Prob Statement

OCR Pipeline Assignment – Handwritten document PII Extraction

1. Objective

Build a simple OCR + PII-extraction pipeline for handwritten documents in JPEG format.

End-to-end flow:
Input (handwritten JPEG) → Pre-processing → OCR → Text Cleaning → PII Detection → (Optional) Redacted Image

2. Input Samples (find samples below in ‘samples’ tab**)**

You will be given:

- 2–3 handwritten document images (JPEG)

Your pipeline must work for:

- Slightly tilted images
- Different handwriting styles
- Basic doctor/clinic-style notes or forms

samples

1. https://drive.google.com/file/d/1fZNjzBZXcjvPSAsuz_zHf5fba2niGuy8/view?usp=sharing

2. https://drive.google.com/file/d/1Ud8x3VouXcj0EfOmmxw8s5OgmGBB9Rec/view?usp=sharing

3. https://drive.google.com/file/d/16o3Ukj8Nz8FhvCXeH4Nc1X2OwIV65WQl/view?usp=sharing

Delivery:

1. Python Notebook file
2. Dependency document
3. Results screenshot for test document
4. We will run with another set of documents for benchmarking