CHUKA  **UNIVERSITY**

UNIVERSITY EXAMINATIONS

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

COSC 447: DATA WAREHOUSING AND DATA MINING

ACSC 411: DATA MINING AND KNOWLEDGE DISCOVERY

STREAMS: **TIME: 2 HOURS**

DAY/DATE: FRIDAY 24/09/2021 **8.30 A.M – 10.30 A.M**

INSTRUCTIONS

Answer Question ONE and ANY other TWO Questions

QUESTION ONE (30 MARKS)

a) Differentiate between knowledge discovery and data mining. (2 marks)
b) Explain how data mining can be used in Telecommunication Industry and Biological Data Analysis. (4 marks)
c) Differentiate between star schema and snowflake schema as applied in data warehousing. (3 marks)
d) Explain Association algorithm in Data mining. (3 marks)
e) The criteria used for the ETL tools comparison are grouped into categories. Briefly describe any six of these categories. (6 marks)
f) Discuss with an aid an of an illustration how you will design an e-mail spam filter? (6 marks)
g) Suppose that you are in the market to purchase a data mining system. Which features would you look for when selecting the system? (6 marks)

QUESTION TWO (20 MARKS)

a) Differentiate between supervised and unsupervised data mining techniques. (4 marks)
b) The table below show a collection of records used by a certain bank to determine if a customer will default on loan payment or not. Use the K-NN algorithm to devise a set of rules for identifying whether a customer will default or not. Use squared Euclidean distance. (10 Marks)

Age	Loan	Default
48	142,000	?

Age	Loan	Default
25	40,000	N
35	60,000	N
45	80,000	N
20	20,000	N
35	120,000	N
52	18,000	N
23	95,000	Y
40	62,000	Y
60	100,000	Y
48	220,000	Y
33	150,000	Y

c) During Machine learning class you were tasked to build a random forest model with 6000 trees. Through its training, your model achieved a training error of 0.00. However, on testing the validation error was 54.23. Explain the phenomena and the solution to this phenomenon. (6 Marks)

QUESTION THREE(20 MARKS)

a) Describe the steps involved in data mining when viewed as a process of knowledge discovery. (8 marks)

b) In class we discussed three different kinds of Unsupervised Learning problems. List the three types of problems and for each name a method which addresses that problem.(6 Marks)

c) The following is a table that shows the items bought in different transactions. Calculate (as a percentage) the support and confidence of: (6 Marks)

i) Steak → Beer

ii) Beer → Steak

Transaction-id	Items bought
10	Steak, Sugar, Beer
20	Steak, Milk, Beer
30	Steak, Beer, Coffee
40	Sugar, Coffee, Bread
50	Sugar, Milk, Beer, Coffee, Bread

QUESTION FOUR (20 MARKS)

a) Suppose a certain bank wants to deploy a new system for assessing credit worthiness of its customers. The new system uses a feed forward network with a supervised learning algorithm. Suggest what the bank should have before the system is used. Discuss problems associated with this requirement. (6 marks)

b) Describes the four major issues data mining. (4 marks)

c) Use the ID3 algorithm to devise a set of rules for identifying bird's eggs (see table below). (10 marks)

Bird	Colour	Texture	Size
Herring Gull	Brown	Speckled	Medium
Starling	Blue	Smooth	Small
Mallard	Grey	Smooth	Medium
Swan	Grey	Smooth	Large
Guillemot	Brown	Speckled	Medium
Sparrow	Grey	Speckled	Small

QUESTION FIVE (20 MARKS)

a) Discuss five applications of clustering in modern business. (5 Marks)

b) With the aid of diagram illustrate the different layers on Convolution Neural Network(CNN). (6 Marks)

c) Explain what is Pooling on CNN, and How discuss how it works in deep learning. (4 Marks)

d) You built a machine learning model, and while training it, you noticed that after a certain number of epochs, the accuracy is decreasing. What's the problem and how to fix it? (5 Marks)
