FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

Tesis

Digitalización basada en reconocimiento óptico de caracteres para mejorar la disponibilidad de actas de evaluación en la UGEL San Martín

Para optar el título profesional de Ingeniero de Sistemas e Informática

Autores:
César David Paredes Torres
https://orcid.org/0000-0002-2047-1454

Aranza Luccia Marcelo Vasquez
https://orcid.org/0000-0002-8768-8113

Asesor:
Ing. Mtro. Cristian Werner García Estrella
https://orcid.org/0000-0002-5687-8694

Coasesor:
Ing. Lloy Pool Pinedo Tuanama
https://orcid.org/0000-0002-5569-8739

Tarapoto, Perú

2024

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

Tesis

Digitalización basada en reconocimiento óptico de caracteres para mejorar la disponibilidad de actas de evaluación en la UGEL San Martín

Para optar el título profesional de Ingeniero de Sistemas e Informática

Presentado por

César David Paredes Torres
Aranza Luccia Marcelo Vasquez

Sustentado y aprobado el 13 de diciembre de 2024 por los siguientes jurados:

Presidente de Jurado
Ing. Dr. Jorge Damián Valverde Iparraguirre

Secretario de Jurado
Ing. Dr. Juan Orlando Riascos Armas

Vocal de Jurado
Ing. Dr. Alberto Alva Arévalo

Asesor
Ing. Mtro. Cristian Werner García Estrella

Coasesor
Ing. Lloy Pool Pinedo Tuanama

Tarapoto, Perú

2024

Universidad Nacional de San Martín
Facultad de Ingeniería de Sistema e Informática
Ciudad Universitaria - Jr. Amorarca # 315 - Morales

ACTA DE SUSTENTACIÓN
PARA OPTAR EL TÍTULO DE INGENIERO DE SISTEMAS E INFORMÁTICA

Resolución Nº 36-2024-UNSM/FISI-D (10.12.2024)

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA – ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

A las 10:00 horas del día Viernes, 13 de diciembre del año 2024, se inició el acto público de sustentación de la tesis titulada: DIGITALIZACIÓN BASADA EN RECONOCIMIENTO ÓPTICO DE CARACTERES PARA MEJORAR LA DISPONIBILIDAD DE ACTAS DE EVALUACIÓN EN LA UGEL SAN MARTÍN; presentado por CÉSAR DAVID PAREDES TORRES y ARANZA LUCCIA MARCELO VASQUEZ, con el Asesor: Ing. Mtro. Cristian Werner García Estrella y Co asesor: Ing. Lloy Pool Pinedo Tuanama.

Instalado los miembros de jurado calificador conformado por:

Presidente	:	Ing. Dr. JORGE DAMIAN VALVERDE IPARRAGUIRRE
Secretario	:	Ing. Dr. JUAN ORLANDO RIASCOS ARMAS
Vocal	:	Ing. Dr. ALBERTO ALVA ARÉVALO

El presidente del jurado dirigió brevemente unas palabras y a continuación el secretario dio lectura a la Resolución Nº 36-2024-UNSM/FISI-D.

Seguidamente el autor expuso el trabajo de investigación y el jurado realizó las preguntas pertinentes, respondidas por el sustentante y eventualmente por el asesor, con la venia del jurado.

Una vez terminada la ronda de preguntas el jurado procedió a deliberar para determinar la calificación final, para lo cual dispuso un receso de quince (15) minutos, con participación del asesor con voz, pero sin voto y sin la presencia del sustentante y otros participantes del acto público.

Luego de aplicar los criterios de calificación con estricta observancia del principio de objetividad y de acuerdo con los puntajes en escala vigesimal (de 0 a 20), según el Anexo 4.2. del RG-CTI, la nota de sustentación otorgada resultante del promedio aritmético de los calificativos emitidos por cada uno de los miembros del jurado fue diecisiete (17).

De acuerdo con el Artículo 40° del RG – CTI, la nota obtenida es APROBADO y correspondiente a la calificación de MUY BUENO; leído este resultado en presencia de todos los participantes del acto de sustentación, el secretario dio lectura a las observaciones subsanables al informe final que el autor deberá corregir y alcanzar al jurado en un plazo máximo de treinta (30) días calendario.

Firman los integrantes del jurado calificador, asesor y el autor de la tesis en señal de conformidad, dando por concluido el acto a las 10:15 horas, el mismo día 13 de diciembre del 2024.

Ing. Dr. JORGE DAMIAN VALVERDE IPARRAGUIRRE
Presidente

Ing. Dr. JUAN ORLANDO RIASCOS ARMAS
Secretario

Ing. Dr. ALBERTO ALVA ARÉVALO
Vocal

Ing. Mtro. CRISTIAN WERNER GARCÍA ESTRELLA
Asesor

Ing. LLOY POOL PINEDO TUANAMA
Co Asesor

CÉSAR DAVID PAREDES TORRES
Autor

ARANZA LUCCIA MARCELO VASQUEZ
Autor

Declaratoria de autenticidad

César David Paredes Torres, con DNI N° 70194642, y Aranza Luccia Marcelo Vasquez, con DNI N° 71597049, egresados de la Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional de San Martín, autores de la tesis titulada: Digitalización basada en reconocimiento óptico de caracteres para mejorar la disponibilidad de actas de evaluación en la UGEL San Martín.

Declaramos bajo juramento que:

1. La tesis presentada es de nuestra autoría.
2. La redacción fue realizada respetando las citas y referencia de las fuentes bibliográficas consultadas, siguiendo las normas APA actuales.
3. Toda información que contiene la tesis no ha sido plagiada.
4. Los datos presentados en los resultados son reales, no han sido alterados ni copiados, por tanto, la información de esta investigación debe considerarse como aporte a la realidad investigada.
5. Por lo antes mencionado, asumimos bajo responsabilidad las consecuencias que deriven de mi accionar, sometiéndome a las leyes de nuestro país y normas vigentes de la Universidad Nacional de San Martín.

Tarapoto, 13 de diciembre del 2024

César David Paredes Torres
70194642
Autor

Aranza Luccia Marcelo Vasquez
71597049
Autor

Ficha de identificación

Título del proyecto Digitalización basada en reconocimiento óptico de caracteres para mejorar la disponibilidad de actas de evaluación en la UGEL San Martín	**Área de investigación:** Ciencias Naturales **Línea de investigación:** Ciencias de la Computación **Sublínea de investigación:** Inteligencia Artificial y Recuperación de la Información **Tipo de investigación:** Básica ☐, Aplicada ☒, Desarrollo experimental ☐

Autores: César David Paredes Torres Aranza Luccia Marcelo Vasquez	Facultad de Ingeniería de Sistemas e Informática Escuela Profesional de Ingeniería de Sistemas e Informática https://orcid.org/0000-0002-2047-1454 https://orcid.org/0000-0002-8768-8113

Asesor: Ing. Mtro. Cristian Werner García Estrella	**Dependencia local de soporte:** Facultad de Ingeniería de Sistemas e Informática Escuela Profesional de Ingeniería de Sistemas e Informática Unidad o Laboratorio Ingeniería de Sistemas e Informática https://orcid.org/0000-0002-5687-8694

Coasesor: Ing. Lloy Pool Pinedo Tuanama	**Contraparte científica:** Ingeniero de Sistemas e Informática Grupo de Investigación IA Investigador Renacyt https://orcid.org/0000-0002-5569-8739

Dedicatoria

A mis padres, Luis Marcelo y Giovanna Vásquez, les agradezco de todo corazón por su amor incondicional y su apoyo constante. Ustedes han sido mi inspiración y mi mayor fortaleza, guiándome con sus enseñanzas y valores en cada paso de este camino. A mis hermanos, Nietch y Naiara, gracias por ser mis cómplices y motivadores; su apoyo ha hecho de esta experiencia algo aún más especial.

A mis abuelos, por transmitir sabiduría y amor, y a mi tía Brenda, por su constante aliento y acompañamiento en cada momento. Y, por supuesto, a mí mismo, por el esfuerzo y la dedicación que he invertido en este proyecto. Este logro es el resultado del trabajo en equipo y de un amor que trasciende, y me siento agradecido por cada uno de ustedes en mi vida.

Aranza Luccia Marcelo Vasquez

Dedico este trabajo a mi padre, mi mentor y tutor de vida, por su apoyo incondicional, por siempre confiar en mí, por su ejemplo que me lleva por el buen camino; también a mis amigos, que siempre están en las buenas y las malas.

César David Paredes Torres

Agradecimientos

A los funcionarios de la UGEL San Martín, por su colaboración en las diferentes actividades del proyecto.

Asimismo, al Ing. Cristian Werner García Estrella e Ing. Lloy Pool Pinedo Tuanama, por la paciencia y asesoría brindada durante la ejecución del estudio.

Los autores

Índice general

Índice de tablas

Índice de figuras

RESUMEN

Digitalización basada en reconocimiento óptico de caracteres para mejorar la disponibilidad de actas de evaluación en la UGEL San Martín

La presente investigación tuvo como objetivo general determinar en qué medida la digitalización basada en reconocimiento óptico de caracteres mejora la disponibilidad de actas de evaluación en la UGEL San Martín. El estudio fue de tipo aplicada, enfoque cuantitativo y un diseño pre-experimental, donde se evaluó la accesibilidad y disponibilidad de las actas antes y después de la digitalización. Para ello, se utilizó el software OCR Tesseract para convertir las actas físicas en documentos digitales almacenados en una base de datos SQL, lo que facilitó su búsqueda y manipulación. En cuanto a los resultados, antes de la digitalización, el tiempo promedio para localizar las actas era de 903,53 seg, y el 55% de los funcionarios consideraba la disponibilidad de las actas como "regular". Tras la implementación de OCR, el tiempo de localización se redujo a 118,07 seg, evidenciando una mejora notable en la optimización del proceso de visado de certificados de estudios. Además, la percepción sobre la disponibilidad de las actas cambió, con un 82% de los funcionarios calificándola como "alta". Se concluye que la digitalización basada en OCR mejoró significativamente la accesibilidad y disponibilidad de las actas de evaluación en la UGEL San Martín, debido a que el p-valor (0,000 y 0,035, respectivamente) fue menor al nivel de significancia (0,05). Por lo tanto, este estudio resalta la relevancia de implementar tecnologías OCR en la gestión administrativa para mejorar la eficiencia y accesibilidad documental.

