**Docker – Report**

* **Contains a very simple code from pathways website. Docker was set up successfully and gave the expected output**

**KYC Checker – Report**

* **A very simple program that uses PAN name as a base and finds that string in all other documents for verification.**

**1. Functions**

* **File Upload:**
  + Accepts **PAN card** (mandatory) and **at least one other document**.
  + Supports .pdf, .txt, .png, .jpg, .jpeg.
* **Text Extraction (extract\_text):**
  + PDFs: Uses **PyMuPDF** to read text.
  + TXT: Reads directly.
  + Images: Uses **EasyOCR** to extract text.
* **PAN Card Extraction (extract\_info\_pan):**
  + PAN number: line after “Permanent Account Number Card” matching regex [A-Z]{5}\d{4}[A-Z].
  + Name: line after PAN number.
  + DOB: line after “DOB” or “Date of Birth”.
* **Other Document Extraction (extract\_info):**
  + Aadhaar: Detects 12-digit number (with or without spaces).
  + Other IDs: Any 9–12 digit number not matching common phone numbers.
  + Name: Line after name: if present.
* **Verification (verify\_with\_pan):**
  + Compares **first name from PAN** (case-insensitive) with other documents.
  + Flags mismatch if first name not found.
  + Ignores DOB.
* **Cleanup:**
  + Deletes all uploaded files after verification.

**2. How It Works**

1. **Upload:** User uploads PAN card and one or more documents.
2. **Text Extraction:**
   * PDFs/TXT: read directly.
   * Images: OCR with EasyOCR.
3. **PAN Parsing:** Extract name, PAN number, and DOB from PAN card.
4. **Other Docs Parsing:** Extract IDs (Aadhaar-specific or generic 9–12 digit) and optional names.
5. **Verification:** Compare **first name from PAN** against other documents.
6. **Output:** Shows:
   * PAN card info (name, DOB, PAN number)
   * Other document IDs
   * Verification summary: “Verified” or “Mismatch Found”
   * Flags for mismatched documents
7. **Cleanup:** All uploaded files removed automatically.

**3. Limitations**

* **First-name only verification:** Does not compare full names; middle/last names ignored.
* **DOB ignored:** DOB mismatch cannot be detected.
* **OCR accuracy:** Image text extraction can fail with poor-quality scans or unusual fonts.
* **ID detection heuristic:** Generic 9–12 digit fallback may pick irrelevant numbers if multiple numbers exist.
* **No persistent storage:** Cannot track users across multiple sessions.
* **No advanced fraud detection:** Only basic name matching and ID extraction.
* **File types limited:** Only PDFs, TXT, and common image formats.