# TRABAJO PRÁCTICO N° 1 - AÑO 2025 - 2° SEMESTRE - PROCESAMIENTO DE IMÁGENES I - IA 4.4

ALUMNOS : SANZ ALFREDO – DI PRINZIO ELVIO

# TRABAJO PRÁCTICO N° 1 - AÑO 2025 - 2° SEMESTRE

## TP1_2025-C2

**Informe Técnico: Validación y Procesamiento de Formularios Escaneados**

Este informe describe el proceso y la lógica implementada en el código Python proporcionado, cuyo objetivo principal es la **validación automática** del llenado de formularios escaneados. El sistema utiliza técnicas de **Procesamiento Digital de Imágenes (PDI)** con la librería cv2 (OpenCV) para localizar campos y analizar su contenido, y pandas para estructurar y reportar los resultados.

---

**1. Propósito del Código**

El código está diseñado para:

1. **Localizar y segmentar** los campos de datos dentro de una imagen de formulario escaneado.
2. **Extraer y analizar** el contenido de cada campo (texto o selección).
3. **Validar** si el contenido cumple con criterios predefinidos para considerar el campo como **"OK"** o **"MAL"**.
4. Generar un **informe consolidado** y un archivo **CSV** (validacion_resultados_final.csv) con el resultado de la validación por campo y una validación global para cada formulario.

---

**2. Archivos y Estructura de Datos**

El sistema opera sobre una lista de archivos de imagen de formulario (.png) y define la estructura de datos para el reporte:

| Variable | Descripción |
|---|---|
| ARCHIVOS_FORMULARIO | Lista de archivos de entrada (ej: formulario_01.png a formulario_05.png). |

| Variable | Descripción |
|---|---|
| CAMPO_NOMBRES | Nombres de los campos del formulario (Nombre y Apellido, Edad, Mail, Legajo, Pregunta 1, Pregunta 2, Pregunta 3, Comentarios). |
| CSV_OUTPUT | Nombre del archivo de salida para el reporte final (validacion_resultados_final.csv). |

---

**3. Descripción de las Funciones Clave y Lógica de Procesamiento**

El flujo de procesamiento principal se realiza a través de la función procesar_formulario(file_path).

**3.1. Detección de Coordenadas de Campos (detectar_coordenadas_campos)**

**Paso de Localización (OCR Layout Analysis):**

1. **Umbralización y Proyección:** La imagen en escala de grises se umbraliza inversamente (cv2.THRESH_BINARY_INV). Se calculan las **proyecciones horizontales y verticales** (sumas de píxeles) para detectar las líneas que delimitan los campos del formulario.
2. **Identificación de Líneas:** Se aplican umbrales a las proyecciones para identificar las coordenadas (filas y columnas) donde las líneas son más densas.
3. **Filtrado y Asignación:** Las coordenadas de línea detectadas (y_lines y x_lines) se utilizan para calcular las regiones (coordenadas $(y_1, y_2, x_1, x_2)$) correspondientes a cada campo definido en CAMPO_NOMBRES.
    - **Campos de Pregunta (Checkboxes):** Para las preguntas (1 a 3), se definen **dos celdas adicionales** con nombre P#_Si y P#_No, que corresponden a las áreas de las casillas de verificación.

**3.2. Extracción de Contenido de la Celda (extraer_contenido_celda)**

**Paso de Análisis de Contenido:**

1. **Segmentación:** Se recorta la región de interés (celda) de la imagen.
2. **Preprocesamiento:** Se aplica un filtro de mediana (cv2.medianBlur) y una **Umbralización de Otsu** (cv2.THRESH_OTSU) para aislar la tinta (texto o marcas) del fondo.
3. **Análisis de Componentes Conectados (CC):** Se utiliza cv2.connectedComponentsWithStats para encontrar todas las regiones de tinta (píxeles blancos) en la celda umbralizada. Cada región representa un posible carácter, parte de un carácter, o una marca.
4. **Filtrado y Conteo:** Se filtran los componentes conectados por un área mínima (MIN_CHAR_AREA).
    - El número de componentes conectados (num_componentes) es la métrica clave para inferir la presencia y cantidad de texto.