Palabras clave: Accesibilidad documental, Procesamiento de imágenes, Tesseract, SQL

ABSTRACT

Digitization based on optical character recognition to improve the availability of evaluation records at UGEL San Martín

The general objective of this research was to determine to what extent digitization based on optical character recognition improves the availability of evaluation reports in the UGEL San Martin. The study was applied, with a quantitative approach and a pre-experimental design, where the accessibility and availability of the minutes before and after digitization was evaluated. For this purpose, the OCR Tesseract software was used to convert the physical minutes into digital documents stored in a SQL database, which facilitated their search and manipulation. In terms of results, before digitization, the average time to locate the transcripts was 903.53 sec, and 55% of the officers considered the availability of the transcripts as “regular”. After the implementation of OCR, the localization time was reduced to 118.07 sec, evidencing a notable improvement in the optimization of the transcript visa process. In addition, the perception of the availability of transcripts changed, with 82% of the officers rating it as “high”. It is concluded that OCR-based digitization significantly improved the accessibility and availability of evaluation records at UGEL San Martin, since the p-value (0.000 and 0.035, respectively) was lower than the significance level (0.05). Therefore, this study highlights the relevance of implementing OCR technologies in administrative management to improve the efficiency and accessibility of documents.

Keywords: Document accessibility, Image processing, Tesseract, SQL

CAPÍTULO I
INTRODUCCIÓN A LA INVESTIGACIÓN

Dentro del contexto académico actual, la gestión documentaria adquiere un papel importante al garantizar no sólo un funcionamiento fluido, sino también una estructuración óptima de los procesos educativos (Kruchinin & Bagrova, 2019). En este marco, las unidades de gestión educativa emergen como entidades fundamentales al estar encargadas de supervisar y gestionar todas las actividades educativas en una jurisdicción determinada; sin embargo, la forma habitual de manejar los documentos en formato impreso y a menudo almacenadas en archivos físicos, plantea obstáculos para su disponibilidad (Makhnevich, 2023).

Este desafío no ocurre únicamente en entidades educativas, sino que se manifiesta en varias organizaciones, tanto públicas como privadas, que se ocupan de procesos documentales. Tal es el caso de la Registraduría de la Municipalidad del Cantón Chone (Ecuador) que según Cañarte-Aizprua et al. (2022) genera abundante cantidad de información que atraviesa diversos requisitos y etapas antes de ser aprobada o rechazada; no obstante, la modalidad de mantener estos registros de forma impresa dificulta el acceso a los mismos y, además, conlleva un riesgo al conservar la información en su forma original en papel por el entorno húmedo del área y la presencia de roedores (Muñoz Soro & Nogueras Iso, 2014).

Por su parte, Martínez Cano (2021) enfatiza la alta cantidad de documentación generada y recibida en la Universidad de las Ciencias Informáticas (Cuba), que tienen relevancia tanto en el ámbito administrativo interno como en su proyección hacia el exterior. Esta extensión se fundamenta en que la gestión administrativa universitaria abarca más que la simple manipulación física de documentos, involucrando la administración de información con consecuencias directas en la salvaguardia de los derechos de los usuarios y en la habilidad de la institución para estar sujeta a evaluación y supervisión. A pesar de esta importancia, se identifica una amenaza para la disponibilidad de los registros, ya que están expuestos al deterioro influenciado por su prolongada retención en formato físico. Asimismo, resalta la fragilidad de estos archivos frente a factores perjudiciales como los ácaros, que pueden causar daños materiales y, por consiguiente, generar riesgos para la salud de aquellos que interactúan con los documentos.

En otro ámbito, Carrillo Morales & Chávez Chiquillan (2022) manifiestan que en la Municipalidad Provincial de Chincheros (Apurímac-Perú) existe una amplia gama de

documentos que son generados, gestionados y almacenados internamente, que resultan en la acumulación de una gran cantidad de documentos físicos y plantea desafíos en términos de almacenamiento, así como en el monitoreo y control de la documentación en uso; sin embargo, no se han planteado o implementado estrategias que permitan gestionar los documentos para garantizar su disponibilidad y proporcionar información pertinente a los usuarios.

Ahora bien, en esta investigación se ha identificado que en la Unidad de Gestión Educativa Local (UGEL) de San Martín, particularmente en el área de Secretaría General, se han recepcionado actas de evaluación de diversas instituciones educativas de nivel básico-regular hasta el año 2017, las cuales se han archivado en formato físico, organizados en libros y posteriormente almacenado en estanterías. Esta situación ha generado complicaciones en la disponibilidad de las actas de evaluación para efectuar las solicitudes de visado de certificados de estudios, dado que el gran volumen de actas almacenadas dificulta su acceso, lo que a su vez provoca un aumento en el tiempo requerido para su localización afectando la eficiencia administrativa.

Este proceso convencional conlleva una carga en términos de recursos humanos para acceder a la documentación, ya que el personal administrativo debe destinar tiempo a la recuperación manual de los registros (Buctuanon et al., 2021). Además, la acumulación constante de actas de evaluación a lo largo del tiempo ha reducido la capacidad de almacenamiento, conduciendo a problemas de organización en los registros existentes, al mismo tiempo que las actas pierden nitidez (Azzam et al., 2023). Adicionalmente, se identifica la ausencia de medidas de resguardo frente a posibles riesgos, como incendios, pues las actas de evaluación están expuestas a pérdidas (Martínez Cano, 2021).

Por lo tanto, la capacidad de la institución para atender las necesidades de los solicitantes de certificados de estudios presenta deficiencias derivadas de la baja disponibilidad de las actas de evaluación que prolonga los plazos necesarios para completar el proceso de obtención del visado. Esta situación genera insatisfacción y obstaculiza oportunidades educativas y laborales que dependen de la presentación de documentos oficiales (Carrillo Morales & Chávez Chiquillan, 2022).

Frente a esta problemática, se planteó utilizar técnicas de digitalización basada en reconocimiento óptico de caracteres tal como aplicaron Promin & Suriyachai (2019), Jayoma et al. (2020) y Viet Anh et al. (2022), de esta forma se buscó mejorar la disponibilidad de las actas de evaluación en la UGEL San Martín para garantizar su

almacenamiento, seguridad y accesibilidad oportuna. En este marco, se formuló el problema de investigación: ¿En qué medida la digitalización basada en reconocimiento óptico de caracteres mejora la disponibilidad de actas de evaluación en la UGEL San Martín?; siendo la hipótesis de investigación: La digitalización basada en reconocimiento óptico de caracteres mejora significativamente la disponibilidad de actas de evaluación en la UGEL San Martín.

En cuanto al objetivo general se planteó: Determinar en qué medida la digitalización basada en reconocimiento óptico de caracteres mejora la disponibilidad de actas de evaluación en la UGEL San Martín; y los objetivos específicos fueron: 1) Evaluar la disponibilidad de actas de evaluación en la UGEL San Martín, 2) Digitalizar las actas de evaluación en la UGEL San Martín mediante técnicas de reconocimiento óptico de caracteres y 3) Medir la disponibilidad de actas de evaluación en la UGEL San Martín después de su digitalización basada en reconocimiento óptico de caracteres.

CAPÍTULO II
MARCO TEÓRICO

2.1. Antecedentes de la investigación

En Tailandia, Chumwatana & Rattana-umnuaychai (2021) aplicaron la técnica de OCR para reconocer el texto de un documento físico a formato digital, intentando extraer todo el texto de las fotocopias a una estructura de base de datos. Los estudios experimentales demostraron que la técnica propuesta hace que los documentos digitalizados sean completamente buscables y editables con un rendimiento de precisión promedio de alrededor del 75,38% para la extracción de atributos y 66,92% para extraer valores de documentos impresos. Mencionan que esta tecnología proporciona importantes beneficios a todas las organizaciones aportando a buscar fácilmente información muy útil en todo el documento y, además, reduce la cantidad de papel que ocupa espacio en oficinas.

En la India, Sai Rakesh Kamisetty et al. (2022) identificaron un sistema de digitalización de facturas que incluye principalmente la extracción de datos, pero no aplica técnicas de preprocesamiento de imágenes, por lo cual propusieron la conversión a blanco y negro, la inversión de colores, la eliminación de ruido, la escala de grises, el realce de fuentes y la optimización de la imagen. Una vez mejorada la calidad, llevaron a cabo procedimientos adicionales utilizando OpenCV. En etapas posteriores, utilizaron tres sistemas OCR diferentes: Keras OCR, Easy OCR y Tesseract OCR, siendo este último el que arrojó resultados más precisos. Después de completar las fases iniciales, eliminaron símbolos no deseados (/t, /n) para obtener un texto escalable como resultado final. Por último, lograron desarrollar un método único que demostró una alta precisión al generar formatos JSON y CSV.

En Ecuador, Ayala-Churo et al. (2022) propusieron un modelo de detección diseñado para extraer y contextualizar campos presentes en documentos de identificación. Con el objetivo de llevar a cabo el entrenamiento de los modelos de detección, procedieron a construir un conjunto de datos que incluyera imágenes reales de documentos previamente etiquetadas por expertos en el área. La tarea de etiquetado de imágenes fue llevada a cabo mediante el uso de plataformas que permitieron la identificación y contextualización de los campos a través de cuadros delimitadores. Durante la fase de entrenamiento del modelo mediante el aprendizaje del algoritmo YOLOv5 y la Liberia Tesseract-OCR, lograron obtener resultados que corresponden a una precisión de 93.7%, recall del 99.2% y un mAP (Precisión Media de Entrenamiento) de 100%.

En Lima-Perú, Carrillo Fuertes (2020) llevó a cabo la creación de un modelo algorítmico con la finalidad de extraer información de individuos a partir de las imágenes de sus DNI electrónicos. Para lograr este propósito, emplearon algoritmos de procesamiento de imagen que permitieron identificar los datos personales presentes en el DNI. En un primer paso, estos algoritmos reconocieron los datos de la persona en la imagen del DNI; posteriormente, descompusieron cada conjunto de datos en palabras, y finalmente, fragmentaron cada palabra en letras individuales. Las imágenes que contenían letras fueron sometidas a una clasificación por parte de un modelo, donde determinó que letra se trataba. Asimismo, evaluaron tres modelos para la clasificación de las letras: Adaboost, basado en árboles de decisiones, y YOLO (v3 tiny), una arquitectura neuronal inspirada en GoogLeNet. Tras analizar una muestra de 17 DNI electrónicos, se obtuvo un acierto del 87% en la detección correcta de letras utilizando Adaboost, y un 98% de precisión con el modelo YOLO. Concluyeron que los modelos Adaboost y YOLO tienen la capacidad de mejorar la extracción de información de individuos a partir de las imágenes de sus DNI electrónicos.

En Moquegua, Apaza Flores (2022) realizó la automatización de los procesos de digitalización de datos utilizando estrategias basadas en Inteligencia Artificial, específicamente en técnicas de Machine Learning, en documentos aduaneros como Facturas Comerciales y Documentos de Embarque, los cuales son esenciales para emitir una Declaración Aduanera de Mercancías (DAM) a la SUNAT. Se aprovecharon servicios en la nube, como Computer Vision y Form Recognizer de la plataforma Microsoft Azure para crear modelos personalizados de reconocimiento. La implementación tuvo un impacto significativo en los tiempos de respuesta del área de Liquidación, logrando una reducción de hasta un 75%. Esto permitió que se pudieran manejar más despachos en el mismo intervalo de tiempo que anteriormente se destinaba a uno solo. Además, logró una disminución en los errores de digitalización, que previamente daban lugar a procesos de rectificación y, en algunos casos, a la imposición de multas.

