PDF Data Extraction

Objective:

Develop an application that extracts data from customer-filled PDF forms, converts the extracted data into Excel or CSV format. The application must accurately extract various types of customer and shop details, including those indicated by tick marks, cross marks, or shaded boxes.

Requirements:

Input:

PDF forms filled by customers, containing details such as:

- Customer Name

- Location

- Phone Number

- Shop details (e.g., floor materials used, roof type, door materials, alarm system)

These details may be selected by tick marks, cross marks, or shaded boxes against respective options. PDFs may also contain images.

Output:

An Excel or CSV file containing all the extracted details.

Features:

1. **Form Data Extraction:**

- Accurately extract text and selection marks from PDF forms

. - Identify and interpret tick marks, cross marks, and shaded boxes to determine selected options.

2. **Image Processing:**

- Handle and process images within the PDF if they contain relevant information.

3. **Data Conversion:**

- Convert the extracted data into structured formats such as Excel or CSV.

4. **Multi-language Support:**

- Ensure the solution can handle text in various languages and scripts.

Technologies to Use:

1. **Programming Language:**

- Python (for data processing and scripting)

2. **Libraries and Frameworks:**

- `PyPDF2` or `pdfminer` (for PDF text extraction)

- `Pillow` or `OpenCV` (for image processing)

- `pytesseract` (for OCR) - `pandas` (for data manipulation and conversion to Excel/CSV)

3. **Machine Learning:**

- Pre-trained models or custom models for recognizing and interpreting marks on forms.

4. **API Integration:**

- Gemini API (if applicable for any additional processing needs)

5. **Frontend (Optional):**

- Any technology (to create a simple user interface for uploading PDFs and downloading processed files)

Deliverables:

1. Source code of the application.

2. Documentation on how to set up and run the application.

3. A demo video showing the application in action.