UNIVERSIDAD
MARIANO GALVEZ

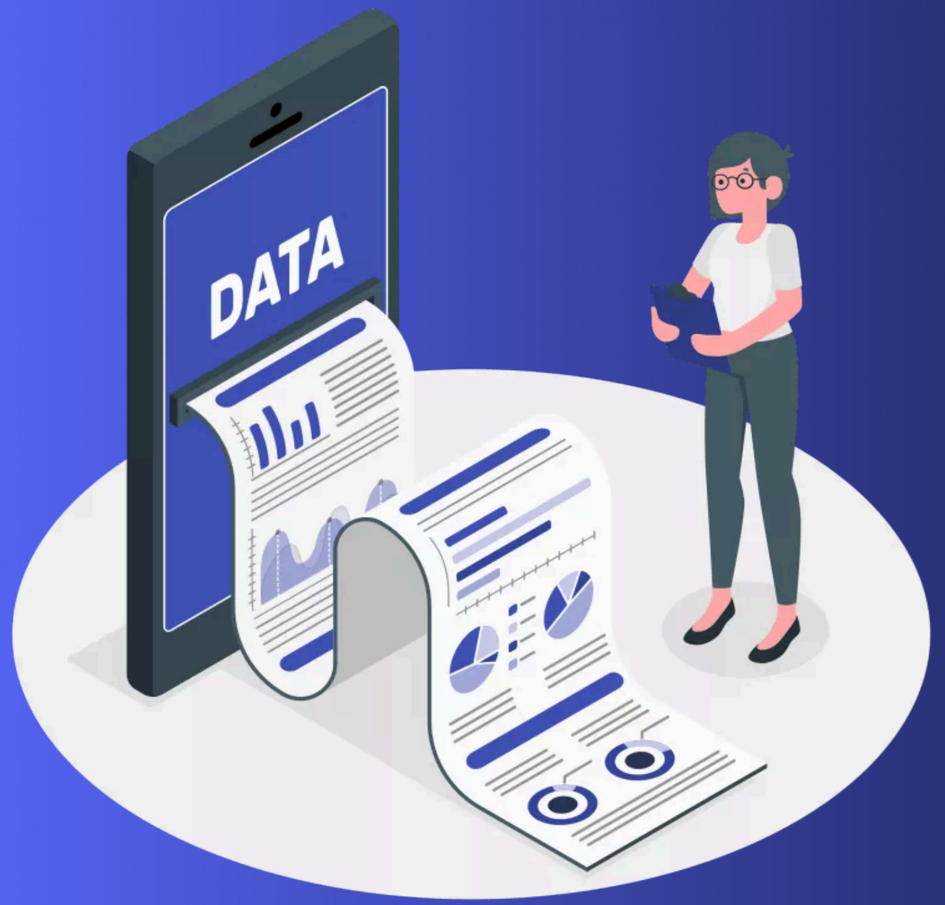
SISTEMA DE GESTIÓN Y EXTRACCIÓN DE INFORMACIÓN EN DOCUMENTOS DIGITALES MEDIANTE PLANTILLAS PERSONALIZADAS Y PROCESAMIENTO AVANZADO DE IMÁGENES PARA EL MANEJO DE GRANDES VOLÚMENES DE DATOS

JAIME ALEXANDER RAX CAAL
0902 20 15240

UNIVERSIDAD
MARIANO GALVEZ

INTRODUCCIÓN

Esta presentación aborda el desarrollo y resultados de un proyecto orientado a la automatización de la extracción de datos desde imágenes y documentos escaneados mediante el uso de tecnologías de OCR avanzadas y plantillas personalizables. El propósito es mejorar la eficiencia y precisión en la captura de datos estructurados en distintos tipos de documentos.



UNIVERSIDAD MARIANO GALVEZ

OBJETIVOS

OBJETIVO GENERAL

Crear un sistema de captura y extracción de datos automatizada basado en plantillas.

OBJETIVOS ESPECIFICOS

- Implementar herramientas avanzadas para la digitalización de datos.

- Reducir trabajo manual mediante un procesamiento automatizado de documentos.

- Mejorar la precisión y agilidad en la extracción de información en imágenes.

DESARROLLO DEL SISTEMA

UNIVERSIDAD MARIANO GALVEZ

Arquitectura del Sistema:

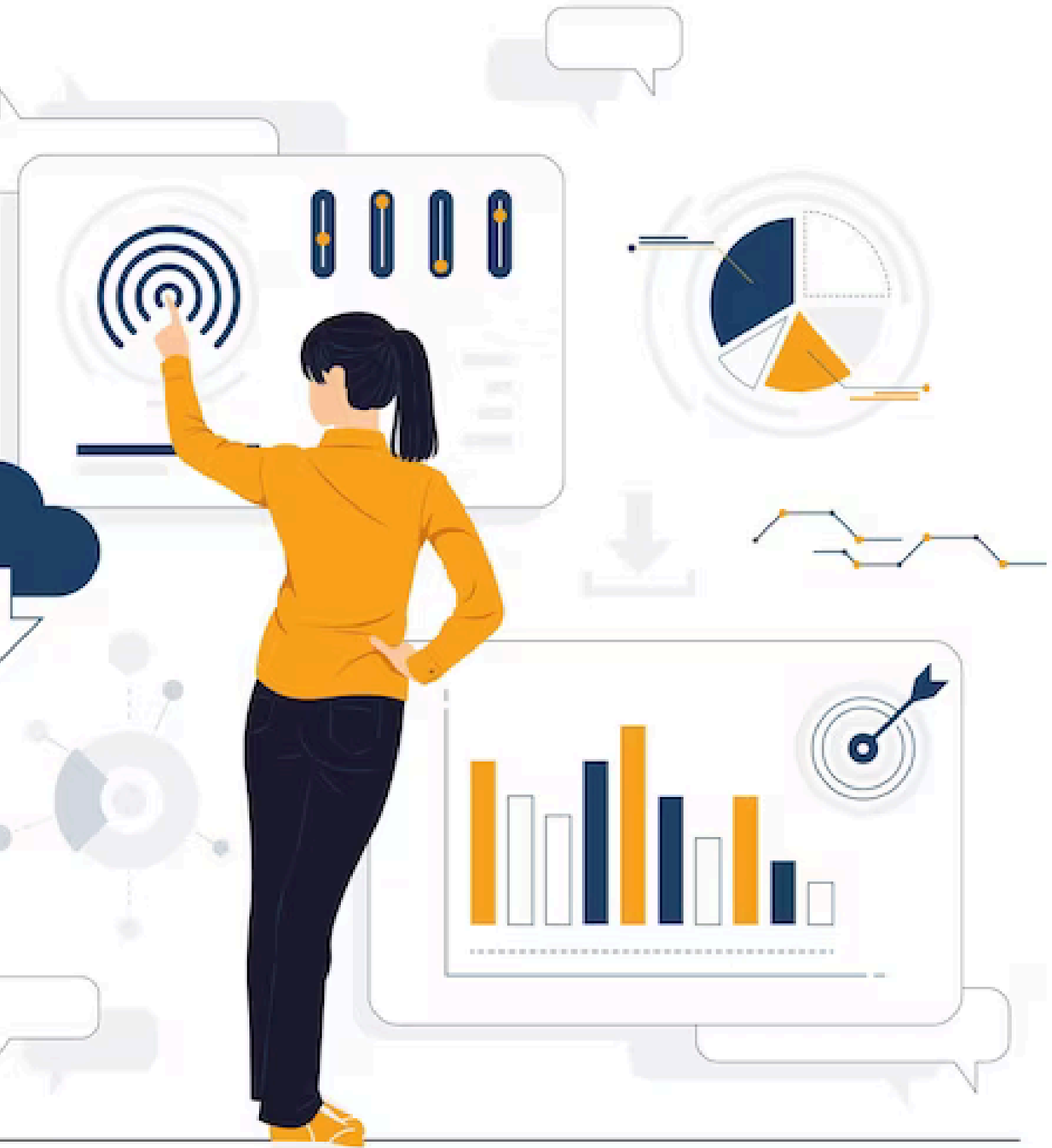
- La aplicación se compone de dos componentes principales: un **backend** desarrollado en Python con Flask y un **frontend** implementado en Next.js y React.
- El backend se encarga del procesamiento y extracción de datos mediante OCR, mientras que el frontend ofrece una interfaz para que los usuarios suban documentos y definan áreas de interés de forma visual.

Componentes del Backend:

- OCR con Tesseract: Utilizado para extraer texto desde imágenes, optimizado para trabajar con plantillas personalizadas y configurado con datos específicos de idioma.
- API REST en Flask: Se desarrolla un conjunto de endpoints en Flask que permite la comunicación entre el frontend y el backend. A través de esta API, se gestionan la carga de imágenes, la selección de áreas de interés y el retorno de resultados en formato JSON y ZIP.
- Almacenamiento de datos y resultados: El backend gestiona y organiza los datos extraídos en una estructura de archivos que incluye tanto las imágenes recortadas como archivos JSON y ZIP, que el usuario puede descargar.

Componentes del Frontend:

- Interfaz de Usuario en Next.js y React: Proporciona un sistema intuitivo de carga de imágenes, selección de áreas de interés mediante anotaciones visuales y gestión de las solicitudes de extracción de datos.
- Integración con el backend: Utiliza Axios para conectar la interfaz con la API, gestionando la carga de documentos y el procesamiento en tiempo real.



UNIVERSIDAD MARIANO GALVEZ

DESPLIEGUE

Servidores y Despliegue:

- Backend en VPS: El servidor de la API está desplegado en una VPS, lo cual permite el procesamiento intensivo de OCR y garantiza la seguridad de los datos mediante HTTPS.
- Frontend en Vercel: La interfaz está alojada en Vercel, proporcionando escalabilidad y una experiencia de usuario optimizada sin comprometer el rendimiento.

Enfoque Modular y Escalable:

El código fue diseñado de manera modular para facilitar su adaptación a distintos tipos de documentos. La arquitectura modular permite añadir funcionalidades adicionales de extracción o nuevas plantillas de documentos según sea necesario.

UNIVERSIDAD MARIANO GALVEZ

RESULTADOS

la implementación del sistema de extracción de datos automatizada demostró mejoras notables en eficiencia, precisión y manejo de grandes volúmenes de información. Gracias a la tecnología OCR y las plantillas personalizables, el sistema agilizó la captura de datos, redujo errores y minimizó la intervención manual, optimizando el flujo de trabajo en documentos visuales complejos.

- Eficiencia: El sistema permitió una extracción de datos hasta un 70% más rápida en comparación con métodos manuales.
- Precisión: Un 85% de precisión en pruebas controladas de extracción de texto.
- Usabilidad: Los usuarios reportaron una mejora en la agilidad y en la precisión del manejo de grandes volúmenes de datos.
- Automatización: Reducción significativa del trabajo manual en la captura de datos de documentos densos.

UNIVERSIDAD
MARIANO GALVEZ

CONCLUISIÓN

- **Impacto**: El sistema contribuye a la reducción de costos y mejora la precisión en la gestión de datos.

- **Importancia**: Las plantillas personalizables y la tecnología OCR proporcionan un sistema adaptable y eficiente para entornos donde se manejen documentos visuales densos.

- **Futuro**: Oportunidad de expandir el sistema para su uso en otros sectores donde el procesamiento de documentos visuales es clave.