En el contexto local, no se han documentado investigaciones que aborden esta misma perspectiva. Es importante destacar que este enfoque resulta sumamente relevante al abordar la problemática de la disponibilidad de actas de evaluación mediante la implementación de la tecnología OCR en la UGEL San Martín, lo cual cobra aún más peso al considerar los antecedentes existentes que demuestran la eficacia de estos modelos en la optimización de tiempos y recursos.

2.2. Fundamentos teóricos

2.2.1. Fundamentos de la variable independiente

Digitalización

La digitalización es el proceso de convertir información en formato analógico, como documentos impresos o imágenes físicas, en un formato digital, es decir, en datos digitales que pueden ser almacenados, procesados y transmitidos por medios electrónicos y computacionales. Esta transformación permite que la información sea más accesible, manipulable y compartible en entornos electrónicos. La digitalización es una respuesta a la creciente necesidad de gestionar y manipular grandes volúmenes de información de manera eficiente, siendo especialmente importante en la era de la tecnología de la información, donde los sistemas informáticos han revolucionado la forma en que se almacenan y comparten los datos (Vrana & Singh, 2021).

Los procesos de digitalización generalmente involucran el uso de dispositivos como escáneres, cámaras digitales u otros dispositivos de captura para convertir imágenes analógicas en digitales, los cuales pueden ser almacenados en diferentes formatos, como documentos PDF, imágenes JPEG, archivos de texto o bases de datos (Verhoef et al., 2021). La digitalización tiene una serie de ventajas, entre las que se incluyen:

- Accesibilidad: Los datos digitales son fácilmente accesibles desde cualquier lugar con una conexión a internet permitiendo compartir información de manera rápida y eficiente.
- Almacenamiento: Los datos digitales ocupan menos espacio físico que los documentos en papel, facilitando el almacenamiento y la organización.
- Búsqueda y recuperación: Los datos digitales se pueden buscar y recuperar de manera más rápida y eficiente utilizando herramientas de búsqueda electrónica.
- Edición y manipulación: Los datos digitales son más fáciles de editar y manipular que los documentos impresos facilitando la corrección, actualización y personalización de la información.
- Preservación y respaldo: Los datos digitales pueden ser respaldados y preservados mediante copias de seguridad y sistemas de almacenamiento.
- Reducción de costos: La digitalización reduce los costos asociados con la impresión, el almacenamiento físico y la distribución de documentos en papel.

La digitalización también presenta desafíos, como la necesidad de garantizar la seguridad y privacidad de los datos digitales, así como la posible obsolescencia de

formatos y tecnologías en el futuro. Sin embargo, en general, la digitalización ha transformado la forma en que interactuamos con la información y ha revolucionado la gestión de datos en diversos campos, desde la administración empresarial hasta la investigación científica (Morze & Strutynska, 2021).

Reconocimiento Óptico de Caracteres (OCR)

El Reconocimiento Óptico de Caracteres (OCR, por sus siglas en inglés Optical Character Recognition) es una tecnología que permite la conversión de textos impresos o escritos a mano en imágenes en formato digital en caracteres procesables por una computadora. En otras palabras, el OCR es capaz de interpretar y traducir la información contenida en imágenes o documentos físicos en texto que puede ser editado, buscado y almacenado electrónicamente (Chaudhuri et al., 2017).

El concepto de OCR comenzó a tomar forma en la década de 1920 con los trabajos pioneros del científico y empresario Emanuel Goldberg. En 1929, Goldberg presentó una patente en Alemania para un "dispositivo para el reconocimiento de caracteres impresos", que se considera uno de los primeros intentos registrados de automatizar el proceso de lectura de caracteres impresos. Aunque la tecnología y los algoritmos de OCR han avanzado significativamente desde entonces, las contribuciones iniciales de Emanuel Goldberg sentaron las bases para la investigación y el desarrollo continuos en este campo. Su visión y trabajo sentaron las bases para la automatización del proceso de conversión de texto impreso en caracteres digitales, lo que ha revolucionado la forma en que interactuamos con documentos impresos y digitalizados en la actualidad (Berchmans & Kumar, 2014).

El proceso de OCR implica varias etapas. En primer lugar, se captura una imagen de un documento impreso, que puede ser una página de un libro, un artículo de revista, un recibo, etc. Luego, el software OCR analiza la imagen y utiliza algoritmos para identificar patrones y formas que representan caracteres individuales. A medida que identifica los caracteres, el software los convierte en texto digital, que luego se puede editar en un procesador de texto, buscar en un motor de búsqueda o almacenar en una base de datos (Cheung et al., 2001).

El OCR ha avanzado significativamente en términos de precisión y capacidad para manejar diferentes tipos de fuentes, tamaños y formatos de texto. La tecnología OCR es especialmente útil en diversas aplicaciones, como la digitalización de libros y documentos antiguos para su preservación, la conversión de documentos en papel a formatos electrónicos para facilitar la búsqueda y recuperación de información, y la automatización de procesos empresariales que involucran la extracción de datos de

documentos impresos. Además de la conversión de texto impreso, algunos sistemas OCR también son capaces de reconocer otros elementos visuales, como tablas, gráficos e imágenes. Esto hace que el OCR sea una herramienta versátil para la transformación de documentos físicos en contenido digital en una amplia gama de contextos y aplicaciones (Thorat et al., 2022).

Herramientas de OCR

Jain et al. (2021) desarrollaron una investigación comparativa en el que describen un conjunto de herramientas de OCR de acuerdo a sus finalidades y eficiencias, las cuales se resumen en la Figura 1.

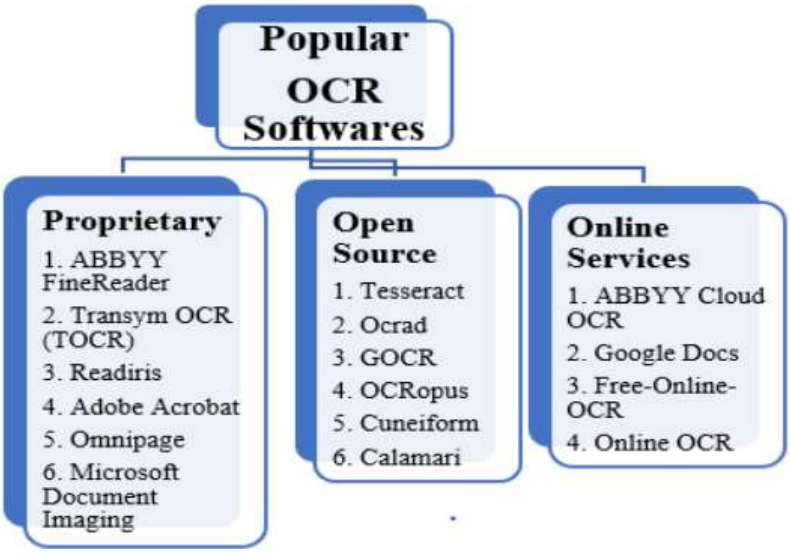

Figura 1.
Software OCR populares
Fuente: Tomado de Jain et al. (2021)

Los autores manifiestan que las herramientas de OCR tienen capacidades diferentes y que un único conjunto de herramientas de OCR puede no encajar en todos los dominios. La calidad de la imagen juega un papel importante en el reconocimiento de texto. Para imágenes de alta calidad, ABBYY es el mejor software de OCR de su clase. Para libros impresos después del siglo XIX, ABBYY ofrece mejores resultados, pero para libros impresos antiguos, OCRopus3 proporciona resultados de OCR mucho mejores. Para impresiones en inglés moderno, en la categoría de código abierto, Calamari produce mejores resultados de reconocimiento en comparación con OCRopus3 y Tesseract4.0. Pero cuando se requiere un único software que pueda realizar todo el proceso de OCR de una sola vez, Tesseract4.0 es la opción más popular entre la categoría de código abierto. Cuando se trata de un gran volumen de

imágenes, Rossetta proporciona resultados de reconocimiento más rápidos y facilita el reconocimiento natural del texto de la escena.

Para entradas distorsionadas con fondos ruidosos, Namysl & Konya (2019) presentan un nuevo conjunto de herramientas de OCR sin léxico, que proporciona resultados de reconocimiento de texto mucho mejores en comparación con los conjuntos de herramientas de OCR establecidos. Hay muchos servicios de OCR en línea gratuitos disponibles, incluidos Google Docs, Online-OCR, Free-Online-OCR, etc., que permiten a los usuarios convertir imágenes en archivos de texto sin descargar los conjuntos de herramientas de OCR en sus máquinas, pero la seguridad de los archivos puede ser una preocupación utilizando estos servicios en línea.

Modelos integrados en OCR

El OCR utiliza una variedad de algoritmos de inteligencia artificial y técnicas de procesamiento de imágenes para convertir imágenes de texto en digital. Algunos de los algoritmos y enfoques de inteligencia artificial que se aplican incluyen (Moudgil et al., 2022; Srivastava et al., 2022):

- Redes Neuronales Convolucionales (CNN): Las CNN son ampliamente utilizadas en OCR debido a su capacidad para capturar patrones visuales en imágenes. Estas redes están diseñadas para procesar datos de matriz, como imágenes, y se utilizan para reconocer patrones como formas y características de texto.
- Redes Neuronales Recurrentes (RNN): Las RNN son útiles para el reconocimiento de secuencias, lo que es esencial en OCR para reconocer secuencias de caracteres en palabras y oraciones. Las LSTM (Memoria de Corto y Largo Plazo) y las GRU (Unidades de Recurrencia Gated) son variantes de RNN que se utilizan a menudo en OCR.
- Redes Neuronales Profundas (DNN): Las DNN son arquitecturas de redes neuronales que contienen múltiples capas ocultas. Estas redes profundas son capaces de aprender representaciones más abstractas y complejas de los datos, lo que es esencial para el reconocimiento de caracteres en una variedad de fuentes y estilos.
- Aprendizaje de Transferencia: Se trata de utilizar modelos preentrenados en grandes conjuntos de datos para tareas relacionadas y luego ajustarlos para el OCR. Por ejemplo, se pueden utilizar modelos entrenados en reconocimiento de imágenes generales y luego afinarlos para el reconocimiento de caracteres.

