　b.　Explain non-hierarchical clustering techniques.　　　　　　　**(06 Marks)**
　　c.　Explain the different terms used in genetic algorithms. Also explain the working of genetic
　　　　algorithms.　　　　　　　　　　　　　　　　　　　　　**(10 Marks)**

8　a.　Let the chromosome strings of parent 1 and parent 2 be 1010101 and 1000001 respectively.
　　　　If the crossover point is after 4 bits from the most significant bit, determine the chromosome
　　　　string of the son. If the crossover point is after 4 bits from the least significant bit, determine
　　　　the chromosome string of the daughter.　　　　　　　　　　**(04 Marks)**
　　b.　Discuss cost minimization for traveling salesman problem.　　　**(06 Marks)**
　　c.　Describe "optimizing predictive customer segments", using genetic algorithms.　**(10 Marks)**

* * * * *