UNIVERSITY OF RUHUNA

BACHELOR OF COMPUTER SCIENCE (BCS) (GENERAL) DEGREE

LEVEL III (SEMESTER I) EXAMINATION - JULY 2016

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

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

1.

a. Explain the purpose of Data Mining.

b.

i. Describe the term "Big Data".
ii. List down the characteristics that describe Big Data.
iii. State four sources from which the Big Data can be generated.

c. Compare and contrast the followings.
i. a data warehouse and a database
ii. classification and prediction techniques

d. Briefly explain the steps in the process of knowledge discovery (KDD – Knowledge Discovery in Databases).

e. Discuss two challenges in mining a huge amount of data with compare to mining a small amount of data.

2.

a. Discuss the importance of Data Preprocessing.

b. Describe the difference between a data warehouse and a data mart?

c. A data warehouse can be modeled by using a *star schema* or a *snowflake schema*.

i. Briefly describe the similarities and the differences of above schemas.
ii. Analyze their advantages and disadvantages with regard to one another.
iii. Which schema might be more empirically useful? Justify your answer.

d. Consider the following a group of 12 sales price records:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.

Partition them into three (3) bins by each of the following methods:

i. Equal-frequency partitioning.
ii. Equal-width partitioning.

3.

a. Briefly explain the purpose of using Association Rules.

b. Define the **Apriori Property** in the Apriori algorithm.

c. Consider the following transactional data of TDB Company to find the interesting patterns. A database has six transactions. Assume that the minimum support threshold is 33.33% and minimum confidence threshold is 60%. Find all the frequent itemsets using Apriori algorithm and derive all strong association rules from the final frequent itemsets.

Transaction ID	Items
T1	HotDogs, Buns, Ketchup
T2	HotDogs, Buns
T3	HotDogs, Coke, Chips
T4	Chips, Coke
T5	Chips, Ketchup
T6	HotDogs, Coke, Chips

4.

a. Explain the reasons for the popularity of decision trees classifiers in data mining.

b. Consider the following training data from the "AllElectronics" customer database.

RID	Age	Income	Student	Credit_rating	Buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

i. What is the class label for the records in the above table?

ii. Which attribute is 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.

c. Explain the strengths and weaknesses of *k-means* clustering algorithm.

d. Describe *Agglomerative* and *Divisive* hierarchical clustering methods.

****