Reg. No.										

MANIPAL ACADEMY OF HIGHER EDUCATION
MANIPAL SCHOOL OF INFORMATION SCIENCES

THIRD SEMESTER MASTER OF SCIENCE – M.Sc. (INFORMATION SCIENCE) DEGREE EXAMINATION – DECEMBER 2020

SUBJECT: MIS 603 – DATA MINING AND WAREHOUSING

Saturday, 5th December 2020

Time: 10.00 – 13.00 Hrs. Max. Marks: 100

Answer ALL the questions

All questions carry 10 marks.

1. Describe what is KDD process with a neat diagram and briefly write about various steps of a KDD PROCESS

2. Given the Initial attribute set: {A1, A2, A3, A4, A5, A6} Discuss Attribute subset selection for the following methods (3+3+4=10 Marks)
   a. Step-wise forward selection
   b. Step-wise backward elimination
   c. Decision-tree induction

3. Describe conceptual modeling of DWH

4. What tasks should be considered in the design GUIs based on a data mining query language? Write short notes on Architecture of data mining systems?

5. Calculate information Gain for each attribute (gender, major, birth_country, age_range, gpa) using the following Target and Contrasting classes.

   a. Mine general characteristics describing graduate students using analytical characterization

gender	major	birth_country	age_range	gpa	count
M	Science	Canada	20-25	Very_good	16
F	Science	Foreign	25-30	Excellent	22
M	Engineering	Foreign	25-30	Excellent	18
F	Science	Foreign	25-30	Excellent	25
M	Science	Canada	20-25	Excellent	21
F	Engineering	Canada	20-25	Excellent	18

Candidate relation for Target class: Graduate students (Σ=120)

gender	major	birth_country	age_range	gpa	count
M	Science	Foreign	<20	Very_good	18
F	Business	Canada	<20	Fair	20
M	Business	Canada	<20	Fair	22
F	Science	Canada	20-25	Fair	24
M	Engineering	Foreign	20-25	Very_good	22
F	Engineering	Canada	<20	Excellent	24

Candidate relation for Contrasting class: Undergraduate students (Σ=130)

6. Give comparison of Descriptive vs. predictive data mining? What is concept description? Briefly write on Data generalization and summarization-based characterization.

7. Give A Decision Tree for "*buys_computer*" *using the following training dataset? Calculate Gain for age, income, Student, credit_rating*

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

8. Write short notes on A) Attribute Selection Measures B) Over fitting and Tree Pruning

9. Explain a Neuron A Hidden/Output Layer Unit with a neat diagram

10. Cluster the following 8 points (with (X, Y) representing locations) into three Clusters A1(2,10), A2(2,5), A3(8,4), A4(5,8), A5(7,5), A6(6,4), A7(1,2), A8(4,9).
Initial cluster centers are: A1, A4 & A7

******