MERU UNIVERSITY OF SCIENCE AND TECHNOLOGY
P.O. Box 972-60200 – Meru-Kenya
Tel: +254(0) 799 529 958, +254(0) 799 529 959, + 254 (0) 712 524 293,
Website: info@must.ac.ke Email: info@must.ac.ke

University Examinations 2023/2024

FIRST YEAR SECOND SEMESTER EXAMINATION FOR THE DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE

CCD 7152: ADVANCED DATA MINING

DATE: APRIL 2024 **TIME: 3 HOURS**

INSTRUCTIONS: *Answer question **one** and any other **two** questions*

QUESTION ONE (30 MARKS)

a) Explain the importance of data preprocessing as a step in the Knowledge Discovery cycle (4 Marks)

b) Correlation analysis is an approach that can be used is detection of data redundancy.

i. Distinguish between the Pearson Product Moment coefficient and the x^2 (chi-square) test as measures of correlation relationship (4 Marks)

ii. Suppose that a group of 7,500 students was surveyed and the gender of each person noted. In the poll, each person responded as to whether their preferred type of movie was action, romance, drama or documentary. Thus, we have two attributes, gender and preferred watch. The observed frequency (or count) of each possible joint event is summarized in the Table 2.1 below;

	Male	Female	**Total**
Action	1776 ($e_{1,1}$)	467 ($e_{l,2}$)	**2243**
Romance	145 ($e_{2,1}$)	1205 ($e_{2,2}$)	**1350**
Drama	507 ($e_{3, 1}$)	600 ($e_{3,2}$)	**1107**
Documentary	1500($e_{4,1}$)	1300($e_{4,2}$)	**2800**
Total	**3928**	**3572**	**7500**

Table 2.1: Preferred Watch Frequency Table

Required:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \qquad (2.1)$$

Where oij is the observed frequency (actual count) of the joint event (A_i, B_j) and ei,j is the expected frequency of(A_i, B_j), which can be computed as

$$eij = \frac{count(A = a_i)xcount(B = bj)}{N} \qquad (2.2)$$

Where *N* is the number of data tuples, count (A=a_i) is the number of tuples having value a_i for A, and count (B=b_j) is the number of tuples having value b_j for B

i. Using eqn. 2.2 fill in the expected frequency count in the table ($e_{1,l}$ — $e_{4,2}$) (4 Marks)

ii. Compute the x^2 for the set of data in Table 2.1 (3 Marks)

iii. Calculate the degrees of freedom for this data set and the accepted significance level (2 Marks)

iv. Based on your computation, what is the interpretation on the correlation of these variables (3 Marks)

QUESTION TWO (20 MARKS)

a) Real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogeneous sources. Consequently, low-quality data will lead to low-quality mining results.

i. Explain what you understand by the term 'noise ' in reference to data (2 Marks)

ii. Consider the following sorted set of data;

44, 58, 65, 72, 81, 84, 95, 98, 114, 116, 119, 153, 165, 189

a) Explain the three binning approaches that may be used for data smoothin (3 marks)

b) Demonstrate how any two of the binning approaches explained in a. above can be practically applied to this data set (6 Marks)

iii. Explain how a data miner can use 'the most probable value ' to fill in the missing value and its advantage over the other potential approaches (2 marks)

b) Using appropriate examples, distinguish between each of the following concepts as used in data mining

i. Supervised and hybrid learning (2 marks)

ii. Classification and clustering (2 marks)

c) Explain how each of the OLAP operations function in a data warehouse

i. Roll up (1 mark)

ii. Slice and dice (1 mark)

iii. Pivot (1 mark)

QUESTION THREE (20 MARKS)

a) Using an appropriate illustration, explain the various stages of Knowledge Discovery (KDD) (8 Marks)

b) Consider a company that pays its employees a salary ranging from kshs. 15,000 to kshs. 150,000. Using this data, you are required to transform the salaries for analysis and fit a salary of kshs. 67,500 using each of the following approaches;

i. Min-max normalization to [0.0, 1.0] (4 Marks)

ii. Z-score normalization (Let μ =66000, σ =10000) (4 Marks)

iii. Normalization by decimal scaling (4 Marks)

QUESTION FOUR (20 MARKS)

a) Enumerate five differences between the OLTP and OLAP technologies (5 Marks)

b) Icon based visualization techniques have gained a significant acceptance across various data mining applications. Briefly, outline the building blocks of this type of approach and describe the how Chernoff faces fit into this category (6 Marks)

c) Meru University wishes to establish a data warehouse for its enormous data being generated from various campuses and research centers. As an expert in data warehousing and mining, you are hired to perform a viability study of the intended project. Provide a summary of your report detailing the potential models that the University can adopt and justify why it may be necessary for the full implementation of the project (9 marks)

QUESTION FIVE (20 MARKS)

a)

i. Explain the motivations behind the establishment of a data warehouse in any organization (4 Marks)

ii. Why should an organization consider keeping a separate transactional database alongside a data warehouse (4 Marks)

b) Meru University keeps track of all its expenditure across its various campuses. For each location, it tracks its expenditure on the items teaching (*teach*), computers (*comp*), library (lib) and staff (*staff)* as shown in Table 3.1 *(Kshs. in thousands).*

Time	Location = "Main"				Location = "Nairobi"				Location = "Kapsabet"			
	Teach.	Comp.	Lib.	Staff	Teach.	Comp.	Lib.	Staff	Teach.	Comp.	Lib.	Staff
Q1	2854	1882	189	1623	1087	968	38	872	818	446	43	591
Q2	2943	1890	164	1698	1130	1024	41	925	894	469	52	682
Q3	4032	1924	159	1789	1034	1048	65	1002	940	495	58	728
Q4	1129	1992	163	1870	1142	1091	74	984	978	564	59	784

Table 3.1: MUST Expenditure Tracking

Using this data, explain using appropriate illustrations (*where applicable*) each of the following Data Warehousing Concepts;

a. Multidimensional Data Cube (4 Marks)

b.OLAP (4 Marks)

c. Star Schema of a multidimensional database (4 Marks)