- Modelos de Segmentación: Algunos algoritmos se centran en la segmentación de la imagen para identificar regiones que contienen caracteres. Luego, estos caracteres segmentados se pasan a modelos de reconocimiento específicos.
- Modelos de Lenguaje Natural (NLP): Para mejorar el contexto y la coherencia en la interpretación de secuencias de caracteres, se pueden aplicar técnicas de procesamiento de lenguaje natural.
- Modelos Generativos Adversarios (GAN): Los GAN se han utilizado para generar imágenes sintéticas de texto y mejorar la calidad de las imágenes de entrada antes del proceso de reconocimiento.
- Enfoques de Detección de Objetos: En el caso de documentos con formatos estructurados, los enfoques de detección de objetos también pueden aplicarse para localizar y extraer áreas de texto.

2.2.2. Fundamentos de la variable dependiente

Gestión documentaria

La gestión documentaria es el conjunto de enfoques y procedimientos diseñados para dirigir de manera efectiva la creación, captura, almacenamiento, organización, recuperación y disposición de documentos en una entidad, ya sea una empresa, una institución gubernamental o un entorno académico (Obukhov et al., 2020). Su objetivo principal es manejar la información y los documentos de manera eficiente y coherente a lo largo de su ciclo de vida, desde su creación hasta su eventual eliminación o archivo (Díaz Jiménez & Mena Mujica, 2022).

Según Mazon-Fierro et al. (2023), en un mundo donde la generación de información es cada vez más abundante, la gestión documentaria se convierte en un componente crucial para mantener el orden, garantizar la integridad y la disponibilidad de la información, y asegurar que los documentos estén accesibles para aquellos que los necesitan, al tiempo que se protege la seguridad y confidencialidad de la información sensible.

Disponibilidad documentaria

La disponibilidad documentaria es un pilar fundamental en la gestión documentaria, ya que garantiza que los documentos estén a disposición de las personas autorizadas en el momento exacto en que se requieran (Camilo Momblanc & Castro Milán, 2020). Para lograr esto, es esencial contar con sistemas de almacenamiento eficientes y estructurados que permitan la rápida recuperación de documentos. Esto no sólo

mejora la eficiencia en las operaciones, sino que también evita retrasos en la toma de decisiones basadas en información documental (Sosa del Angel et al., 2022).

Accesibilidad documentaria

La accesibilidad documentaria es un componente esencial dentro de la gestión de la información en cualquier entidad u organización. Su objetivo fundamental radica en simplificar y agilizar el proceso mediante el cual los usuarios autorizados pueden buscar, acceder y aprovechar de manera eficiente los documentos relevantes para sus labores o actividades específicas (Das et al., 2022).

Esta práctica abarca la implementación de sistemas de clasificación y categorización coherentes, que a su vez deben estar alineados con las necesidades particulares de la organización y reflejar la estructura lógica que la define. A través de una estructuración cuidadosa y bien concebida, se facilita la navegación y ubicación de la información requerida. Esto desencadena una serie de beneficios directos para la operatividad de la entidad (Abdul Hamid et al., 2023).

Cumplimiento normativo y legal

El cumplimiento normativo y legal desempeña un papel fundamental en el ámbito de la gestión documentaria, adquiriendo una importancia aún mayor en entornos sometidos a regulaciones específicas. En diversos sectores y bajo distintas jurisdicciones, se establecen requisitos particulares que dictaminan la manera en que ciertos tipos de documentos deben ser tratados y preservados. La adecuada gestión documentaria debe, por lo tanto, ajustarse a estas normativas con el fin de eludir posibles consecuencias legales y preservar la impecabilidad de los documentos en situaciones de auditoría y litigio (Díaz Jiménez & Mena Mujica, 2022).

La necesidad de una gestión documentaria rigurosa y alineada con las regulaciones se acentúa aún más por las posibles consecuencias negativas de un incumplimiento normativo. Las sanciones legales, multas y otras repercusiones financieras pueden tener un impacto significativo en las finanzas y la reputación de una empresa. Además, la integridad de los documentos adquiere un valor insustituible en situaciones de auditoría, donde la capacidad de proporcionar registros precisos y completos puede marcar la diferencia entre la confianza y el escepticismo (Khalil & Fakhratov, 2023).

Ciclo de vida documental

De acuerdo con Abbasova (2020), el ciclo de vida documental es un concepto central en la gestión documentaria que reconoce que los documentos evolucionan a lo largo

del tiempo. Comienza con la creación de un documento, cuando la necesidad o la idea impulsan su origen, seguida de su captura y registro en el sistema de gestión, donde adquiere oficialidad. Durante su vigencia, el documento se convierte en un actor activo, utilizado para soportar procesos operativos y facilitar la toma de decisiones fundamentadas en información concreta y verificable. Este período intermedio no sólo refleja la utilidad del documento, sino que también puede generar versiones actualizadas que enriquezcan su contenido original.

Luego, llega el momento crucial de evaluar su valor y relevancia en el contexto en el que se desenvuelve. Esta evaluación no sólo considera su importancia inmediata, sino también su potencial para contribuir a la historia y al cumplimiento legal. Al final de su ciclo, los documentos pueden ser eliminados de acuerdo con políticas de retención y disposición, liberando espacio para nuevos registros sin valor (Mazon-Fierro et al., 2023).

2.2.3. Definición de términos básicos

Acta de evaluación: Documento para el seguimiento y registro de evaluaciones que sirve para mantener un registro preciso del rendimiento y los logros en diversos contextos (Golesworthy et al., 2022).

Algoritmo: Conjunto de instrucciones diseñado para resolver un problema o realizar una tarea en computación o matemáticas (Millán Naveas & Vargas Guzmán, 2020).

Certificado de estudio: Documento oficial que certifica la finalización de un programa educativo y detalla las materias cursadas y las calificaciones obtenidas (Chavez & Salinas, 2021).

Deep learning: Subcampo de la inteligencia artificial que utiliza redes neuronales profundas para aprender y representar patrones complejos en datos (Nakaura et al., 2020).

Inteligencia artificial: Simulación de procesos cognitivos humanos por sistemas informáticos, permitiéndoles realizar tareas que normalmente requieren inteligencia humana (Mejías et al., 2022).

Machine learning: Enfoque de la inteligencia artificial donde las computadoras aprenden automáticamente y mejoran su rendimiento en una tarea específica a través de la experiencia y datos (Greener et al., 2022).

Trámite documentario: Proceso organizado para recibir, procesar y archivar documentos en una institución u organización, asegurando su gestión eficiente y seguimiento adecuado (Salas-Tanchiva, 2022).

Transformación digital: Adopción estratégica de tecnologías digitales para cambiar fundamentalmente procesos, modelos de negocio y experiencias, buscando mejorar la eficiencia y el valor (García Peñalvo, 2021).

Visión artificial: Campo de la inteligencia artificial que capacita a las computadoras para interpretar y comprender el mundo visual, emulando la capacidad humana de procesar imágenes y videos (Sucari León et al., 2020).

CAPÍTULO III
MATERIALES Y MÉTODOS

3.1. Ámbito y condiciones de la investigación

3.1.1 Contexto de la investigación

La investigación se desarrolló en la UGEL de San Martín, una entidad peruana encargada de supervisar y administrar las actividades educativas en la provincia de San Martín. La UGEL es responsable de implementar las políticas educativas del Ministerio de Educación en su jurisdicción, supervisando las instituciones educativas y brindando apoyo a los docentes y estudiantes. Su función principal es coordinar y promover la mejora de la calidad de la educación en la provincia, así como en garantizar que los estándares educativos se cumplieran en todas las instituciones.

3.1.2 Periodo de ejecución

El estudio se ejecutó en un periodo de ocho meses, desde enero hasta agosto de 2024.

3.1.3 Autorizaciones y permisos

Para acceder a las actas de evaluación y llevar a cabo su manipulación, se solicitó la autorización al director de la UGEL San Martín para proceder con la intervención (Anexo 4). Además, dado que se aplicó una encuesta como parte de la investigación, se proporcionó un consentimiento informado para asegurar una participación consciente y voluntaria.

3.1.4 Control ambiental y protocolos de bioseguridad

Las actas de evaluación, al ser documentos físicos, estuvieron vulnerables a factores ambientales como la humedad y la presencia de plagas, deteriorando su calidad y legibilidad. Para minimizar estos riesgos en la salud de los investigadores, se utilizó guantes y mascarillas, asegurando así la protección de los documentos y la prevención de cualquier potencial afectación a la salud.

3.1.5 Aplicación de principios éticos internacionales

En esta investigación se hizo hincapié en dos principios éticos esenciales, por un lado, la confidencialidad fue un pilar fundamental, ya que las actas de evaluación contenían información sensible y personal de los estudiantes. Se garantizó que cualquier dato

recopilado o digitalizado se manejara con el más alto grado de confidencialidad, lo que implicó la adopción de medidas para proteger la privacidad de los individuos, evitando cualquier divulgación no autorizada de información personal. Los datos fueron tratados de manera anónima y solo se utilizaron con fines de investigación.

Por otra parte, el principio de beneficencia se refirió a la responsabilidad de los investigadores de maximizar los beneficios y minimizar los riesgos para los participantes y la institución. En este contexto, la institución UGEL San Martín se benefició de la investigación mediante la implementación de un sistema de digitalización que mejoró la disponibilidad de las actas de evaluación, generando un impacto positivo en la eficiencia de la gestión administrativa, agilizando los procesos y garantizando una mejor respuesta a las solicitudes de certificados de estudios.

3.2. Sistema de variables

3.2.1 Variables principales

Tabla 1.
Descripción de variables por objetivo general

Objetivo general: Determinar en qué medida la digitalización basada en reconocimiento óptico de caracteres mejora la disponibilidad de actas de evaluación en la UGEL San Martín.			
Variable abstracta	**Variable concreta**	**Medio de registro**	**Unidad de medida**
Independiente Digitalización basada en reconocimiento óptico de caracteres	Tiempo de digitalización Tasa de error de digitalización Tasa de precisión	Ficha de registro	Cuantitativo-Discreto Cuantitativo-Continuo
Dependiente Disponibilidad de actas de evaluación	**Accesibilidad:** Tiempo de localización	Ficha de registro	Cuantitativo-Discreto
	Disponibilidad: Búsqueda de actas Organización de actas Protección de actas ante daños Respaldo de actas Conservación de actas Almacenamiento de actas Eficiencia de atención	Encuesta	Cualitativo-Ordinal

3.2.2 Variables secundarias

No corresponde.

3.3 Procedimientos de la investigación

a) Tipo y nivel de investigación

La investigación fue de tipo aplicada, ya que abordó de manera concreta y práctica el problema de la disponibilidad de actas de evaluación en la UGEL San Martín. Con este propósito, se empleó un enfoque cuantitativo, dado que se recolectaron y analizaron datos numéricos para evaluar el impacto de la digitalización. En este contexto, la investigación se situó en un nivel explicativo, con el objetivo de comprender las causas subyacentes de los cambios observados, más allá de una simple descripción de los fenómenos. Se exploraron las relaciones causales entre la implementación de la solución de digitalización y la mejora en la disponibilidad de las actas de evaluación.

b) Población y muestra

La población se conformó por dos grupos de análisis: i) las solicitudes de certificados de estudios, que representaron el proceso para medir la dimensión de accesibilidad a las actas de evaluación en la UGEL San Martín; y ii) los funcionarios de la institución, quienes aportaron su percepción para evaluar la dimensión de disponibilidad de las actas de evaluación.

