NAMIBIA UNIVERSITY
OF SCIENCE AND TECHNOLOGY

FACULTY OF COMPUTING AND INFORMATICS

DEPARTMENT OF INFORMATICS

QUALIFICATIONS: Bachelor of Informatics Honours	
QUALIFICATION CODE: 08BIFH, 08BIHB	**LEVEL: 8**
COURSE CODE: DSA821S	**COURSE:** Data Science and Analytics
DATE: December 2025	**SESSION: 1**
DURATION: 3 Hours	**MARKS: 100**

SUPPLEMENTARY/SECOND OPPORTUNITY EXAMINATION QUESTION PAPER	
EXAMINERS:	Prof. Stephen Fashoto
MODERATOR(S):	Ms. Emilia Shikeenga

THIS EXAMINATION PAPER CONSISTS OF 4 PAGES

(INCLUDING THIS FRONT PAGE)

INSTRUCTIONS FOR THE CANDIDATE

1. Answer any four QUESTIONS.
2. When writing, take into account: The style should inform than impress, it should be formal, in third person, paragraphs set out according to ideas or issues, and the paragraphs flowing in a logical order.
3. Information should be brief and accurate.

Please ensure that your writing is legible, neat and presentable

QUESTION ONE **[25Marks]**

a) Why is 5V's the standard characteristics for the big data technologies and not 3V's? Explain 10marks

b) Write out how to represent the binary class below using a numpy array in python

Index	1	2	3	4	5	6	7	8	9	10
Actual	Dog	Not dog	Dog	Dog	Not dog	Dog	Dog	Dog	Not dog	Dog
predicted	Dog	Dog	Not dog	Dog	Not dog	Not dog	Dog	Dog	Not dog	Not dog

5marks

c) Write short note on how to apply the following with your Data science knowledge

i) Normalization 2marks

ii) Discretization 2marks

iii) Feature selection 2marks

iv) Feature importance 2marks

v) Standardization 2marks

QUESTION TWO **[25Marks]**

a) Differentiate between a binary and multiclass in supervised learning 2marks

b) A set of 1100 pens contains 700 pens of the Parker brand, and the remaining pens are of other brands. A binary classifier correctly identified the 700 Parker pens and incorrectly identified 100 non-Parker pens as Parker.

(i) How many non-Parker pens were correctly identified? 2marks

(ii) Construct the confusion matrix of the classifier 2marks

(iii) Calculate the following based on the confusion matrix in question 1b(ii)

1) Accuracy 2marks

2) Recall 2marks

3) Precision 2marks

4) F1-Score 2marks

5) Specificity 2marks

c) Write short note on the key components of Reinforcement learning with the support of a diagram. 9marks

QUESTION THREE **[25Marks]**

a) Write out the algorithm for implementing K-means clustering 5marks

b) List and explain five reasons why Data Quality is important in Big data technologies? 10marks

c) Given the data point in the table below, initialize the k-means clustering algorithm with two cluster centers c1 =(2,10) and c2=(8,4) using Squared Euclidean distance. What are the values of c1 and c2 after one iteration of k-means clustering? What are the values of c1, and c2 after the second iteration of k-means clustering? 10marks

Squared Euclidean distance formula $d(x,y)=\sum_{i=1}^{n}(x_i - y_i)^2$

Point	Coordinates
X1	(2,10)
X2	(2,5)
X3	(8,4)
X4	(5,8)
X5	(7,5)
X6	(6,4)
X7	(1,2)
X8	(4,9)

QUESTION FOUR **[25Marks]**

a) List and explain the three key skills of a data scientist with the support of a diagram 7marks

b) Explain how SEMMA as a data science methodology can be applied in research 10marks

c) Write short note on the associative rules' terminologies listed below
   i. Antecedent 2marks
   ii. Consequent 2marks
   iii. Support 2marks
   iv. Confidence 2marks

QUESTION FIVE **[25Marks]**

Consider the following dataset in the table below using Apriori algorithm with a minimum support threshold of 55% and minimum confidence threshold of 60%.

Transaction ID	Items bought
T1	Bread, butter, milk
T2	Bread, butter
T3	Bread, milk
T4	Butter, milk
T5	Bread, milk

i)	Find all the frequent itemsets	10marks
ii)	Generate the association rules	10marks
iii)	Calculate the lift results with the interpretation	5marks