



Riara School of Computing
Nurturing business innovators

JAN- MAY 2026 TRIMESTER
EXAMINATION FOR MASTER OF SCIENCE IN DATA SCIENCE AND.
COMPUTATIONAL
INTELLIGENCE
MRDC 911: FUNDAMENTALS OF DATA SCIENCE & COMPUTATIONAL
INTELLIGENCE

DATE: APRIL 2026

TIME: 2 HOURS

GENERAL INSTRUCTIONS:

Students are NOT permitted to write on the examination paper during examination time.

This is a closed book examination. Text book/Reference books/notes are not permitted.

SPECIAL INSTRUCTIONS:

This examination paper consists Questions in Section A followed by section B.

Answer **Question 1 and any Other Two** questions.

QUESTIONS in ALL Sections should be answered in answer booklet(s).

1. **PLEASE start the answer to EACH question on a NEW PAGE.**
2. **Keep your phone(s) switched off at the front of the examination room.**
3. **Keep ALL bags and caps at the front of the examination room and DO NOT refer to ANY unauthorized material before or during the course of the examination.**
4. **ALWAYS show your working.**
5. **Marks indicated in parenthesis i.e. () will be awarded for clear and logical answers.**
6. **Write your REGISTRATION No. clearly on the answer booklet(s).**
7. **For the Questions, write the number of the question on the answer booklet(s) in the order you answered them.**
8. **DO NOT use your PHONE as a CALCULATOR.**
9. **YOU are ONLY ALLOWED to leave the exam room 30minutes to the end of the Exam.**
10. **DO NOT write on the QUESTION PAPER. Use the back of your BOOKLET for any calculations or rough work.**

SECTION A (COMPULSORY)

QUESTION 1: (30 MARKS)

- A. A public health researcher is building a dataset to study malaria spread across sub-Saharan Africa. The dataset will include; patient age, GPS coordinates of residence, number of previous infections, mosquito net usage (Yes/No), average rainfall (mm), and district name. **[4mks]**
- Classify each variable by data type and justify each classification. **[2mks]**
 - The researcher plans to collect this data via door-to-door surveys, satellite imagery, and hospital records. Identify one data quality challenge specific to each collection method and suggest how it could be mitigated. **[2mks]**
- B. Below is a summary statistics table produced for three variables in a Kenyan e-commerce sales dataset containing 4,820 transactions:

Statistic	Amount (KSh)	Delivery Days	Rating
Count	4,820.00	4,820.00	3,941.00
Mean	312.45	4.18	3.72
Std Dev	889.23	2.01	0.98
Min	5	1	1
25th Percentile	75	3	3
Median (50th)	145	4	4
75th Percentile	310	5	4
Max	18,500.00	21	5

- The count for 'rating' is 3,941 while the other two variables each show 4,820. What does this discrepancy indicate, and how should it be addressed before building a predictive model? **[2mks]**
 - The 'amount' variable has a mean of KSh 312.45, yet a maximum of KSh 18,500 and a standard deviation of KSh 889.23. Name the statistical concept that explains this pattern, and describe two data transformation approaches that could address it **[4mks]**
 - Using the percentile values provided, calculate the Interquartile Range (IQR) for 'delivery_days' and determine the upper fence for outlier detection using the standard $1.5 \times \text{IQR}$ rule. Show your working. **[2mks]**
- C. A mobile lending company analyses repayment behaviour and finds the following:
- 60% of borrowers are salaried employees (S), 40% are self-employed (SE)
 - Among salaried borrowers, 8% default
 - Among self-employed borrowers, 22% default

You are part of the analytics team advising management.

- i. Without performing detailed calculations, explain:
 - a) Whether employment type appears to be a strong predictor of default risk **[2mks]**
 - b) Whether employment type alone is sufficient to build a reliable credit risk model **[2mks]**
- ii. The risk team argues that because most defaulters are self-employed, the company should stop lending to self-employed people entirely. Critically evaluate this reasoning from both a statistical and an ethical standpoint. **[3mks]**

D. A data analyst presents two visualizations to the Board of Directors:

- A **pie chart** showing revenue share across 12 product categories.
- A **scatter plot with a fitted trend line** showing the relationship between *time on site* and *purchase value*.

Critically evaluate the effectiveness of each visualization for a board-level audience.

