

CHUKA



UNIVERSITY

UNIVERSITY EXAMINATIONS

**EXAMINATION FOR THE AWARD OF DEGREE OF
BACHELOR OF SCIENCE IN APPLIED COMPUTER SCIENCE, BACHELOR OF
COMPUTER SCIENCE**

**COSC 447/ ACSC 441: DATA MINING AND DATA WAREHOUSE/ DATA MINING
AND KNOWLEDGE DISCOVERY**

STREAMS: BSC (COSC, ACSC) Y4S2

TIME: 2 HOURS

DAY/DATE: THURSDAY 09/04/2020

8.30 AM – 1.30 PM

INSTRUCTIONS:

Answer **question 1** and **any TWO**.

Question 1(30 marks)

- a) Explain the following technologies as used in data mining: (6 marks)
- i. Classification
 - ii. Decision tree
 - iii. Prediction
 - iv. Data cleaning
 - v. Clustering
 - vi. Machine learning
- b) Differentiate between descriptive from predictive mining tasks. (2 Marks)
- c) Explain the roles played by data mining in business environment. (4 marks)
- d) Determine the advantages of the decision tree approach over other approaches for performing classification. (4 Marks)
- e) New applications of data mining are text mining and web mining. Enumerate six examples of text mining. (6 marks)
- f) With the aid of a well labelled diagram explain how Data Mining is a step to Knowledge Discovery. (8 Marks)

Question 2 (20 marks)

- a) With an aid of an example explain what a classifier is as used in data mining. (2 marks)
- b) Discuss the difference between Data Mining and Machine learning? (4 marks)
- c) Identify and explain FOUR defining characteristics of a data warehouse. (6 Marks)
- d) ANN is a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs. Discuss the basic structure and working of ANN. (8 marks)

Question 3 (20 marks)

- a) Differentiate between Feedforward and Feedback ANN. (2 marks)
- b) Explain FOUR applications of association analysis. (4 Marks)
- c) Discuss any TWO types of clusters commonly used in data mining. (4 marks)
- d) Discuss THREE types of deep learning and their appropriate application areas. 6 marks)
- e) Differentiate Clustering and Regression when employed to smooth data. Use Illustrations where necessary. (4 Marks)

Question 4 (20 marks)

- a) Hypothesize the impact of data mining in the area of medical science. (4 marks)
- b) Explain the concept of Data Preprocessing and why we need to Preprocess data. (4 Marks)
- c) Explain what is the function of supervised learning and unsupervised learning? (4 marks)
- d) Explain the term ‘overfitting’ as studied in machine learning? Why does overfitting occur? How can overfitting be controlled. (4 Marks)
- e) Differentiate between K-means clustering algorithm and K-Nearest Neighbor. (4 marks)

Question 5(20 marks)

You have been hired by agricultural research institute to help them create an AI based system for Mushroom classification as either poisonous or not. You have the following data to consider.

Example	Is_Heavy	Is_Smelly	Is_Spotted	Is_Smooth	Is_Poisonous
A	NO	NO	NO	NO	NO
B	NO	NO	YES	NO	NO
C	YES	YES	NO	YES	NO
D	YES	NO	NO	YES	YES
E	NO	YES	YES	NO	YES
F	NO	NO	YES	YES	YES
G	NO	NO	NO	YES	YES
H	YES	YES	NO	NO	YES
U	YES	YES	YES	YES	?
V	NO	YES	NO	YES	?
W	YES	YES	NO	NO	?

You know whether or not mushrooms A through H are poisonous, but you do not know about U through W.

- a) What is the entropy of Is_Poisonous? (2 Marks)
- b) Which attribute should you choose as the root of a decision tree? (3 Marks)
- c) What is the information gain of the attribute you chose in the previous question? (2 Marks)
- d) Build a decision tree to classify mushrooms as poisonous or not. (10 Marks)
- e) Classify mushrooms U, V, and W using this decision tree as Poisonous or not. (3 Marks)
