HW 1 — Introduction to Data Mining

Deadline: 02/17/2026, midnight

Instructions to Students

- You may only use **NumPy** and **matplotlib**
- **Do NOT use sklearn metrics or roc utilities**
- Do NOT change function names or return formats
- Do NOT change Python file name
- Your functions will be tested using **multiple hidden test cases**

Task 1: Confusion Matrix & Evaluation Metrics (40 points)

You are given a **confusion matrix**, which may be:

- Binary (2×2), or
- Multi-class (K×K)

You must:

1. Convert a multi-class confusion matrix into **binary confusion matrices** using **one-vs-rest**
2. Compute the following metrics **for each class**:
   - Accuracy
   - Recall
   - Precision
   - F1-score
3. Return results in a **NumPy array**

Notes for Students

- Handle **division by zero** properly (return 0.0)
- Metrics must be **float**
- Do **not** assume binary-only input
- Output order **must match the specification**
- confusion_matrix_metrics() function in the student_hw1.py file.

Task 2: ROC Curve from Scratch (40 points)

Given:

- Ground-truth labels (binary: 0 or 1)
- Predicted probabilities for the input (range from 0 to 1)

You must:

1. Sweep **~100 thresholds** from 1 → 0
2. Compute **TPR** and **FPR** at each threshold
3. Plot the ROC curve
4. Plot two red points with threshold 0.5 and 0.8 in the plot.

Notes for Students

- Do **not** use sklearn's `roc_curve`
- Use `>= threshold` for positive prediction

- ROC must be **monotonically increasing**
- Code should work for **any valid input size**
- plot_roc_curve() function in the student_hw1.py file.

Task 3: [Open Question] Can ROC curve handle imbalanced datasets? (20 points)

Requirements:

(1) Empirical Analysis

You must:

- Construct [ground-truth labels, predicted probabilities] for both balanced and imbalanced dataset for binary classification.
- Reuse the code in Task 2 and plot the ROC curves for both datasets on the same or separate figures.

(2) Conceptual Explanation

In your own words, answer whether ROC curves can handle imbalanced datasets, and explain why?

Submission:

- Include ROC plots
- Include a short written explanation (≈ 1–2 paragraphs)
- In a **PDF** file.

Submission Checklist (for Students)

- Functions implemented
- No sklearn metrics used
- Correct return shapes
- Code runs without errors
- Submitted files: [HW1-YourFirstNameLastName.zip]
  - student_hw1.py
  - task-3.pdf