NATIONAL UNIVERSITY OF SCIENCE AND TECHNOLOGY

FACULTY OF APPLIED SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

DATA MINING AND WAREHOUSING

SIA 1202

SECOND SEMESTER EXAMINATION PAPER

SECOND SEMESTER 2025

This examination paper consists of 4 pages.

Time Allowed: 3 Hours
Total Marks: 100
Special Requirements: None
Examiner's Name: Mr. S Ncube
External Examiner: Dr LC Sakala

Instruction to Candidates
1. This examination paper consists of Five (5) questions.
2. All questions carry equal marks.
3. Answer any Four (4) questions.

Mark Allocation

Question	Marks
1.	25
2.	25
3.	25
4.	25
5.	25

QUESTION ONE

a) Explain the importance of data mining in modern businesses. [5 marks]
b) With the aid of a diagram, illustrate the general data mining process and explain its key stages. [10 marks]

c) Discuss five (5) major application areas of data mining, providing real-world examples for each. [10 marks]

QUESTION TWO

a) Why is data preprocessing an important step in data mining? [5 marks]
b) Describe any <u>two</u> causes of missing data values you know. [4 Marks]

c) The table 1 below shows a dataset with missing and inconsistent values.

Table 1

ID	Age	Salary	Location	Purchased
1	25	5000	Harare	Yes
2		6000	Bulawayo	No
3	40	?	Harare	Yes
4	35	7000	Gweru	?

Perform the following data pre-processing tasks on the dataset above:
i) Handle missing values using any three methods. [6 marks]

ii) Normalise the Salary column using Min-Max Normalisation where the new range is [0,1]. [5 marks]

iii) Perform data discretisation on the Age column using the binning method (equal-width binning, 2 bins). [5 marks]

QUESTION THREE

a) Define the following terms:

i) Database
ii) Data Mart
iii) Data Warehouse
iv) OLAP
(v) ROLAP [10 marks]

b) Explain the three-tier data warehouse architecture with a diagram.

[9 marks]

c) Describe three (3) major characteristics of a data warehouse.

QUESTION FOUR

a) Explain how Python can be used for data mining, mentioning three (3) key libraries and their functions. [10 marks]
b) The following Python code performs data loading and preprocessing. Fill in the missing parts and explain each step:

```
python
CopyEdit
import pandas as pd

# Load dataset
data = pd._____("<_____>")

# Handle missing values
data.fillna(_____, inplace=True)

# Normalize a column
data['Salary'] = (data['Salary'] -
min(data['Salary'])) / (max(data['Salary']) -
min(data['Salary']))
```

[15 marks]

QUESTION FIVE

a) Differentiate between hierarchical clustering and k-means clustering. [6 marks]

b) The table below contains data points for customer segmentation:

Table 2

Customer ID	Age	Annual Income ($000s)
1	25	30
2	30	40
3	45	80
4	50	90

Using k-means clustering (k=2), perform the first iteration of centroid calculation and cluster assignment. [10 marks]

c) With the aid of an example, explain how Naïve Bayes classification works. [9 marks]

END