DEDAN KIMATHI UNIVERSITY OF TECHNOLOGY

UNIVERSITY EXAMINATION ACADEMIC YEAR 2021/2022

THIRD YEAR FIRST & SECOND SEMESTER EXAMINATION FOR THE DEGREE OF BACHELOR OF SCIENCE (CRIMINOLOGY AND SECURITY MANAGEMENT)

CSM 4206: DATA MINING

TIME: 2 HOURS

Instructions: Answer Question 1 and Any Other Two.

Question One (30 Marks)

a) Define the following terms
   i. Data Mart (1 mark)
   ii. Data Mining (1 mark)
   iii. Machine Learning (1 mark)

b) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. (3 marks)

c) Outliers are often discarded as noise. However, one person's garbage could be another's treasure. Exceptions in credit card transactions can help us detect fraudulent use of credit cards. Taking fraudulence detection as an example, propose and explain two methods that can be used to detect outliers and discuss which one is more reliable. (4 marks)

d) Association rule mining often generates a large number of rules. Discuss effective methods that can be used to reduce the number of rules generated while still preserving most of the interesting rules. (3 marks)

e) Data mining can be defined as "Data Mining is a confluence of multiple disciplines". Justify the above statement (2 marks)

f) Explain the capabilities of online analytical processing that can be used to browse a hyper cube in a multidimensional ware house. (3 marks)

g) John is a database administrator at Kenya police headquarters administering heterogeneous databases and information repositories. The company is in the process of implementing a data warehouse. Discuss some of the issues John is likely to face during data integration. (3 marks)

h) "We are drowning in data, but starving for knowledge'', clearly explain possible solutions. (2 marks)

i) Data mining has matured over time with the development of reliable and scalable tools that outperform older classical statistical methods. Despite this rapid evolution, many issues are still pending yet to be to be addressed. Explain any three such issues (3 marks)

j) Suppose Naivas Supermarket has an operational database management system and a data warehouse at the same time, using the data about the customers who purchased sugar, bread and milk last month, explain the difference OLTP and OLAP (4 marks)

QUESTION TWO (20 Marks)

a) Distinguish between association analysis and outlier analysis as used in data mining (4 marks)

b) Sinai hospital is implementing a data warehouse based on patient and doctor information. It is important to know the number of patients who were attended to at any one time and the charge (fee) per patient so as monitor profits. For each patient and doctor we need to note their name, address, contacts, gender and description of illness. Doctor appointments can be monthly, quarterly or yearly and date and day should be specified

i. Discuss the three classes of schemas that are popularly used for modeling data warehouses. (3 marks)

ii. Draw a schema diagram for the above data warehouse. (10 marks)

iii. Starting with the base cuboid [day; doctor; patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2014? (3 marks)

Question THREE (20 Marks)

a) i) Distinguish between data warehouse and data warehousing (2 marks)

ii) Outline FOUR characteristics exhibited by a data warehouse that supports the management's decision- making process. (4 marks)

b) Explain ONE reason why clustering is required in data mining (1 marks)

c) The Kenya Anti-Terrorism commission collects data about the potential attacks that may happen in the country. The commission is looking for an employee who can be able to assist in the mining of Knowledge from the large amount of data they have collected. As a candidate for the position, explain the knowledge discovery steps that they can follow to acquire useful information. (7 marks)

d) With the aid of a well labeled diagram, explain the data warehouse three-tier architecture (6 marks)

QUESTION FOUR (20 Marks)

a) Consider the table below extracted from Mathai supermarket cashiers. The supermarket wants to mine the information so as to put the closely related items or those with strong relationships together.

Transaction ID	Items
1001	{Jacket, Shoes}
1002	{Milk, Cheese, Bread, Shoes}
1003	{Jacket, Bread}
1004	{Milk, Bread, Shoes}
1005	{Bread, Shoes}
1006	{Jacket, Cheese}

i. What are the supports and confidences of the following two rules? (4 marks)

ii. Using association rule mining and minimum support of 30%, how many large itemsets will be found? Show all your workings (10 marks)

b) Suppose a group of 12 sales price records has been sorted as follows: 5*;* 10*;* 11*;* 13*;* 15*;* 35*;* 50*;* 55*;* 72*;* 92*;* 204*;* 215*:* Partition them into three bins by each of the following methods.

i. Equal-frequency partitioning (2 marks)

ii. Smoothing by bin boundaries (2 marks)

iii. Clustering (2 marks)