WISCONSIN INTERNATIONAL UNIVERSITY COLLEGE-GHANA

MSc. INFORMATION TECHNOLOGY, MSc. CYBERSECURITY, MSc. BUSINESS COMPUTING
FIRST SEMESTER EXAMINATIONS: 2025/2026

SCHOOL OF COMPUTING AND TECHNOLOGY

WMI607: ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING (3 CREDITS)

TAKE-HOME EXAMINATION

INSTRUCTIONS:

1. *This term paper is in seven (7) sections, namely Sections A, B, C, D, E, F and G.*
2. *Answer all questions in the 7 sections and type the answers in PDF or MS Word.*
3. *For sections A to F, you must work with ONE dataset that is directly relevant to your proposed MSc Project.*
4. *For Section G, each student is expected to generate a unique dataset following a specified rule indicated in that section. Answers that match another student's values will be treated as academic misconduct.*
5. *All results must be reproducible.*
6. *You must submit:*
    - *A written report (PDF or MS Word) with maximum of 20 pages*
    - *Dataset, description and/or access proof*
    - *Source code (Python or any one of your choice)*
    - *Experiment logs / outputs*
7. *You may be invited for an oral defense to validate authorship.*
8. *Plagiarism or use of AI-generated content will attract severe penalties.*

TIME ALLOWED:
FORTY-EIGHT (48) HOURS

SECTION A: Dataset–MSc Project Alignment

Question 1 (10 Marks)

Provide a technical description of the dataset you selected and justify its relevance to your MSc Project.
Your answer must include:

- Dataset source and ownership
- Data size (records, features, storage size)
- Data type (tabular, text, image, time series, graph)
- Big data characteristics addressed (Volume, Velocity, Variety, Veracity)
- Why this dataset is non-trivial for standard machine learning approaches

Answers without dataset-specific statistics will score zero.

SECTION B: Data Engineering & Big Data Challenges

Question 2 (15 Marks)

Critically analyze the big data challenges encountered when working with your dataset.

Discuss:

- Data ingestion and preprocessing constraints
- Memory, storage, or computational bottlenecks
- Data imbalance, sparsity, noise, or missingness
- Scaling strategies used (batching, sampling, distributed processing)

You must reference actual runtime behavior or system limitations observed during experimentation.

SECTION C: Feature Engineering & Representation

Question 3 (10 Marks)

Describe and justify the feature engineering or data representation techniques applied.

Your response must include:

- At least two alternative feature representations
- Justification grounded in domain knowledge
- Impact of feature choices on model performance
- Evidence from experiments (tables or plots)

Feature choices must be traceable to your project problem.

SECTION D: Machine Learning Model Design

Question 4 (20 Marks)

Select and implement two machine learning models appropriate for your dataset.

At least one model must be:

- scalable (e.g., tree-based ensembles, etc)
- unsuitable for small toy datasets

For each model:

- Explain why it fits the data characteristics
- State hyperparameters and tuning strategy
- Analyze training time and computational cost
- Discuss convergence or optimization issues

SECTION E: Evaluation, Robustness & Scalability

Question 5 (10 Marks)

Evaluate your models using metrics appropriate to your problem domain.

You must:

- Justify metric selection
- Compare model performance across data scales
- Analyze robustness to noise or imbalance
- Discuss generalization limitations

Use numerical evidence from your experiments.

SECTION F: Project-Driven Insight & Reflection

Question 6 (10 Marks)

Reflect critically on how this machine learning analysis contributes to your proposed MSc Project.

Address:

- What new insight was gained from the dataset?
- How results influence your project methodology
- Limitations discovered through experimentation
- How the approach will be extended in your project work

This section must show research maturity, not technical reporting.

SECTION G: Dataset Generation & Entropy Computation

Question 7 (4 Marks)

Generate a unique dataset of 24 instances with 3 categorical features and 1 binary class label using the following procedure:

- Use your student ID as the random seed
- Features must each have 2–3 possible values
- Class label must be binary (Yes/No or 0/1)

You must:

- Clearly present the dataset in tabular form
- State the seed used and justify feature choices

Question 8 (2.5 Marks)

a. Compute the entropy of the class label.
b. Show all intermediate steps and logarithmic calculations.
c. Briefly interpret what the entropy value implies about class uncertainty.

Question 9 (6 Marks)

For each feature in your dataset:
a. Compute the conditional entropy
b. Calculate Information Gain
c. Rank features from highest to lowest IG

All calculations must be shown symbolically and numerically. Spreadsheet-only answers will not be awarded full marks. Thus, handwritten entropy calculations scanned or a snapshot taken is recommended.

Question 10 (1.5 Marks)

a. Select the root node based on Information Gain
b. Construct the first two levels of the decision tree manually
c. Draw the tree clearly and label:

- Splitting attributes
- Branch values
- Class outcomes

Question 11 (6 Marks)

Modify one data instance in your dataset:

a. Recompute entropy and Information Gain
b. Discuss how and why the root node changes or remains the same

Question 12 (5 Marks)

In 300–400 words, answer:

- Why Information Gain can be biased?
- How alternative criteria (Gain Ratio, Gini Index) might behave on your dataset?
- Whether a different splitting criterion would be preferable?

Note: Each dataset generated in Section G is unique. Answers that match another student's values will be treated as academic misconduct.