THE CO-OPERATIVE UNIVERSITY OF KENYA

END OF SEMESTER EXAMINATION APRIL/MAY -2025

EXAMINATION FOR THE DEGREE OF BACHELOR OF SCIENCE IN DATA SCIENCE, APPLIED STATISTICS AND DATA SCIENCE
(YR II SEM II)

UNIT CODE: BDSC 2203
UNIT TITLE: DATA PREPARATION

DATE: TUESDAY 29TH MAY, 2025
TIME: 8:30 AM – 10:30 AM

INSTRUCTIONS:

- **Answer Question ONE (Compulsory) and choose any other TWO Questions**

QUESTION ONE (30 MARKS)

a)

i) State the data preparation term that suits each of the following descriptions

I. The process of extracting and organizing the important features from raw data in such a way that it fits the purpose of the machine learning model. (1 mark)

II. the process of bringing together data from multiple sources across an organization to provide a complete, accurate, and up-to-date dataset (1 mark)

III. the process of taking raw data that has been extracted from data sources and turning it into usable datasets (1 mark)

ii) State the type of schema integration conflict that suits each of the following descriptions

I. One attribute **employeeNo** is declared as an integer in one schema and a character string in another schema. (1 mark)
II. Entity type CUSTOMER in one schema describes an entity type CLIENT in another schema (1 mark)
III. The KEY of an entity type differs in two in schemas being integrated (1 mark)
IV. similar concept represented in two schemas by different modeling constructs. For example, DEPARTMENT is represented as entity type in one schema and an attribute in another. (1 mark)

b) Given the data set 10, 15, 18, 20, 31, 34, 41, 46, 51, 53, 54. You are required to use equal width Binning to discretize the data

i) Determine the width of each bin (2 marks)
ii) Show how you would determine the values to include in each of the 4 bins (2 marks)

iii) Write down the values in each of the four bins Bin 1, Bin 2, Bin 3 and Bin 4 (2 marks)

c) Given the data set 4, 7, 13, 16, 20, 24, 27, 29, 31, 33, 38, 42. you are required to use equal frequency binning to discretize the data into 3 bins then smooth the data by bin boundaries.

i) Explain the purpose of data smoothing (2 marks)

ii) what will be the data in each of the 3 bins before smoothing (2 marks)

iii) determine the data in each of the 3 bins after smoothing (2 marks)

d)

i) Explain what you understand by data cleaning or data scrabbing and state why it is an important step in data preparation (3 marks)

ii) Explain what you understand by data imputation (2 marks)

e) Explain how each of the following models of data integration works

i) Federated Databases (2 marks)

ii) Data Warehouse (2 marks)

iii) Mediatization (2 marks)

QUESTION TWO (20 MARKS)

a) Explain THREE methods of dealing with missing values (6 marks)

b) Consider the following data set with missing values

Number	City	Gender	Age	Income	Illness
1	New York	Male	41	40367	No
2	Los Angele	Male	54	45084	No
3	New York	Male	42	52483	No
4	Los Angele	Male	40	40941	No
5	New York	Male	46	50289	No
6	Dallas	Female		50786	No
7	Dallas	Female	32	33155	No
8	Los Angele	Male	39	30914	No
9	Los Angele	Male	51	68667	No
10	Los Angele	Female	30	45919	No
11	Dallas	Female	48	41524	Yes
12	New York	Male	47	54777	No
13	New York	Male	46	62749	No
14	Dallas		42	50894	No
15	Boston	Female	61	38429	No
16	Boston	Male	43	34074	No
17	Dallas	Male	27	50398	No
18	Dallas	Male		46373	Yes
19	New York	Male	47	51137	No
20	New York	Female	35	23688	No
21	New York	Male	57	17378	No

Explain the method used by the following pandas code snippet to treat the missing values (8 marks)

i). Code snippet 1

```
df_1 = df.dropna( subset =['Gender','Age'] )
```

ii). Code snippet 2

```
df['Income'] = df['Income'].fillna((df['Income'].mean()))
```

iii). Code snippet 3

```
df['Age'] = df['Age'].fillna((df['Age'].median()))
```

iv). Code snippet 4

```
df['Age'] = df['Age'].fillna((df['Age'].mode()))
```

c) Explain the role of the following code snippets in data cleaning and write down the output before data treatment and after data treatment

i). Code snippet 1 (4 marks)

```
Import pandas as pd
df= pd.DataFrame ([[1, 2, 3],[none, none, 6], [none, 9, none]])
print (df)
df1 = df.fillna( method =”ffill)
print (df1)
```

ii). Code snippet 2 (2 marks)

```
import pandas as pd
df = pd.DataFrame(['a','b','c','d','a','b','e'])
df[df.duplicated(keep=False)]
```

QUESTION THREE (20 MARKS)

a) Explain THREE benefits of data transformation (6 marks)

b) Give ONE example of each of the following type of data transformation

i) Constructive transformation (1 mark)
ii) Structural Transformation (1 mark)
iii) Aesthetic Transformation (1 mark)
iv) Destructive Transformation (1 mark)

c) Given the simple exponential smoothing forecast equation as $S_{t+1} = S_t + \alpha (y_t - S_t)$

i). State what each of the parameters S_{t+1}, S_t, y_t and α (alpha) represent (4 marks)

ii). Given a data set with 39 observations determine the value of α (alpha) (1 mark)

d) Sales of water purifiers in a shop in 4 consecutive months are as given below. Compute the forecast for the next month using the exponential smoothening method with a smoothening constant of 0.3 and weighted moving average method with the weights as 0.1, 0.2, 0.3 and 0.4

MONTH	**D_j**
January	1080
February	950
March	1050
April	1120

(3 marks)

e) An electric manufacturer underestimated the January car sales by 20 units while the actual sale was 120 units. If the manufacturer uses exponential smoothing method with a smoothing constant of $\alpha = 0.2$, compute the sales forecast for the month of February for the same year (2 marks)

QUESTION FOUR (20 MARKS)

a)

i) Explain the goal of data normalization in building machine learning models (2 marks)

ii) Suppose that the minimum and maximum values for the attribute income are \$12,000 and \$98,000, respectively. We would like to map income to the range [0.0, 1.0]. By min-max normalization, determine the transformed value of the income \$73,600 . (3 marks)

iii) Suppose that the mean and standard deviation of the values for the attribute income are \$54,000 and \$16,000, respectively. Use the z-score normalization to determine the transformed value of the income \$73,600. (2 marks)

iv) Explain with aid of an example how normalization using decimal scaling is done (3 marks)

b)

i) Explain the importance of data aggregation (2 marks)

ii) Given a relation ***Country (name, continent, population),*** write a valid SQL statement that would aggregate total population per continent (2 marks)

iii) Differentiate between a data cube and a spreadsheets (4 marks)

iv) What are the Key elements of a data cube (2 marks)

QUESTION FIVE (20 MARKS)

a)

i) Explain what you understand by data reduction (2 marks)

ii) Give FOUR benefits of data reduction in data mining (4 marks)

b) explain each of the following data reduction methods

i) Dimensionality reduction (2 marks)

ii) Numerosity reduction (2 marks)

iii) Data compression. (2 marks)

c)

i) Distinguish between lossless and lossy compression (4 marks)

ii) Distinguish between parametric and non-parametric numerosity reduction (4 marks)