- Se realiza una estimación rudimentaria del número de palabras (num_palabras_estimado) y caracteres a partir del conteo de componentes y el ancho total de los *bounding boxes*.

5. **Resultado:** Devuelve un diccionario con el tipo de contenido (vacío o texto) y las estadísticas de los componentes conectados.

**3.3. Validación del Campo (validar_campo)**

**Paso de Lógica de Negocio/Validación:**

Esta función aplica reglas de validación específicas para cada tipo de campo, basándose principalmente en el **número de componentes conectados** (num_componentes) detectados en la celda.

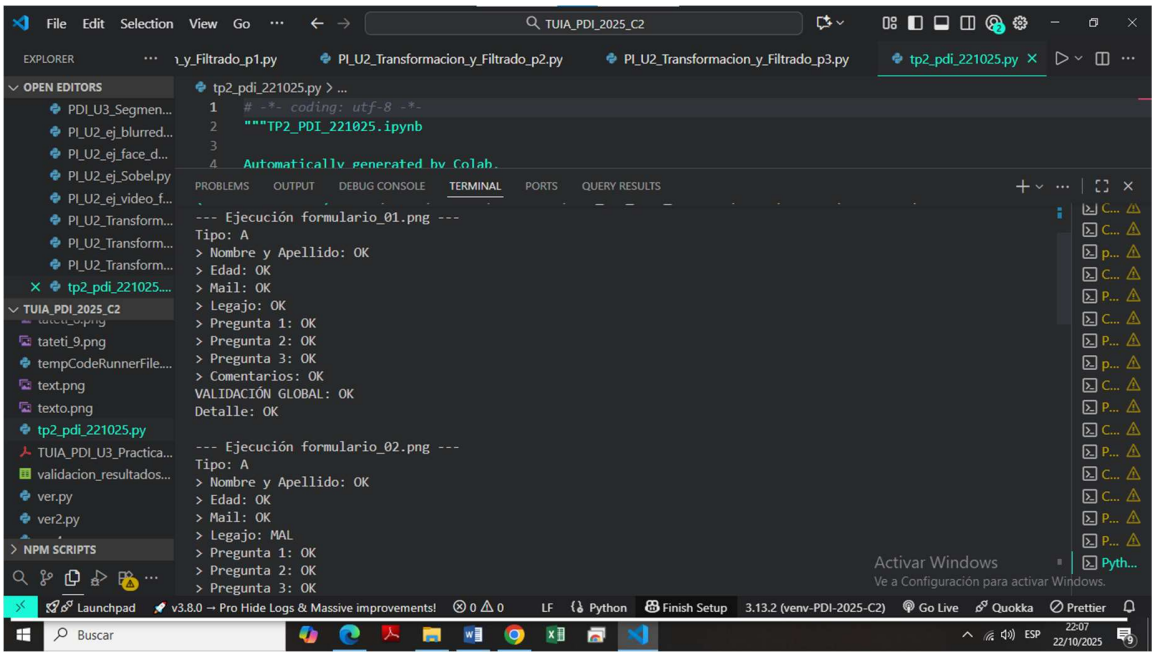
| **Campo** | **Criterio de Validación (Métricas)** | **Reglas (Basadas en Conteo de Componentes)** |
|---|---|---|
| **Nombre, Edad, Legajo, Comentarios** | Conteo de Componentes en la celda de datos. | Se definen rangos mínimos y máximos (ej: MIN_CC_ESTRICTO2, MAX_CC_LARGO) para inferir si hay texto razonable. Un campo "Edad" debe tener un conteo menor que "Comentarios". |
| **Mail** | Conteo de Componentes en la celda de datos. | Criterios más laxos. **Excepción:** Si el campo está vacío (0 componentes), se considera **"OK"** (posiblemente aceptando que el campo puede ser opcional). |
| **Pregunta 1, 2, 3** | Conteo de Componentes en las sub-celdas **Si** (info_si) y **No** (info_no). | Se compara el conteo de componentes entre las celdas Si y No. Se requiere que **una sea dominante** sobre la otra por un factor de RATIO_THRESHOLD (1.1) y que el conteo no exceda MAX_CC_CHECK para evitar falsos positivos por ruido. |
| **Resultado** | | Devuelve **"OK"** si cumple con las reglas, o **"MAL"** si no las cumple. |

