Document Ingestion Pipeline

Objective

Design and implement a **document ingestion pipeline** that:

- Takes **PDFs** (could be **scanned**, **multi-column**, or **containing tables and diagrams**)
- Extracts structured content (text, tables, images, and metadata)
- Converts them into **Markdown** format with proper sectioning

Input

- A folder of **PDFs** (you can provide 3–5 samples)
  - Some normal digital PDFs
  - Some **scanned PDFs** (images)
  - Some with **multi-column layout** or **embedded tables**

2. Processing Steps

a. Text Extraction

- Extract text while preserving **headings**, **paragraph structure**, and **column flow**
- Handle scanned PDFs via **OCR (e.g., Tesseract)**
- Identify tables and convert them into Markdown tables

b. Diagram / Figure Extraction

- Extract images or diagrams and save them separately
- In the Markdown output, insert image references (e.g.,[figure1](./figures/figure1.png))

c. Content Structuring

- Create a **Markdown (.md)** file for each document:
  - Include title, author, metadata (if available)
  - Extract and format:
    - Table of Contents (if exists)
    - Sections and subsections
    - Tables in Markdown format
    - Image references