Dado que no se disponía de la cantidad precisa de solicitudes de certificados de estudios durante el período de intervención, la muestra se seleccionó por conveniencia (no probabilístico), tomando en consideración 30 solicitudes. Asimismo, la selección de los participantes se limitó a aquellos directamente involucrados en la gestión documentaria para la emisión de certificados de estudios. Este grupo de trabajadores estuvo compuesto por un total de 11 individuos, distribuidos entre 4 en el área de Secretaría, 2 en Mesa de Partes, 2 en Dirección y 3 en el departamento de Informática.

c) Diseño experimental

El diseño experimental fue de tipo pre-experimental, ya que se recopilaron mediciones de la variable dependiente antes y después de implementar la variable independiente (Pimienta & de la Orden, 2017). De esta manera, el esquema siguió el siguiente patrón:

$$G: O1 - - - - X - - - - O2$$

En donde:

G: Unidad de análisis

O1: Pretest de la disponibilidad de acta de evaluación

X: Digitalización de las actas de evaluación mediante técnicas de reconocimiento ópticos de caracteres

O2: Postest de la disponibilidad de acta de evaluación

d) Procedimientos de la investigación

3.3.1 Objetivo específico 1

Para iniciar con la evaluación de la disponibilidad de las actas de evaluación en su estado actual (Figura 2), se midieron las dimensiones de accesibilidad y disponibilidad. Para el primer caso, se empleó como técnica la observación directa y el instrumento fue una ficha de registro (Anexo 2A). En el segundo caso, la técnica fue la encuesta y como instrumento se utilizó un cuestionario *ad hoc* (Anexo 2B), es decir, elaborado específicamente para propósitos del presente estudio.

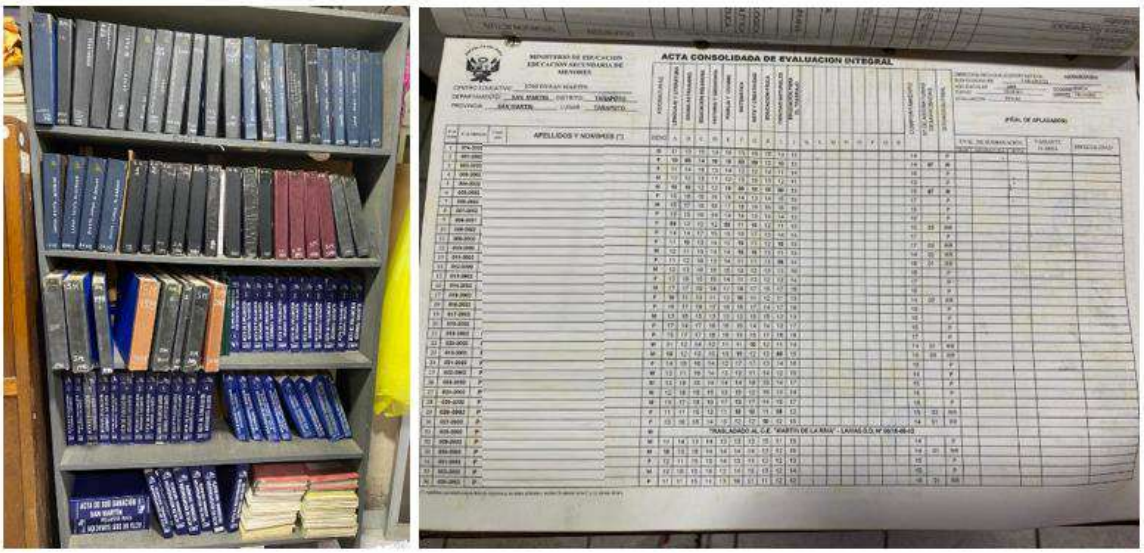

Figura 2.
Estado actual de las actas de evaluación en la UGEL San Martín

Los datos recopilados fueron tabulados en Excel 2019, luego importados desde el programa estadístico SPSS versión 26. En el caso de los datos provenientes de la ficha de registro se identificó la media y la desviación estándar; para el caso de los datos provenientes del cuestionario se analizaron las frecuencias.

3.3.2 Objetivo específico 2

Para digitalizar las actas de evaluación de la UGEL San Martín mediante técnicas de OCR, se siguió un enfoque metodológico estructurado. Primero, se recolectaron las actas físicas y se escanearon en alta resolución para obtener imágenes digitales claras. Estas imágenes se preprocesaron mediante técnicas como la conversión a escala de grises, binarización y corrección de perspectiva para mejorar la calidad y legibilidad. Luego, se segmentaron las regiones de interés que contienen texto y se aplicó el software OCR, como Tesseract, para convertir el texto en formato digital

editable. Este proceso incluyó el entrenamiento del modelo OCR con ejemplos específicos de las actas para aumentar la precisión del reconocimiento.

Posteriormente, el texto digitalizado se sometió a un postprocesamiento para corregir errores y dar formato al texto según el diseño original de las actas. Las actas digitalizadas se almacenarán en una base de datos SQL, asegurando su seguridad y realizando copias de seguridad periódicas. Finalmente, se validó la precisión del proceso comparando una muestra de las actas digitalizadas con los documentos originales y se recogió retroalimentación para realizar mejoras continuas.

3.3.3 Objetivo específico 3

En primer lugar, se llevó a cabo un análisis descriptivo de los datos mediante el uso de estadística descriptiva, esto permitió realizar comparaciones entre los resultados antes y después de la intervención. En segundo lugar, se aplicó estadística inferencial para abordar la hipótesis de investigación y así responde el objetivo general establecido.

En el caso de la dimensión de accesibilidad, considerando que la escala de respuesta fue cuantitativa-discreta, se procedió a evaluar la normalidad de los datos. Posteriormente, se seleccionó la prueba estadística prueba T Student ya que la distribución fue normal. Por otro lado, en lo que respecta a la dimensión de disponibilidad, que involucra una escala cualitativa-ordinal aplicada a muestras emparejadas, se utilizó la prueba de Wilcoxon para datos categóricos, tal como lo describen Juárez García et al. (2002).

CAPÍTULO IV
RESULTADOS Y DISCUSIÓN

4.1 Resultado específico 1: Evaluar la disponibilidad de actas de evaluación en la UGEL San Martín.

De acuerdo con la Tabla 2, en el pretest se identificó que la accesibilidad a las actas de evaluación en la UGEL San Martín, medida a través del tiempo de localización manual por parte de los funcionarios, presentó un promedio de 903,53 seg con una desviación estándar de 161,43 seg. Asimismo, el tiempo mínimo registrado para localizar las actas fue de 602 seg, y el máximo, de 1167 seg, es decir, se requiere aproximadamente 15 minutos para localizar un acta de evaluación y proceder con su visado.

Tabla 2.
Pretest de la accesibilidad a las actas de evaluación

Indicador	N	Mínimo	Máximo	Media	Desv. Estándar
Tiempo de localización de actas de evaluación	30	602	1167	903,53	161,43

En cuanto a la disponibilidad de las actas de evaluación en la UGEL San Martín, medida a través de una encuesta aplicada a los funcionarios involucrados en el proceso de visado de actas, se encontró que la mayoría (55%) considera la disponibilidad de las actas como "regular" (Figura 3), lo cual indica que aspectos como la búsqueda, organización, respaldo y protección de las actas, entre otros indicadores, presentan debilidades que requieren mejoras.

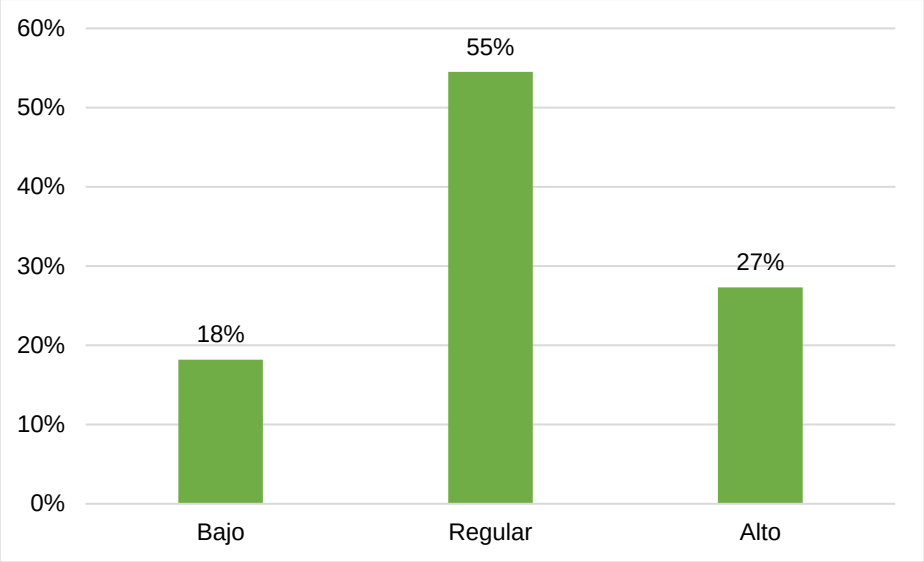

Figura 3.
Pretest de la disponibilidad de las actas de evaluación

Estos resultados revelan que la disponibilidad de las actas de evaluación en la UGEL San Martín es insuficiente debido al método tradicional de almacenamiento físico y a la organización manual que según Buctuanon et al. (2021) conlleva a una sobrecarga en términos de recursos humanos para acceder a la documentación, ya que el personal administrativo debe destinar tiempo a la recuperación manual de los registros. Además, de acuerdo a Azzam et al. (2023), la acumulación constante de actas de evaluación a lo largo del tiempo reduce la capacidad de almacenamiento, conduciendo a problemas de organización en los registros existentes, al mismo tiempo que las actas pierden nitidez.

4.2 Resultado específico 2: Digitalizar las actas de evaluación en la UGEL San Martín mediante técnicas de reconocimiento óptico de caracteres.

4.2.1. Conversión de PDF a JPG

Para comenzar con la digitalización de actas de evaluación, realizamos la conversión de los documentos de formato PDF a formato JPG. Esto nos permitió preparar los documentos para el procesamiento de imágenes y el OCR, ya que las imágenes en formato JPG son las adecuadas para las técnicas de procesamiento de imágenes que se utilizarán en pasos posteriores.

