UNIVERSITY OF RUHUNA
BACHELOR OF COMPUTER SCIENCE (BCS) (GENERAL) DEGREE
LEVEL III (SEMESTER I) EXAMINATION – June/July 2015

COURSE UNIT: CSC3132–Data Warehousing and Data Mining **TIME:2 Hours**

*Answer **all four (04)** questions.*

1.

a. Define the term **"Big Data"** and give two (02) examples for Big Data.

b. Briefly explain three (03) major potential application areas in data mining.

c. Explain the difference between the star schema and the fact constellation schema using a suitable example.

d. Consider the following group of data:

200, 400, 800, 1000, 2000 and 2200

i. Normalize them with min = 0 and max = 100.

ii. For the above group of data, partition them into two bins using each of the following methods:

(1). Equal-width partitioning
(2). Equal-frequency partitioning

2. Consider the transactions shown in the Table 1.

Transaction ID	Items Sold
1	Bread, Butter, Banana, Jam, Ham, Chocolate, Milk, Egg
2	Butter, Banana, Bread, Wheat Flour, Jam
3	Egg, Bread, Milk, Jam, Butter, Chocolate, Banana
4	Cheese, Bread
5	Chocolate, Egg, Jam, Banana, Butter, Bread
6	Bread, Banana, Butter
7	Cashew Nuts, Butter, Bread
8	Cake
9	Grapes, Bread, Egg, Jam, Banana, Butter

Table 1

a. Find the support and the confidence of the associations given below. Indicate whether they are interesting or not. Assume that **minimum support threshold** is 20% and **minimum confidence threshold** is 75%.

i. Butter → Banana
ii. Jam, Banana → Butter
iii. Egg, Milk, Jam → Grapes
iv. Jam, Banana →Egg
v. Egg, Bread, Butter, Jam → Chocolate

b. Find the largest frequent itemset that can be extracted using **Frequent Pattern Tree** method.

c. Briefly explain three (03) benefits of **Frequent Pattern Tree** structure.

3.

a. Briefly explain the two-step process of classification.

b. Consider the training examples shown in Table 2 for a binary classification problem.

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

Table 2

i. Draw the contingency tables and compute the Gini Index for the attributes Gender, Car Type and Shirt Size.

ii. Which attribute is best for the root node of the decision tree, Gender, Car Type, or Shirt Size? Justify your selection.

iii. Explain why Customer ID should not be used in any node of the decision tree as an attribute for the condition test even though it has the lowest Gini.

4.

a. Consider the 3-means algorithm on a set S consisting of the following six points in the plane: **a**=(0,0), **b**=(8,0), **c**=(16,0), **d**=(0,6), **e**=(8,6), **f**=(16,6). The algorithm uses the Euclidean distance metric to assign each point to its nearest centroid. A starting configuration is a subset of three starting points from S that form the initial centroids. Consider **a, e** and **f** as initial cluster centers.

i. Identify the points in the three clusters after applying the algorithm for one iteration.

ii. Calculate the new cluster centers.

b. Answer the following questions. **Do not** use more than two sentences.

i. Briefly explain how data mining is used in many different areas in manufacturing engineering.

ii. Write down the two (02) methods which can be used to evaluate the result of classification.

iii. Briefly explain how data mining can be used in customer identification phase of Customer Relationship Management (CRM).

iv. Usage of a data mining applications in medical diagnosis is one of the emerging technologies in the world. List down five (05) data mining methods which can be used for classification category of the breast cancer disease and for prediction techniques to assign patients in to either a "benign" group that is non-cancerous or a "malignant" group that is cancerous and generate rules for the same.

v. Briefly explain spatial point patterns in crime analysis.

vi. Briefly explain how Neural Networks can be used as a data mining technique for a Credit Card Fraud Detection System.

vii. "*Data mining techniques have been used to uncover hidden patterns and predict future trends and behaviors in financial markets.*" Explain the above statement using a suitable example.

****