- i) Identify one key weakness of each visualization and recommend a more appropriate alternative for each and justify how it would better support executive decision-making. **[2mks]**
- ii) The scatter plot reports a correlation of 0.87 between *time on site* and *purchase value*. A junior analyst concludes that increasing customers' time on the site will directly increase sales. **[2mks]**

Critically assess this conclusion by:

- iii) Identifying the logical error involved and recommend one appropriate analytical approach that would help determine whether the relationship is causal and suitable for strategic action. **[3mks]**

QUESTION 2 (15 MARKS)

A. In data science, datasets with hundreds of features often cause the 'curse of dimensionality'.

- i. Explain what the curse of dimensionality is and describe two concrete problems it creates when building machine learning models. **[3mks]**
- ii. Principal Component Analysis (PCA) is a linear algebra technique used to reduce dimensionality. Explain how PCA works, what 'variance explained' means, and how a practitioner decides how many components to retain. **[4mks]**

B. A hospital deploys a machine learning model to screen patients for a rare but treatable disease that affects 2% of the population. The model achieves 98% accuracy on the test set.

- i. Without doing any calculations, explain why 98% accuracy is a misleading success metric for this problem. **[3mks]**
- ii. Between precision and recall, which metric should the hospital prioritise for this screening tool and why? What real-world consequence follows from optimising the wrong one? **[3mks]**
- iii. Sketch (describe in words) what the ROC curve for a perfect model, a random model, and a useful but imperfect model would each look like, and explain the significance of the AUC value. **[2mks]**

QUESTION 3 (15 MARKS)

A. A telecommunications company wants to segment its 2 million prepaid subscribers into behavioural groups for targeted marketing. The available features per subscriber are average monthly spend (KSh), average daily data usage (MB), number of calls per week, months since last top-up, and region (one of 8 regions).

- i. Before applying k-means clustering, identify two preprocessing steps that are essential for this particular feature set and explain why each is necessary. **[3mks]**
- ii. Describe the elbow method and the silhouette score as two different approaches for choosing k. Under what circumstances would the elbow method give an ambiguous result, and what would you do in that case? **[3mks]**
- iii. After clustering, the marketing team asks: 'Which cluster should we target with a premium data bundle offer?' Describe what cluster profile characteristics you would look for to answer this question. **[3mks]**

B. A bank collects data on seven customers to train a model predicting whether a customer will subscribe to a savings account. The dataset is shown below:

No.	Age Group	Income Level	Has Existing Product	Subscribed?
-----	-----------	--------------	----------------------	-------------

1	Young	Low	No	No
2	Young	Low	Yes	Yes
3	Middle	High	No	Yes
4	Senior	Medium	No	No
5	Senior	High	Yes	Yes
6	Middle	Low	Yes	No
7	Young	High	No	Yes

- i. Calculate the entropy of the target variable 'Subscribed' for this dataset. Show your working using $H = -\sum p \log_2 p$. **[3mks]**
- ii. A decision tree built on this data achieves 100% accuracy on training data. Explain the concept of overfitting in this context and describe two pruning strategies that could improve generalisation. **[3mks]**

QUESTION 4 (15 MARKS)

A. A county government in Kenya partners with a private tech company to digitise citizen records including land ownership, tax compliance, health history, and household income. The data is stored on the company's cloud servers. Citizens are not formally informed, and there is no data-sharing agreement made public. **[5mks]**

- i. Identify three distinct ethical principles that are violated in this scenario and explain how each is breached.

B. A university in East Africa uses an algorithm to award scholarship shortlists. An audit reveals that female applicants from rural counties are shortlisted at half the rate of male applicants from urban areas, even when academic scores are equivalent. The developers claim the model is 'objective because it only uses historical data'.

- i. Critique the developers' claim. Explain how historical data can encode and amplify existing social inequalities in a model's outputs. **[3mks]**

- ii. Define demographic parity and equalised odds as two fairness metrics. For this scholarship scenario, argue which metric is more appropriate and why. **[2mks]**
 - C. A regional airline collects flight operations data (fuel consumption, delay reasons, maintenance logs, passenger loads) but currently uses it only for monthly compliance reports. A consultant proposes a five-stage roadmap toward predictive operations.
 - i. Map the airline's current state to a level in the Business Information Maturity Model and justify your placement. **[3mks]**
 - ii. Describe what the airline's operations would look like at the highest maturity level, giving two specific data science applications that would be running. **[2mks]**
-

END OF EXAMINATION