Cuando un usuario sube un archivo PDF a través de la interfaz web, la función **read_document()** se encarga de recibir el archivo y guardarlo en una ubicación específica en el servidor. Este paso asegura que el archivo PDF esté disponible para

su procesamiento posterior. La función maneja tanto las solicitudes GET como POST, permitiendo al usuario cargar el archivo PDF y recibir una respuesta con la lista de imágenes generadas en la misma ruta.

Una vez que el archivo PDF se ha guardado en el servidor, procedemos a la conversión de cada página del PDF en una imagen JPG. Esta conversión se realiza utilizando una función específica, **convert_pdf_to_jpg()** de la librería **pdf2jpg**, que transforma cada página del documento en una imagen de alta calidad.

Después de la conversión, la función **read_document()** lista todas las imágenes JPG generadas y las devuelve al usuario en forma de una respuesta JSON. Esta lista de imágenes sirve como índice con el que se procesarán individualmente para extraer la información relevante. Al proporcionar una lista de imágenes, se asegura que el usuario pueda verificar y acceder a cada página del documento convertido.

```
    def read_document(request):
        if request.method == 'GET':
            return render(request, 'read_document.html')
        elif request.method == 'POST':
            uploaded_file = request.FILES['pdf_file']
            file_path = f'{STATIC_ROOT}/pdf/{uploaded_file.name}'
            # Guardar archivo
            with open(file_path, 'wb+') as destination:
                for chunk in uploaded_file.chunks():
                    destination.write(chunk)
            # Convertir PDF a JPG
                                                convert_pdf_to_jpg(file_path,
output_folder=f'{STATIC_ROOT}/ocr_jpg')
            lista_imagenes = os.listdir(f'{STATIC_ROOT}/ocr_jpg')
            return JsonResponse({'lista_imagenes': lista_imagenes})
```

4.2.2. Importación de Librerías y módulos

Empezamos mediante la importación de librerías y módulos esenciales. Estos recursos, proporcionan la base para todas las operaciones de procesamiento de imágenes y OCR que se llevarán a cabo durante el procesamiento principal. La primera línea de importación, corresponde a la biblioteca **os**, la cual nos permite interactuar con el sistema operativo, facilitando operaciones como listar archivos en un directorio y verificar la existencia de archivos específicos.

A continuación, importamos la biblioteca **cv2**, que forma parte de OpenCV, el cual se encargará de la mayoría de tareas correspondientes al procesamiento de imágenes. OpenCV es una biblioteca de código abierto ampliamente utilizada en visión por

computadora, que proporciona una amplia gama de funciones para manipular y transformar imágenes. Durante este proyecto, utilizamos cv2 para cargar imágenes, convertirlas a diferentes formatos de color, aplicar filtros, detectar bordes y realizar transformaciones geométricas, entre otras operaciones.

Importamos también la biblioteca **numpy** que junto a OpenCV, se complementan proporcionando capacidades avanzadas de computación numérica, permitiendo realizar operaciones eficientes en matrices y arreglos que representan las imágenes.

Además, importamos también la librería **imutils**, siendo esta una colección de funciones de conveniencia que simplifican tareas comunes de manipulación de imágenes, como redimensionar, rotar y traducir imágenes. Esta biblioteca facilita la implementación de operaciones básicas de procesamiento de imágenes de manera más eficiente y con menos código. Por otro lado, **pytesseract** actúa como un enlace de Python para **Tesseract**, el motor de OCR de código abierto utilizado durante el presente trabajo. Esta herramienta es esencial para extraer texto de las imágenes procesadas, permitiendo convertir la información visual en datos digitales editables.

Finalmente, incluimos los módulos locales denominados **funciones.py** y **extracción.py**, las cuales contienen funciones personalizadas desarrolladas específicamente para este sistema. Estas funciones encapsulan la lógica necesaria para tareas específicas como la corrección de perspectiva y el procesamiento de la cabecera y el cuerpo de las fichas digitalizadas, asegurando que el flujo de trabajo sea modular y fácil de mantener.

```
import os
import cv2
import numpy as np
from .funciones import corregir_perspectiva, obtener_cabecera
from .extraccion import procesar_cabecera, procesar_cuerpo
import imutils
```

4.2.3. Funciones de Procesamiento de Imágenes

Las funciones de procesamiento de imágenes, son esenciales para preparar las imágenes de las actas de evaluación antes de aplicar técnicas de OCR. Dichas funciones permiten transformar y mejorar las imágenes para asegurar que el texto sea legible y que el OCR pueda extraer la información con precisión.

Una de las funciones clave es **corregir_perspectiva()**, que se encarga de ajustar la perspectiva de las imágenes. Esta función utiliza los puntos de referencia de la imagen para calcular el rectángulo de área mínima que encierra estos puntos y luego aplica

una rotación para alinear correctamente la imagen. Este paso es crucial para asegurar que las actas de evaluación estén correctamente orientadas y que el texto sea legible. De esta manera es posible aplicar márgenes y dimensiones estándares para cada imagen procesada de manera individual.

```
    def corregir_perspectiva(imagen, puntos):
        # Calcula los datos del rectángulo de área mínima
        centro, (ancho, alto), angulo = cv2.minAreaRect(puntos)
        if angulo == 0:
            return imagen
        if angulo < -45:
            angulo = -(90 + angulo)
        if angulo > 45:
            angulo = angulo-90
        else:
            angulo = -angulo
        # Obtiene las dimensiones de la imagen
        (h, w) = imagen.shape[:2]

        # Calcula el centro de la imagen
        centro = (w // 2, h // 2)

        # Crea la matriz de rotación
        M = cv2.getRotationMatrix2D(centro, angulo, 1.0)

        # Aplica la rotación
            imagen_rotada = cv2.warpAffine(imagen, M, (w, h),
flags=cv2.INTER_CUBIC, borderMode=cv2.BORDER_REPLICATE)
        return imagen_rotada
```

Cabe destacar el formato con el que cuentan las fichas de evaluación que se analizan por el sistema OCR. Este formato cuenta con una estructura de doble cara, la cual presenta una cabecera de datos distinta para cada cara. Por eso, es importante recalcar el funcionamiento de la función **obtener_cabecera()**, que se encarga de identificar y extraer la cabecera de la imagen. La cabecera contiene información clave como el nombre de la institución, el grado y la sección. Esta función convierte la imagen a escala de grises, aplica un umbral binario para resaltar los bordes y utiliza técnicas de detección de contornos para identificar las áreas de interés. Una vez identificada la cabecera, se extrae y se prepara para el procesamiento posterior.

Las funciones **procesar_cabecera()** y **procesar_cuerpo()** se encargan de extraer y procesar la información contenida en la cabecera y el cuerpo de las actas de evaluación, respectivamente. La primera de estas funciones, **procesar_cabecera()**

utiliza coordenadas predefinidas para extraer áreas específicas de la cabecera y aplica OCR para convertir el texto en datos digitales.

```
        def procesar_cabecera(imagen):
            coordenadas = {
                        'nombre_institucion':   (0.268463,   0.111111,   0.215633,
0.074074),
                'grado': (0.335803, 0.362433, 0.020954, 0.063492),
                'seccion': (0.457142, 0.362433, 0.026954, 0.063492),
                'turno': (0.457064, 0.438507, 0.028032, 0.066783),
                'periodo': (0.665229, 0.015873, 0.073854, 0.074074),
            }
            datos_cabecera = {}
            for key, value in coordenadas.items():
                x, y, w, h = value
                                                                celda             =
imagen[int(y*imagen.shape[0]):int((y+h)*imagen.shape[0]),
int(x*imagen.shape[1]):int((x+w)*imagen.shape[1])]

                if key == 'seccion':
                                    celda   =   cv2.threshold(cv2.cvtColor(celda,
cv2.COLOR_BGR2GRAY), 120, 255, cv2.THRESH_BINARY)[1]
                                    texto   =   pytesseract.image_to_string(celda,
config=single_character+'                                                       -c
tessedit_char_whitelist=ABCDEFGHIJKLMNOPQRSTUVWXYZ')

                elif key == 'turno':
                    celda = cv2.cvtColor(celda, cv2.COLOR_BGR2GRAY)
                                    celda   =   cv2.threshold(celda,   120,   255,
cv2.THRESH_BINARY)[1]
                                    texto   =   pytesseract.image_to_string(celda,
config=raw_character+' -c tessedit_char_whitelist=MT')
                    texto = re.sub(r'[^MT]', '', texto)
                    if len(texto) == 2:
                        texto = texto[0]
                elif key == 'grado':
                    celda = imutils.resize(celda, width=celda.shape[1]*4)
                                    celda   =   cv2.threshold(cv2.cvtColor(celda,
cv2.COLOR_BGR2GRAY), 120, 255, cv2.THRESH_BINARY)[1]
                                    texto   =   pytesseract.image_to_string(celda,
config=digit_options)
                else:
                    celda = imutils.resize(celda, width=celda.shape[1]*4)
                    _y = celda.shape[0]
                    celda = celda[int(_y*0.1): _y-int(_y*0.1), 5:]

                    if key == 'nombre_institucion':
                        celda = cv2.GaussianBlur(celda, (3, 3), 0)
```

```
                                celda = cv2.threshold(cv2.cvtColor(celda,
cv2.COLOR_BGR2GRAY), 0, 255, cv2.THRESH_BINARY+cv2.THRESH_OTSU)[1]
                        texto = pytesseract.image_to_string(celda, lang="spa",
config="--psm 7 --oem 3")
                    else:
                                celda = cv2.threshold(cv2.cvtColor(celda,
cv2.COLOR_BGR2GRAY), 120, 255, cv2.THRESH_BINARY)[1]
                        texto = pytesseract.image_to_string(celda, lang="spa")
                datos_cabecera[key] = re.sub(r'¥n', '', texto)
            return datos_cabecera
```

Por otro lado, **procesar_cuerpo()** se enfoca en extraer nombres, códigos y notas del cuerpo de la imagen, utilizando técnicas de segmentación y OCR para obtener la información relevante. Para esto, separamos el procesamiento números y posteriormente de las notas en funciones aparte, denominadas **procesar_numeros()** y **procesar_notas()** respectivamente.