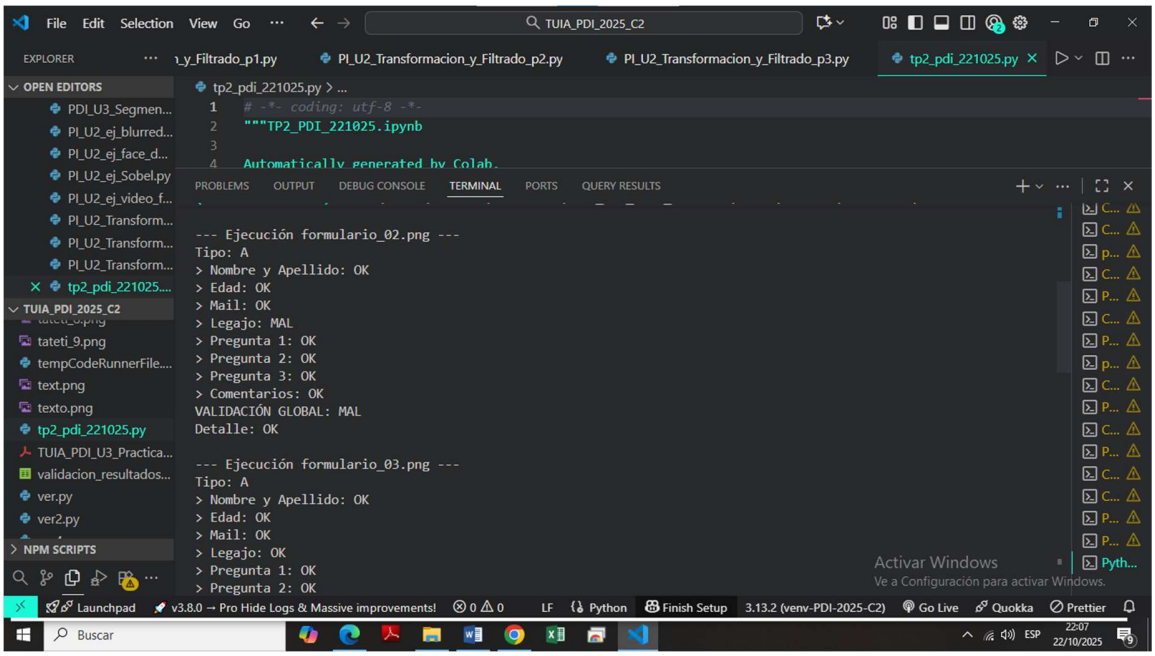
---

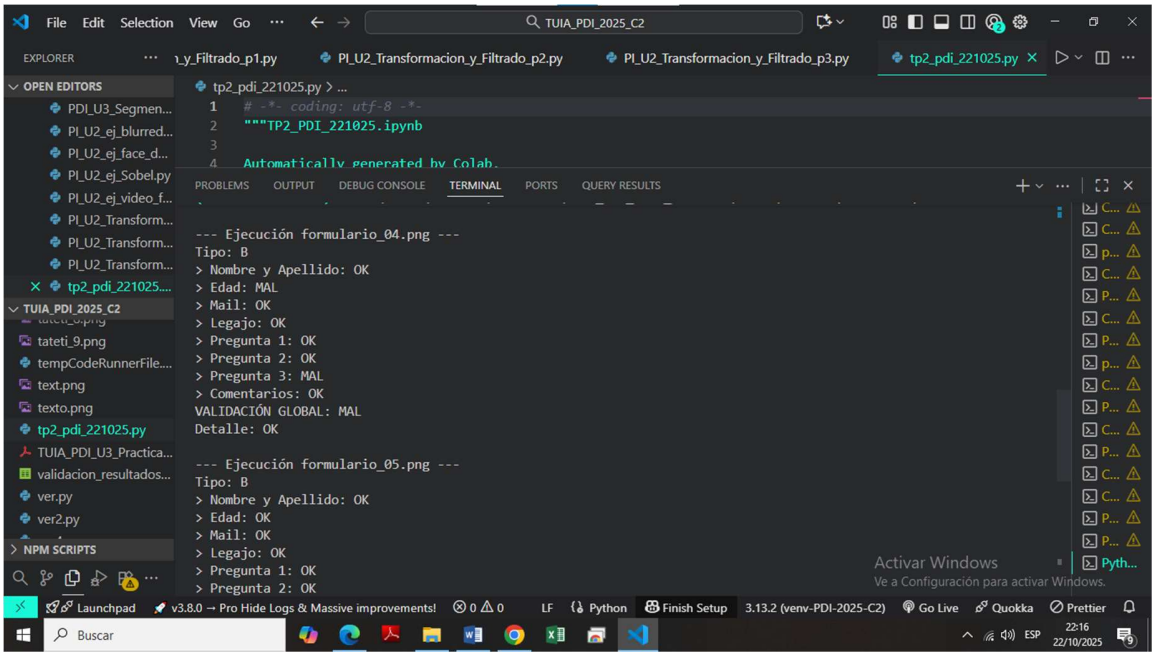
**4. Proceso de Reporte Final**

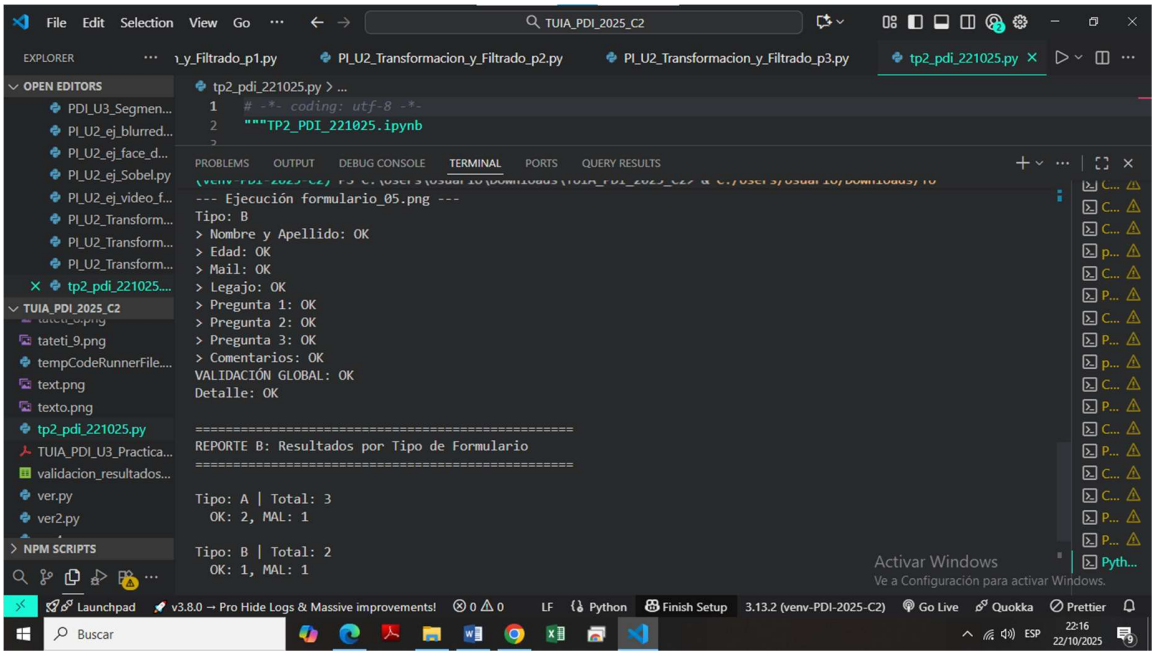
1. **Iteración:** Se ejecuta procesar_formulario para cada archivo en ARCHIVOS_FORMULARIO.
2. **Consolidación: Reporte A**: Los resultados de cada formulario (ID, Tipo, Validación de cada campo y Validación Global) se recopilan en todos_los_resultados.
3. **Reporte B (Por Tipo):** Se crea un *DataFrame* de pandas para agrupar los resultados por Tipo de formulario (A, B, C) y se imprime el conteo de validaciones **OK** y **MAL** por cada tipo.

