Purpose of the Case Study

In this case study, you will step into the role of a data scientist tasked with helping a sustainability initiative improve waste sorting accuracy using computer vision. You will analyze a real dataset (RealWaster), restructure its labeling scheme, and build a simplified convolutional neural network (CNN) to classify images into three broad waste categories: Recyclable, Compostable, and Trash. The purpose of this assignment is for you to practice working with real-world image data, understand the impact of class imbalance and label grouping, and produce a concise deliverable demonstrating your findings.

Task Description

You will complete the following tasks to execute the case study:

1. **Understand the Problem Context**
   Read the hook document and supplemental materials to understand why waste classification matters and how machine learning is used in environmental sustainability.
2. **Load and Explore the Dataset**
   - Examine the RealWaster dataset
   - Review the original nine waste categories
   - Identify issues such as class imbalance and underrepresented classes
3. **Reorganize Labels into Three Groups**
   - Recyclable
   - Compostable
   - Trash (non-recyclable, non-compostable)
     This regrouping helps mitigate class imbalance and improves classification performance
4. **Build a Simple CNN Pipeline**
   Using the provided starter code, construct a lightweight convolutional neural network to classify images into the three new categories. You may modify the architecture if you wish, but the emphasis is on demonstrating understanding, not complexity.
5. **Train, Validate, and Evaluate the Model**
   - Split data appropriately (train/validation)
   - Track accuracy, precision, recall, and F1 score
   - Produce a 3x3 confusion matrix and interpret the model's performance
   - Reflect on uncertainty, misclassified patterns, and the effects of regrouping
6. **Produce the Final Deliverable**
   Create a short, clear report or Jupyter Notebook that includes:
   - The regrouping logic
   - CNN structure
   - Evaluation metrics
   - Confusion matrix
   - A brief discussion of results, limitations, and next steps

Student Deliverables

Your final deliverable must include:

1. **A reproducible notebook or script** implementing the 3-class CNN pipeline.
2. **A written report (max 2 pages)** summarizing:
    a. Your approach to relabeling
    b. Model structure
    c. Performance results
    d. Integration of metrics and confusion matrix
    e. Bias/uncertainty considerations
3. **All code and materials committed to the GitHub repository**

Criteria for Success

You will have succeeded in this case study if you:

- Demonstrate a clear understanding of the problem and motivation
- Correctly reorganize the dataset into three appropriate classes
- Build a functioning CNN pipeline using the provided data
- Evaluate your model using accuracy + precision/recall/F1 + confusion matrix
- Interpret the misclassification patterns and uncertainty thoughtfully
- Present your results concisely and clearly in your written deliverable
- Submit a repository that is clean, organized, and reproducible

Assessment Rubric

Category	Meets Expectations	Below Expectations
Understanding of Context & Motivation	Explains the purpose of waste classification, environmental motivation, and why machine learning is appropriate. Shows strong comprehension of the problem setup.	Provides limited or unclear discussion of context; missing or incomplete understanding of the motivation.
Data Exploration & Relabeling	Correctly identifies original class structure, articulates imbalance, and implements the 3-class regrouping with justification.	Does not clearly explain imbalance, regrouping is missing or incorrect, or justification is weak.
CNN Pipeline Construction	Builds a functioning CNN using the provided structure; shows understanding of	CNN is incomplete, does not run, or shows misunderstanding of basic

	model architecture and training process.	architecture.
Evaluation & Interpretation	Reports accuracy, precision, recall, F1, and confusion matrix. Provides thoughtful interpretation of results and uncertainties.	Metrics missing or misinterpreted; confusion matrix not explained; little to no performance discussion.
Final Deliverable Quality	Clear, well-organized report/notebook with correct results, readable code, and concise reflections.	Report is unclear, disorganized, missing results, or difficult to follow.
Repository Organization	GitHub repo includes all required materials (hook, rubric, data link, scripts, supplemental materials) and is easy to navigate.	Missing required files; repo structure unclear or incomplete.

Length & Format Expectation

- Report: **Maximum 2 pages**
- Deliverable: **Notebook or script + written summary**
- **All materials must be included in the GitHub repo**

Support

Use the provided explainer article and starter code. This case study is designed to be challenging but achievable with the foundational knowledge of Python, image preprocessing, and introductory deep learning.