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

SECOND YEAR SECOND SEMESTER EXAMINATION FOR THE DEGREE OF BACHELOR OF SCIENCE DATA SCIENCE

CDS 3253: ALGORITHMS FOR DATA SCIENCE

DATE: APRIL 2024 **TIME: 2 HOURS**

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

QUESTION ONE (30 MARKS)

a) Distinguish between each of the following terms as used in Data Science;

i. Missing Completely at Random (MCAR) and Missing at Random (MAR) (2 Marks)

ii. Univariate and Multivariate Feature Imputation (2 Marks)

iii. Euclidean and Manhattan Distance (2 Marks)

iv. Data Science and Data Mining (2 Marks)

b) Consider the following dollar prices against the Kenya Shilling for some twelve instances;

146.5, 160, 156.5, 153.15, 150.45, 148.1, 145.4, 142.5, 140.5, 138.5, 135.9, 132.4

i. Smooth the data using Bin means (2 Marks)

ii. Smooth the data using Bin Boundary (4 Marks)

iii. Smooth the data using linear regression (4 Marks)

a) Data preprocessing is a data mining technique commonly used in all projects and plays an important role in data science. However, a certain school of thought argues that this may be an exercise in futility. As a data scientist, you would a different opinion.

i. Define the term Data Preprocessing (1 Mark)

ii. Briefly, outline some of the reasons you would put forward on this argument as to why preprocessing is important (5 Marks)

iii. Describe the three important data pre-processing tasks clearly highlighting their associated sub tasks (6 marks)

QUESTION TWO (20 MARKS)

Consider the table below to answer the questions that follow;

Node ID	**GENDER**	**AGE**	**WEIGHT**	**EYE COLOR**	**HEIGHT**
A	F	51	77	BLACK	5
B	F	45	47	BLACK	5.11
C	M	34	55	WHITE	?
D	M	23	59	PURPLE	5.6
E	F	69	72	BLUE	?
F	M	44	60	BLUE	?
G	M	33	?	WHITE	?

H	F	46	60	BLUE	5
1	M	34	40	BLUE	5.1
J	F	56	62	?	5
K	F	?	45	BLACK	5.5
L	M	45	?	WHITE	5.7

i. Rewrite the resultant table when replacing the missing values using;

i. Drop Method (1 Marks)

ii. Mean (2 Marks)

iii. Mode (2 Marks)

ii. Using K-Nearest neighbors, compute the missing values for C, G and J when K=3, 14=4 and 14=5 (5 Marks each) (15 Marks)

QUESTION THREE (20 MARKS)

a) Data normalization plays a crucial role in Data Science. Assuming that the you collect data of salaries of MUST employees for analysis, you realize that the income attribute ranges between Kes. 24,000 to Kes. 1,400,000.

i. Using the Min-Max Normalization technique map an income of Kes. 260,000 to arange of [0.0, 1.0] (4 Marks)

ii. What would the income in (i) above be mapped to when using the Z-score normalization technique (4 Marks)

iii. Map the income in (i) using the Decimal Scaling technique (4 Marks)

b) Data may contain hundreds of attributes which often may be irrelevant to the task at hand. Consequently, feature selection is applied to aid not only in speeding up training time but also improve on the accuracy of the developed models.

i. Outline the procedure of feature selection in data science (4 Marks)

ii. Highlight any four wrapper algorithms used in feature selection (4 Marks)

QUESTION FOUR (20 MARKS)

a) Using appropriate examples, explain each of the following data reduction methods;

i. Data Cube Aggregation (4 Marks)

ii. Dimension reduction (4 Marks)

iii. Numerosity Reduction (4 Marks)

b) Describe each of the following terms as used in data analytics;

i. Descriptive Analytics (2 Marks)

ii. Diagnostic Analytics (2 Marks)

iii. Predictive Analytics (2 Marks)

iv. Prescriptive Analytics (2 Marks)

QUESTION FIVE (20 MARKS)

a) You have been hired as the lead data scientist to explore the massive image datasets in the NASA warehouse. You are required to present a detailed report with adequate visualizations for the Organization's decision making.

i. Differentiate between the terms data and data-set (2 Marks)

ii. Explain what you understand by the term Data Warehouse (2 Marks)

iii. Outline how you would create a Data Mart from the warehouse (2 Marks)

iv. Highlight all the steps you would follow to complete this task (2 Marks)

v. Supposedly one of the tasks you need to accomplish contains an unbalanced data. Elaborate three measures you would put in place to ensure appropriate data augmentation (3 Marks)

vi. List any two tools you use to render your visualizations (2 Marks)

vii. Describe any three data transformation effects that you may apply during this task

(3 Marks)