La función **procesar_numeros()** se encarga de procesar imágenes binarizadas para resaltar los números y facilitar su extracción mediante OCR.

```
        def procesar_numeros(imagen_binarizada, son_notas=False):
            if not son_notas: mascara = cv2.copyMakeBorder(imagen_binarizada,
2, 2, 2, 2, cv2.BORDER_CONSTANT, value=0)
            mascara = imutils.resize(mascara, height=300)
            mascara = cv2.GaussianBlur(mascara, (3, 3), 0)
            fondo_blanco = np.full_like(mascara, 255)

                contornos_, _ = cv2.findContours(mascara, cv2.RETR_TREE,
cv2.CHAIN_APPROX_SIMPLE)

            for i, c_ in enumerate(contornos_):
              if (mascara.shape[1]//4)**2>cv2.contourArea(c_)>600:
                cv2.drawContours(fondo_blanco, [c_], -1, (0, 0, 0), 5)

            return np.copy(fondo_blanco)
```

Por otra parte **procesar_notas()** se enfoca en extraer las notas de los cursos presentes en las actas de evaluación, utilizando técnicas de procesamiento de imágenes y OCR para obtener resultados precisos.

```
        def procesar_notas(img, gris_ref):
                    cursos = ('Matemática', 'Comunicación', 'Idiomas
extranjero/originario', 'Educación por el Arte', 'Ciencias Sociales',
'Persona, Familia, y Relaciones Humanas', 'Educación Física',
                              'Educación Religiosa', 'Ciencia, Tecnología y
Ambiente', 'Educación por el Trabajo', 'Especialidad Ocupacional', 'Liderazgo
y Autoestima', 'Cantidad Desaprobados')
```

```
            notas_cursos = {}
            img = imutils.resize(img, width=img.shape[1]*4)
            gris_ref = imutils.resize(gris_ref, width=gris_ref.shape[1]*4)

            _, procesado = cv2.threshold(img, 110, 255, cv2.THRESH_BINARY)
               contornos_,   _   =   cv2.findContours(procesado,   cv2.RETR_TREE,
cv2.CHAIN_APPROX_SIMPLE)
            _fondo = np.full_like(procesado, 255)
            _fondo_contorno = np.full_like(procesado, 255)
            for cont in contornos_:
                # obtener altura del contorno
                altura = cv2.boundingRect(cont)[3]
                if altura > img.shape[0]*0.75:
                    _fondo = cv2.fillPoly(_fondo, [cont], (0, 0, 0))

            _fondo_contorno = cv2.bitwise_xor(gris_ref, _fondo)
            blur = cv2.GaussianBlur(_fondo_contorno,(3,3),0)
                                        _,              _fondo_contorno              =
cv2.threshold(blur,0,255,cv2.THRESH_BINARY+cv2.THRESH_OTSU)
                                                  _fondo_contorno                    =
_fondo_contorno[int(_fondo_contorno.shape[0]*0.15):, :]

            celda_notas = _fondo_contorno.copy()
                        celda_notas       =       imutils.resize(celda_notas,
width=celda_notas.shape[1]*2)
            for ce, curso in enumerate(cursos):
                _x = celda_notas.shape[1]//19
                _y = celda_notas.shape[0]
                nota = 0
                 while int(nota) not in range(1, 20) or len(str(nota).strip())
!= 2:
                     celda_unitaria = celda_notas[int(_y*0.08):_y-int(_y*0.2),
_x*ce+int(_x*0.15):_x*(ce+1)-int(_x*0.15)]
                           nota   =   pytesseract.image_to_string(celda_unitaria,
config=digit_options)
                    if nota == '':
                        break
                                if   (int(nota)   not   in   range(1,   20))   or
(len(str(nota).strip()) != 2):
                                    celda_notas   =   cv2.dilate(celda_notas,   None,
iterations=2)

                if nota == 0: nota = ''
                notas_cursos[curso] = nota

            _fondo_letras = _fondo_contorno.copy()
            _fondo_letras = cv2.erode(_fondo_letras, None, iterations=1)
```

```
                            _fondo_letras          =          _fondo_letras[:,
int(_fondo_letras.shape[1]*0.89):]

                            _comportamiento         =         _fondo_letras[:,
:int(_fondo_letras.shape[1]//2)]
                           _situacion_final         =         _fondo_letras[:,
int(_fondo_letras.shape[1]//2):]

              comportamiento   =   pytesseract.image_to_string(_comportamiento,
config=raw_character+' -c tessedit_char_whitelist=ABCD')
           comportamiento = re.sub(r'[^ABCD]', '', comportamiento)
           if len(comportamiento) == 2 and comportamiento == 'AB':
               comportamiento = 'AD'
           elif len(comportamiento) == 1 and comportamiento == 'D':
               comportamiento = 'B'
             situacion_final   =   pytesseract.image_to_string(_situacion_final,
config=raw_character+' -c tessedit_char_whitelist=ADPR')
           situacion_final = re.sub(r'[^ADPR]', '', situacion_final)
            if len(situacion_final) == 2 and situacion_final != 'RR' and
situacion_final != 'PP':
                if situacion_final == 'RB' or situacion_final == 'BR' or
situacion_final == 'BB':
                  situacion_final = 'RR'
               else:
                   situacion_final = situacion_final[0]
           elif len(situacion_final) == 1 and situacion_final == 'B':
               situacion_final = 'D'

           notas_cursos['Comportamiento'] = comportamiento
           notas_cursos['Situación Final'] = situacion_final
           return notas_cursos
```

Estas funciones de procesamiento de imágenes son fundamentales para preparar las actas de evaluación y asegurar que el OCR pueda extraer la información de manera precisa y eficiente.

4.2.4. Configuración de Tesseract

Un componente crucial durante el presente proyecto es Tesseract, el motor de OCR utilizado para extraer texto de las imágenes de las actas de evaluación. Tesseract es una herramienta de código abierto ampliamente utilizada en aplicaciones de OCR debido a su precisión y flexibilidad.

Configuramos Tesseract para optimizar el reconocimiento de diferentes tipos de texto presentes en las actas. Para esto, definimos varias opciones de configuración para Tesseract, adaptadas a las necesidades específicas del formato de las fichas. Cada

una de estas configuraciones utiliza diferentes modos de segmentación de páginas (--psm) y motores de reconocimiento (--oem) para optimizar el OCR según el tipo de texto que se desea extraer. Estas opciones son:

- **single_character:** Corresponde la configuración necesaria para reconocer sólo un carácter dentro de la imagen.

```
single_character = '--psm 10 --oem 3'
```

- **raw_character:** Línea en bruto. Trata la imagen como una única línea de texto.

```
raw_character = '--psm 13 --oem 3'
```

- **digit_options:** Configuración correspondiente a la detección de caracteres numéricos.

```
digit_options = single_character+" -c tessedit_char_whitelist=0123456789"
```

Además, especificamos la ruta del ejecutable de Tesseract en el sistema mediante la variable **pytesseract.pytesseract.tesseract_cmd**. Esto asegura que el motor OCR pueda ser invocado correctamente desde el código Python. Esta configuración fue importante para garantizar que Tesseract funcione de manera eficiente en el entorno de desarrollo.

```
pytesseract.pytesseract.tesseract_cmd = r'C:¥Program Files¥Tesseract-OCR¥tesseract.exe'
```

La configuración de Tesseract nos permite adaptar el motor OCR a las características específicas de las actas de evaluación, mejorando la precisión del reconocimiento de texto. Al definir opciones específicas para diferentes tipos de texto, se optimiza el proceso de extracción, asegurándonos que los datos extraídos sean lo más precisos y completos posible. Esta configuración es un paso muy importante para preparar el entorno de trabajo y asegurar que el OCR pueda manejar la variedad de texto presente en las actas de evaluación.

4.2.5. Función Principal

Establecimos la función principal **leer_imagen()**, que orquesta el flujo de procesamiento de las imágenes de las actas de evaluación. Esta función permite cargar las imágenes, aplicar las transformaciones necesarias, y extraer tanto la cabecera como el cuerpo de las actas utilizando las funciones de procesamiento de imágenes y OCR configuradas previamente.

La función **leer_imagen()** toma como entrada la ruta de una imagen y un número que la identifica. Carga la imagen desde la ruta especificada y obtiene la cabecera

utilizando la función **obtener_cabecera()**. Luego, aplica la corrección de perspectiva a la imagen para asegurar que esté correctamente alineada. Como se mencionó anteriormente, este paso es necesario para aplicar márgenes estándares permitiendo obtener una imagen bien alineada, lo que logra mejorar la precisión del OCR en la extracción de texto.

```
    def leer_imagen(image_path, numero):
        imagen = cv2.imread(image_path)
        cabecera = obtener_cabecera(imagen)[0]
        imagen = corregir_perspectiva(imagen, cabecera)

        contorno = obtener_cabecera(imagen)[1]
        imagen = cv2.drawContours(imagen, [contorno], -1, (0, 255, 0), 2)
        (x, y, w, h) = cv2.boundingRect(contorno)
        cabecera_img = imagen[y:y+h, x:x+w]
        _imagen = cabecera_img.copy()
        cuerpo_img = imagen[y+h:int(w//1.585), x:x+w]
        if 0 in cuerpo_img.shape:
            _imagen = imutils.resize(_imagen, width=1080)
            cv2.imshow('Contorno', _imagen)
            cv2.waitKey(0)
            cv2.destroyAllWindows()

        cabecera = None
        if(numero%2==0):
            cabecera = procesar_cabecera(cabecera_img)
        cuerpo = procesar_cuerpo(cuerpo_img)
        return (cabecera, cuerpo)
```

Una vez corregida la perspectiva, la función detecta y dibuja los contornos de la cabecera en la imagen. Utilizamos estos contornos para calcular el recuadro de la cabecera y extraerla de la imagen. Una vez hemos extraído la cabecera se procesa utilizando la función **procesar_cabecera()**, que aplica OCR para convertir el texto en datos digitales. Dado a que cada ficha de evaluación cuenta con una configuración de doble página, la cabecera sólo es procesada en iteraciones pares, mientras que el cuerpo siempre es procesado.

