University of Eswatini

Department of Computer Science

Examination(Main)

2020/2021

FIRST SEMESTER

Title of Paper: DATA MINING

Course Code: CSC471

Time Allowed: Three (3) Hours

Instructions: Answer Question one and any other three **Questions**.

Don't write anything on the Examination Question paper.

You are not allowed **to open this paper until you have been told to do so by the invigilator.**

QUESTION ONE

a) Define Data Mining 2MARKS

b) List and explain the steps of the Knowledge Discovery in Databases (KDD). 15MARKS

c) The following is an example of customer purchase transaction data set.

CID	TID	Date	Items Purchased
1	1	01/01/2001	10,20
1	2	01/02/2001	10,30,50,70
1	3	01/03/2001	10,20,30,40
2	4	01/03/2001	20,30
2	5	01/04/2001	20,40,70
3	6	01/04/2001	10,30,60,70
3	7	01/05/2001	10,50,70
4	8	01/05/2001	10,20,30
4	9	01/06/2001	20,40,60
5	10	01/11/2001	10,20,30,60

Note: CID = Customer ID and TID = Transactions ID

i. Calculate the support and confidence of the following association rule. Indicate if the items in the association rule are independent of each other or have negative or positive impacts on each other. {10} -> {50,70} 4MARKS

ii. Based on the types of association rules discussed in class, identify which type(s) of rules {10}-> {50,70} is? 4MARK

QUESTION TWO

a) What is the main difference between predictive models and descriptive models? Give an example of each type. 4MARKS

b) List and explain four reasons why multi-dimensional data mining is important 8MARKS

c) Write short note on the following data mining tasks

Classification 2MARKS

Regression 2MARKS

Clustering 2MARKS

Association Rules 2MARKS

d) Suppose you are using a neural network instead of a decision tree. List at least three possible parameters you want to tune to improve its performance during the training period. 5MARKS

QUESTION THREE

a) What is the difference between discrete and continuous attributes? 10MARKS

b) Explain any 5 sequential methods for handling missing values that you can apply on the dataset in Table 1.

Table 1

	Attributes			Decision
case	Temperature	Headache	Nausea	Flu
1	High	?	No	Yes
2	Very High	Yes	Yes	Yes
3	?	No	No	No
4	High	Yes	Yes	Yes
5	High	?	Yes	No
6	Normal	Yes	No	No
7	Normal	No	Yes	No
8	?	Yes	?	Yes

10MARKS

c) Define an attribute and write short note on four common attributes used in Data Mining with the support of an example for each 5MARKS

QUESTION FOUR

a) How can noise be reduced in a dataset? 4MARKS

b) Describe two different approaches to detect outliers in a dataset. 4MARKS

c) Explain three data imputation methods used in Python and when do you think each of these methods would be most appropriate. Write a python code to implement one of the three methods 12MARKS

d) Write out the common properties of similarities between data points in Data Mining 5MARKS

QUESTION FIVE

Assuming we have data containing university CGPA scores of students (ranging from 0 to 6) and their future allowances (in thousands Emalangeni) as presented in Table 2. We noticed that the features of the CGPA and the allowances have different scales, the possibility that the feature with higher value is favour is high. This will impact the performance of the data mining algorithm and obviously, we do not want our algorithm to be biased towards one feature.

Table 2: Student record

s/no	Student ID	CGPA	ALLOWANCES(E)
0	1	4.2	15000
1	2	5.0	24000
2	3	3.4	25000
3	4	2.5	16000
4	5	4.2	12000

a) Why do you think you need to apply normalization method on CGPA and allowances? Apply min-max, Z-score and decimal scaling method on Table 2. 10MARKS

b) Determine the effect of feature scaling when we compare the Euclidean distance between data points for students A and B, and between B and C on Table 2.

15MARKS

QUESTION SIX

a) Write short note on the following

i) Similarity and dissimilarity based on numerical measure and attribute type 5MARKS

ii) Minkowski distance 5MARKS

b) Calculate minkowski distance when λ=1 and λ=2 between object A and B in the matrix K below

$$K=\begin{pmatrix} 2 & 3 \\ 10 & 7 \\ 3 & 2 \end{pmatrix}$$

15MARKS