4. **Reporte C (CSV):** Se genera el archivo **CSV** final con la validación de cada formulario y campo.




File Edit Selection View Go
TUIA_PDI_2025_C2
EXPLORER
PI_U2_Transformacion_y_Filtrado_p1.py
PI_U2_Transformacion_y_Filtrado_p2.py
PI_U2_Transformacion_y_Filtrado_p3.py
tp2_pdi_221025.py
OPEN EDITORS
tp2_pdi_221025.py > ...
# -*- coding: utf-8 -*-
"""TP2_PDI_221025.ipynb
Automatically generated by Colab.
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS QUERY RESULTS
--- Ejecución formulario_04.png ---
Tipo: B
> Nombre y Apellido: OK
> Edad: MAL
> Mail: OK
> Legajo: OK
> Pregunta 1: OK
> Pregunta 2: OK
> Pregunta 3: MAL
> Comentarios: OK
VALIDACIÓN GLOBAL: MAL
Detalle: OK
--- Ejecución formulario_05.png ---
Tipo: B
> Nombre y Apellido: OK
> Edad: OK
> Mail: OK
> Legajo: OK
> Pregunta 1: OK
> Pregunta 2: OK
TUIA_PDI_2025_C2
tateti_9.png
tempCodeRunnerFile....
text.png
texto.png
tp2_pdi_221025.py
TUIA_PDI_U3_Practica...
validacion_resultados...
ver.py
ver2.py
NPM SCRIPTS
Activar Windows
Ve a Configuración para activar Windows.
Launchpad
v3.8.0 → Pro Hide Logs & Massive improvements!
LF
Python
Finish Setup
3.13.2 (venv-PDI-2025-C2)
Go Live
Quokka
Prettier
Buscar
ESP
22:16
22/10/2025


File Edit Selection View Go
TUIA_PDI_2025_C2
EXPLORER
PI_U2_Transformacion_y_Filtrado_p1.py
PI_U2_Transformacion_y_Filtrado_p2.py
PI_U2_Transformacion_y_Filtrado_p3.py
tp2_pdi_221025.py
OPEN EDITORS
tp2_pdi_221025.py > ...
# -*- coding: utf-8 -*-
"""TP2_PDI_221025.ipynb
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS QUERY RESULTS
--- Ejecución formulario_05.png ---
Tipo: B
> Nombre y Apellido: OK
> Edad: OK
> Mail: OK
> Legajo: OK
> Pregunta 1: OK
> Pregunta 2: OK
> Pregunta 3: OK
> Comentarios: OK
VALIDACIÓN GLOBAL: OK
Detalle: OK
==========
REPORTE B: Resultados por Tipo de Formulario
==========
Tipo: A | Total: 3
OK: 2, MAL: 1
Tipo: B | Total: 2
OK: 1, MAL: 1
TUIA_PDI_2025_C2
tateti_9.png
tempCodeRunnerFile....
text.png
texto.png
tp2_pdi_221025.py
TUIA_PDI_U3_Practica...
validacion_resultados...
ver.py
ver2.py
NPM SCRIPTS
Activar Windows
Ve a Configuración para activar Windows.
Launchpad
v3.8.0 → Pro Hide Logs & Massive improvements!
LF
Python
Finish Setup
3.13.2 (venv-PDI-2025-C2)
Go Live
Quokka
Prettier
Buscar
ESP
22:16
22/10/2025