UNIVERSITY OF RUHUNA

BACHELOR OF COMPUTER SCIENCE (BCS) (GENERAL) DEGREE

LEVEL III (SEMESTER I) EXAMINATION – SEPTEMBER 2018

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

Answer __all__ four (04) questions.

1.

a. Explain briefly the steps in *data mining* as a process of knowledge discovery.

b. Discuss the importance of using data mining approaches for a success of a business.

c.

i. What is meant by *Data Reduction* in data mining?
ii. Name and describe the data reduction strategies.
iii. Discuss three important issues that must be considered during data integration.

2.

a. Explain the key differences among an operational database, data warehouses and data mining approaches using examples.

b. Describe how data warehousing and OLAP relate to data mining.

c. Name the schemas which can be used in conceptual modeling of data warehouses. Briefly explain two of them.

d.

i. Describe the difference between ***k-means*** and ***k-medoids*** clustering algorithms.
ii. Explain briefly the problems that can be seen in the k-means clustering.
iii. Consider the following data to answer the question below.

Data: (0, 0), (0, 1), (1, 1), (1, 0), (0.5, 0.5), (5, 5), (5, 6), (6, 6), (6, 5), (5.5, 5.5)

k=2, initial clusters: C1={(0,0)} and C2={(0,1)}, and the Manhattan distance function dis = |x2-x1|+|y2-y1|. Use k-means algorithm to find the final two clusters.

Cont...

3.

a. Describe the term "Frequent Patterns" using suitable examples.

b. Consider the following transactional data (Table 1). Assume that the minimum support is 60% and minimum confidence threshold is 80%.

Table 1

Transaction ID	**List of items_bought**
T100	Mango, Onion, Nintendo, Key-chain, Eggs, Yo-yo
T200	Doll, Onion, Nintendo, Key-chain, Eggs, Yo-yo
T300	Mango, Apple, Key-chain, Eggs
T400	Mango, Umbrella, Corn, Key-chain, Yo-yo
T500	Corn, Onion, Onion, Key-chain, Ice-cream, Eggs

i. Find all the frequent itemsets using ***Apriori algorithm***.

ii. List all the strong association rules matching the following metarule (formular), where X is a variable representing customers, and *$item_i$* denotes variables representing items (ex: item A, item B):

$$\forall x \in transaction,\ buys(X, "item_1") \wedge buys(X, "item_2") \rightarrow buys(X, "item_3")$$

4.

a. Explain the reasons for popularity of decision tree classifiers.

b. Briefly explain the basic steps of decision tree classification.

c. Consider the following training data of an employee database (Table 2) and answer the questions given below.

i. Name the class label for the records in the table.

ii. Which attribute is the best as the splitting attribute for the root node of the decision tree when you use ***Information Gain*** approach as an attribute selection method? Justify your answer.

iii. Construct a decision tree for the given data.

Cont...

Table 2

ID_Code	Working_Area	Age	Income	Loan_decision
A	software	middle-age	medium	safe
B	software	young	low	risk
C	software	middle-age	low	risk
D	systems	young	medium	risk
E	systems	middle-age	high	safe
F	systems	young	medium	risk
G	systems	senior	high	safe
H	marketing	middle-age	medium	safe
I	marketing	middle-age	medium	risk
J	hardware	senior	medium	safe
K	hardware	young	low	risk

****