Name of the Course : **B.Sc. (Hons.) Computer Science**

Semester : **VI**

Name of the paper : **Data Mining**

Unique Paper Code : **32347611**

Year of Admission : **2015, 2016, 2017 & 2018**

Duration: 3 Hours **Maximum Marks: 75**

Instructions for Candidates

1. Attempt any FOUR out of SIX questions.
2. All questions carry *equal* marks.
3. Upload single PDF file for each question.

1. Consider a sample dataset of patients visiting a clinic for consultation:

Patient ID	Patient Name	Blood Pressure	Blood Pressure date	Chest Pain	Age	Exercise	Heart Condition
101	Sarita	120	14-10-1995	1	20	High	Good
102	Maddy	110	16-1-2018	0	40	High	Good
103	Rohit	140	18-10-2018	3	?	Low	Bad
104	Gauhar	172	4-6-2018	3	39	Medium	Bad
105	Himani	150	8-6-2018	2	35	Low	Bad
106	Shubham	110	4-7-2018	1	40	Medium	Good
107	Suresh	120	5-3-2018	0	26	High	Good
108	Anmol	110	5-5-2018	1	27	Medium	Good

- Identify the type of each attribute and give justification.
- Perform min-max normalisation on "Blood Pressure" attribute.
- Identify the data quality issues for each of the following fields. Can these issues be resolved? If yes, how?
    - Blood Pressure date for patient 101.
    - The blood pressure meter is miscalibrated and adds 1 mmHg to each reading.
    - Age of the patient 103.

2. A binary classification problem with class labels: 'Yes' and 'No' that denotes the access to the elevator, has the following set of attributes and attribute values:

   Status = {Faculty, Student}
   Floor = {First, Second, Third, Fourth, Fifth}
   Health = {healthy, unhealthy}

   Consider the following set of records for the above classification problem:

Transaction Number	**Status**	**Floor**	**Health**	**Accessible**
1	Faculty	First	Healthy	Yes
2	Student	First	Healthy	No
3	Student	Third	Healthy	No
4	Faculty	Second	Unhealthy	Yes
5	Student	Fifth	Healthy	Yes
6	Faculty	First	Healthy	No
7	Student	Fifth	Unhealthy	No

   A rule-based classifier produces the following rule set:

   R1: Status = Faculty, Floor = Second → Accessible = Yes
   R2: Status = Student, Floor = Second → Accessible = No
   R3: Floor = First → Accessible = No
   R4: Health = unhealthy → Accessible = Yes
   R5: Status = Student, Floor = Fifth, Health = healthy → Accessible = Yes

   - Are the rules in the above rule set mutually exclusive? Justify.
   - Is the rule set exhaustive? Justify.
   - Is ordering needed for this set of rules? Justify
   - Do you need a default class for the rule set? Justify
   - Compute the coverage and accuracy of rules R1 and R5. Which one do you think is a better rule? Why?

3. Consider the training examples shown in the following table for a classification problem.

Student ID	Admission Category	Admission List	Gender	Predicted Class	Actual Class
1	Sports	First	M	C0	C0
2	Arts	Second	F	C0	C1
3	Arts	Third	F	C0	C0
4	Sports	Fourth	M	C0	C1
5	Academics	First	F	C0	C0
6	Arts	Second	F	C1	C1
7	Arts	Third	M	C1	C1
8	Sports	Fourth	M	C1	C0
9	Sports	Third	F	C1	C1
10	Arts	First	F	C1	C0
11	Sports	Third	M	C1	C1
12	Sports	Second	F	C1	C1
13	Sports	First	M	C1	C0

- Compute the Information Gain for the Student ID attribute.
- Compute the Gini Index for the Admission Category Attribute and Admission List attribute
- Which is a better attribute for split based on the Gini Index: Admission Category or Admission List? Why?
- Create a confusion matrix for the above data set and compute False positive rate, accuracy, recall and precision.

4. Consider the following set of points:

{44, 28, 48, 26, 32, 14, 52, 50}

Assuming that k=2, and initial cluster centres for k-means clustering are 5 and 38, compute the sum of squared errors (SSE) and cluster assignment for each iteration.

5. Consider the dataset given below:

Age	Income	Employed	Credit-rating	Buys Car
young	high	yes	fair	yes
young	high	no	good	no
middle	high	no	fair	yes
old	medium	no	fair	yes
old	medium	no	fair	yes
old	low	yes	good	no
old	medium	no	good	no
middle	high	yes	fair	yes

Compute all class conditional and class prior probabilities. Use Naïve Bayes classifier to predict the class of the following tuple:

X = (age = young, income = medium, employed = yes, credit rating = good)

6. Consider the market basket transactions shown in the following table. Use Apriori algorithm to answer the questions that follow.

TID	Item bought
1	oregano, chocolate, milk, cheese, french fries
2	milk, french fries, cheese, ketchup
3	chocolate, cheese, oregano, ketchup
4	chocolate, cheese, french fries
5	french fries, cheese, oregano, chocolate
6	chocolate, ketchup
7	oregano, french fries, ketchup
8	oregano, french fries, chocolate
9	ketchup, oregano, milk
10	french fries, chocolate

- Assuming the minimum support threshold is fixed at 40%, list the set of frequent 1-itemsets (L_1) and with their respective supports.
- List the itemsets in the set of candidate 2-itemsets (C_2) and calculate their supports.
- Generate all association rules from the itemsets in L_2 and also compute the confidence of these rules.