CHUKA  **UNIVERSITY**

UNIVERSITY EXAMINATIONS

EXAMINATION FOR THE AWARD OF DEGREE OF BACHELOR OF SCIENCE IN APPLIED COMPUTER SCIENCE AND BACHELOR OF SCIENCE IN COMPUTER SCIENCE

COSC 447: DATA WAREHOUSE AND DATA MINING

ACSC 441: DATA MINING AND KNOWLEDGE DISCOVERY

STREAMS: **TIME: 2 HOURS**

DAY/DATE: THURSDAY 13/04/2023 **2.30 P.M. –4.30 P.M.**

Instructions

1. Answer Question 1 and Any Other Two questions.
2. **Mobile phones** are **not allowed** in the examination room.
3. **No Reference Material** is allowed in the exam Room.
4. You are not allowed to write on this examination question paper.
5. Write legibly on both sides of an answer sheet
6. Begin each question on a new page.
7. Write legibly on both sides of an answer sheet

SECTION A:

Answer all the questions in this section.

QUESTION 1 (30 MARKS)

a) Define following concepts: (3 Marks)

   i. Pivoting

   ii. Data tokenization

   iii. Data preprocessing

b) Differentiate between OLTP and OLAP, explain how they have enhanced the business processing. (4 Marks)

c) Differentiate between data sanitization and data scrubbing as applied in data mining. (4 Marks)

d) Explain four contemporary applications of data mining. (4 Marks)

e) Explain why is the KNN Algorithm is known as Lazy Learner. (3 Marks)

f) Describe the process of data visualization, and explain how it can be used to communicate the results of a data mining project to a non-technical audience. (4 Marks)

g) Discuss the main challenges of data mining. (4 Marks)

h) Discuss the main differences between supervised and unsupervised learning, and provide an example of a problem that would be better suited to one approach over the other. (4 Marks)

SECTION B:

Answer any two questions in this section.

QUESTION 2(20 MARKS)

a) Briefly discuss the major difference between classification and clustering. List one real application for each of them. (4 Marks)

b) Explain the difference between K-Means Clustering algorithm and K-Nearest Neighbour data mining techniques. (5 Marks)

c) Table below shows Iris dataset. The dataset have 3 attributes which have sepal length, sepal width, and species. Species is the target attribute with three species (Setosa, Virginia, and Versicolor).

Sepal Length	Sepal Width	Species
5.3	3.7	Setosa
5.1	3.8	Setosa
7.2	3.0	Virginica
5.4	3.4	Setosa
5.1	3.3	Setosa
5.4	3.9	Setosa
7.4	2.8	Virginica
6.1	2.8	Verscicolor
7.3	2.9	Virginica
6.0	2.7	Verscicolor
5.8	2.8	Virginica
6.3	2.3	Verscicolor
5.1	2.5	Verscicolor
6.3	2.5	Verscicolor
5.5	2.4	Verscicolor

i. Use the K-NN algorithm to devise a set of rules for identifying unlabelled species. Use Manhattan distance. (8 Marks)

Sepal Length	Sepal Width	Species
5.2	3.1	?

ii. What will be the appropriate *K* value for the dataset? Support your answer. (3 marks)

QUESTION 3(20 MARKS)

a) Explain the concept of association rule mining and describe how it can be used to identify relationships between items in a large dataset. (4 Marks)

b) A database has 4 transactions, shown below.

TID	Date	items_bought
T100	10/15/23	{K, A, D, B}
T200	10/15/23	{D, A, C, E, B}
T300	10/19/23	{C, A, B, E}
T400	10/22/23	{B, A, D}

Assuming a minimum level of support min_sup = 60% and a minimum level of confidence min_conf = 80%:

i. Find all frequent itemsets using the Apriori algorithm. Show your work—just showing the final answer is not acceptable. For each iteration show the candidate and acceptable frequent itemsets. (8 Marks)

ii. List all of the strong association rules, along with their support and confidence values. (8 Marks)

QUESTION 4 (20 MARKS)

a) With aid of illustration in each case compare decision tree and neural network classification methods. (6 marks)

b) You have a friend who only does one of four things on every Saturday afternoon: go shopping, watch a movie, play tennis, or just stay in. You have observed your friend's behavior over 11 different weekends. On each of these weekends you have noted the weather (sunny, windy, or rainy), whether her parents visit (visit or no-visit), whether she has drawn cash from an ATM machine (rich or poor), and whether she had an exam during the coming week (exam or no-exam). You have built the following data table:

# ex.	Weather	Parents	Cash	Exam	Decision
1	sunny	visit	rich	yes	cinema
2	sunny	no-visit	rich	no	tennis
3	windy	visit	rich	no	cinema
4	rainy	visit	poor	yes	cinema
5	rainy	no-visit	rich	no	stay-in
6	rainy	visit	poor	no	cinema
7	windy	no-visit	poor	yes	cinema
8	windy	no-visit	rich	yes	shopping
9	windy	visit	rich	no	cinema
10	sunny	no-visit	rich	no	tennis
11	sunny	no-visit	poor	yes	tennis

i. What is the entropy of H(Decision)? (4 Marks)

ii. Draw the full decision tree that would be learned for this dataset using ID3 algorithm. You need to show any calculations. (10 Marks)

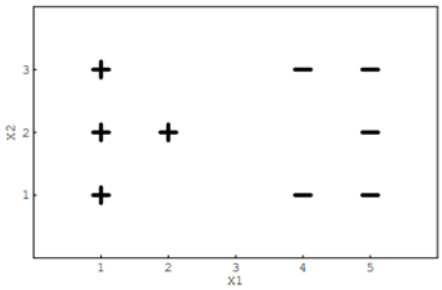
QUESTION 5 (20 MARKS)

a) You are designing a deep learning system to diagnose chest cancer through X-ray images. Explain three most appropriate evaluation metric and why. (3 Marks)

b) Explain the term "Curse of dimensionality". How can you deal with it? (3 Marks)

c) Explain the concept of deep learning and discuss how it is different from traditional machine learning algorithms and how it can be used in data mining. (4 Marks)

d) Suppose you are using a linear Support Vector Machine and is given the following dataset with two classes (+,-). (4 Marks)



e)

Draw the decision boundary of linear SVM. Give a brief explanation.

f) Draw and explain the architecture of Convolutional Neural Network (CNN). (6 Marks)