A continuación, la función extrae el cuerpo de la imagen, que contiene las notas y otros datos relevantes. Utilizamos la función **procesar_cuerpo()** para segmentar y extraer esta información, aplicando OCR para convertir el texto en datos digitales. La función **procesar_cuerpo()** se encarga de identificar las celdas que contienen nombres, códigos y notas, y de aplicar las técnicas de OCR configuradas para extraer esta información con precisión.

```
    def procesar_cuerpo(imagen):
        area_imagen = 1093184
        gris = cv2.cvtColor(imagen, cv2.COLOR_BGR2GRAY)
        _, binarizado = cv2.threshold(gris, 120, 255, cv2.THRESH_BINARY)
        erosionado = cv2.erode(binarizado, None, iterations=1)
        dilatado = cv2.dilate(erosionado, None, iterations=1)
            contornos = cv2.findContours(dilatado, cv2.RETR_EXTERNAL,
cv2.CHAIN_APPROX_SIMPLE)[0]
        output = []
        for contorno in contornos:
            area = cv2.contourArea(contorno)
            if area_imagen*0.015 > area > area_imagen*0.009:
                x, y, w, h = cv2.boundingRect(contorno)
                celda_nombres = gris[y:y+h, x:x+w]
                celda_codigo = binarizado[y:y+h, 34:x]
                celda_notas = binarizado[y-2:y+h, x+w+h+13:x+w+730]
                 if 0 in celda_nombres.shape or 0 in celda_codigo.shape or
0 in celda_notas.shape:
                    continue
                        celda_nombres = imutils.resize(celda_nombres,
width=celda_nombres.shape[1]*4)
                _y = celda_nombres.shape[0]
                                                    celda_nombres =
celda_nombres[int(_y*0.1):celda_nombres.shape[0]-int(_y*0.1),
6:int(celda_nombres.shape[1]*.75)]
                # Pinta el borde del recuadro con 255 de 3px
                celda_nombres = cv2.copyMakeBorder(celda_nombres, 8, 3, 3,
3, cv2.BORDER_CONSTANT, value=255)
                blur = cv2.GaussianBlur(celda_nombres, (3, 3), 0)
                    celda_nombres = cv2.threshold(blur, 50, 255,
cv2.THRESH_BINARY+cv2.THRESH_OTSU)[1]
                # Aplica Tesseract a cada celda
                    texto = pytesseract.image_to_string(celda_nombres,
lang="spa", config="--psm 7 --oem 3
tessedit_char_whitelist=ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz,
")
                texto = re.sub(r'[^a-zA-Z ]', '', texto)
                fondo_blanco = procesar_numeros(celda_codigo)
                    codigo = pytesseract.image_to_string(fondo_blanco,
config=digit_options)
                if codigo == '' or texto == '':
                    continue
                    output.append({'codigo': codigo[-15:], 'nombres':
texto.strip(), 'notas': procesar_notas(celda_notas, gris[y-2:y+h,
x+w+h+13:x+w+730])})
        output.reverse()
        return output
```

Finalmente, la función **leer_imagen()** retorna un diccionario que contiene la cabecera y el cuerpo procesados de la imagen. Este diccionario se puede utilizar para almacenar los datos extraídos en una base de datos estructurada, facilitando su gestión y consulta.

4.2.6. Preprocesamiento de Imágenes

El preprocesamiento de las imágenes es una etapa para preparar las imágenes de las actas de evaluación antes de aplicar técnicas de OCR. El objetivo del preprocesamiento es mejorar la calidad de las imágenes y resaltar las características importantes, facilitando así la extracción precisa de texto.

Comenzamos con la conversión de las imágenes a escala de grises. Esta transformación simplifica la imagen al eliminar la información de color, reduciendo el número de canales a sólo uno, lo que reduce la complejidad del procesamiento posterior. La conversión a escala de grises es especialmente útil para resaltar los contrastes entre el texto y el fondo, siendo un proceso esencial para facilitar la detección de bordes de manera efectiva.

```
    gris = cv2.cvtColor(imagen, cv2.COLOR_BGR2GRAY)
```

Luego de esto, aplicamos binarización, un proceso que convierte la imagen en una representación en blanco y negro mediante el establecimiento de un umbral fijo o adaptativo que decanta el valor de los pixeles en blanco o negro según su valor respecto al umbral. Este paso es fundamental para resaltar el texto y otros elementos importantes de la imagen, eliminando el ruido y las variaciones de color que podrían interferir con el OCR.

```
    _, binarizado = cv2.threshold(gris, 120, 255, cv2.THRESH_BINARY)
```

Otro paso importante en el preprocesamiento es la corrección de perspectiva. Las imágenes de las actas de evaluación pueden estar inclinadas o distorsionadas debido a la forma en que fueron escaneadas o fotografiadas. La corrección de perspectiva ajusta la orientación de la imagen para asegurar que el texto esté alineado horizontalmente, lo que mejora la precisión del OCR. Este proceso utiliza técnicas geométricas para calcular y aplicar la rotación necesaria.

Además, aplicamos técnicas de eliminación de ruido para mejorar la calidad de la imagen. Incluyendo el uso de filtros como el filtro Gaussiano, que suaviza la imagen y reduce el ruido sin perder detalles importantes. La eliminación de ruido sirve para asegurar que el OCR pueda distinguir claramente entre el texto y el fondo.

```
        blur = cv2.GaussianBlur(celda_nombres, (3, 3), 0)
```

4.2.7. Segmentación y detección de contornos

La segmentación y detección de contornos permite identificar y aislar las regiones de interés (ROI) en las imágenes de las actas de evaluación. Durante este proceso localizamos y extrajimos secciones específicas de la imagen, como la cabecera y el cuerpo, que contienen la información relevante para el análisis.

Empezamos con la detección de bordes, que es fundamental para resaltar los límites de los objetos presentes en la imagen. Para esto, utilizamos algoritmos como el de Canny, se identifican los bordes significativos, lo que facilita la posterior detección de contornos. La detección de bordes nos permite convertir la imagen en una representación binaria donde los píxeles de los bordes se destacan, permitiendo una identificación más precisa de las regiones de interés.

```
        bordes = cv2.Canny(humbralizado, 100, 250)
```

Figura 4.
Detección de borde

Tras detectar los bordes, se realizó la detección de los contornos. Estos contornos son curvas que unen todos los puntos continuos a lo largo de un límite, teniendo el mismo color o intensidad. Para esto, utilizamos algunas funciones de OpenCV para encontrar y dibujar estos contornos.

```
        contornos, _ = cv2.findContours(bordes, cv2.RETR_EXTERNAL,
cv2.CHAIN_APPROX_SIMPLE)
```

El siguiente paso fue calcular los *bounding boxes* de los contornos detectados. Un *bounding box* es el rectángulo más pequeño que puede contener un contorno. Utilizamos este rectángulo para aislar y extraer las regiones de interés de la imagen. Por ejemplo, en la cabecera de las actas, los *bounding boxes* permiten extraer secciones específicas como el nombre de la institución, el grado y la sección.

La segmentación y detección de contornos también incluye la eliminación de contornos irrelevantes. Esto lo logramos filtrando los contornos según su área, forma y posición en la imagen. Solo se conservó los contornos que cumplen con ciertos criterios predefinidos, asegurando que solo se procesen las regiones relevantes. Este filtrado ayuda a evitar la extracción de ruido o elementos no deseados que podrían interferir con el OCR.

```
        for contorno in contornos:
                perimetro = cv2.arcLength(contorno, True)
                   approx = cv2.approxPolyDP(contorno, (5*0.01) * perimetro,
True)
                if len(approx) == 4:
                    (x, y, w, h) = cv2.boundingRect(contorno)
                                    if int(area_imagen*29*0.01) > w*h >
int(area_imagen*13*0.01):
                                    c+=1
                                    return (approx, contorno)
```

Finalmente, preparamos las regiones de interés extraídas para el OCR. Esto incluye el redimensionamiento de las celdas, la aplicación filtros adicionales para mejorar la calidad de la imagen y asegurarnos que el texto esté claramente visible.

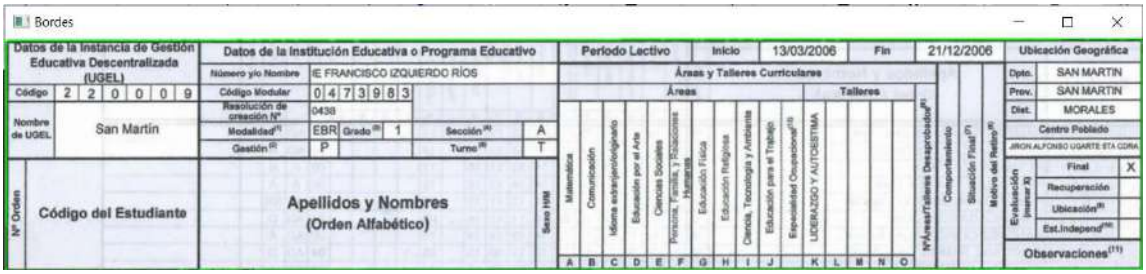

Figura 5.
Regiones de interés extraídas para el OCR

4.2.8. Postprocesamiento de Texto

Es una etapa en el cual limpiamos y formateamos los datos extraídos mediante OCR con el objetivo de asegurar que el texto extraído sea preciso, coherente y de esta manera esté listo para su almacenamiento y análisis. Esta etapa incluye la corrección de errores, la eliminación de caracteres no deseados y el formateo del texto según las necesidades del proyecto.

Una de las primeras tareas en el postprocesamiento de texto es la limpieza de los datos extraídos debido a que, durante el OCR, es común que se introduzcan errores producidos por la baja calidad de la imagen o por la naturaleza del texto. Por ejemplo, caracteres no deseados como símbolos o números pueden aparecer en los nombres

de los estudiantes. Para abordar esto, utilizamos expresiones regulares y otras técnicas de limpieza para eliminar dichos caracteres no deseados y asegurar que solo se conserven los datos relevantes.

```
        texto = re.sub(r'[^a-zA-Z ]', '', texto)
```

Además de la limpieza, el postprocesamiento incluye la corrección de errores comunes. Por ejemplo, si se detectan caracteres no válidos en un código de estudiante, se puede aplicar una lógica de corrección para asegurar que el código sea válido. De manera similar, se pueden aplicar reglas específicas para corregir errores comunes en los nombres y las notas. Estas correcciones son esenciales para asegurar la precisión y coherencia de los datos extraídos.

```
        for ce, curso in enumerate(cursos):
                _x = celda_notas.shape[1]//19
                _y = celda_notas.shape[0]
                nota = 0
                 while int(nota) not in range(1, 20) or len(str(nota).strip())
!= 2:
                     celda_unitaria = celda_notas[int(_y*0.08):_y-int(_y*0.2),
_x*ce+int(_x*0.15):_x*(ce+1)-int(_x*0.15)]
                          nota = pytesseract.image_to_string(celda_unitaria,
config=digit_options)
                    if nota == '':
                        break
                               if (int(nota) not in range(1, 20)) or
(len(str(nota).strip()) != 2):
                                  celda_notas = cv2.dilate(celda_notas, None,
iterations=2)

                if nota == 0: nota = ''
                notas_cursos[curso] = nota
```

4.2.9. Implementación en Django

La implementación de este sistema OCR para la digitalización de actas de evaluación se realizó utilizando el framework Django, que proporciona una estructura robusta y escalable para desarrollar aplicaciones web a través del lenguaje de programación Python, en la versión 3.12. Django facilita la creación de una interfaz web intuitiva y funcional, así como la gestión eficiente de los datos extraídos mediante OCR. A continuación, se describe en detalle la implementación de los principales módulos de la aplicación.

4.2.9.1. Interfaz de Usuario y Autenticación

La aplicación cuenta con una interfaz de usuario que permite a los usuarios interactuar con el sistema de manera intuitiva. La autenticación se maneja mediante un sistema de login, asegurando que solo los usuarios autorizados puedan acceder a las funcionalidades de la aplicación. Django proporciona un sistema de autenticación integrado que maneja la gestión de usuarios y permisos.

4.2.9.2. Módulo de Carga

En el módulo de carga es donde los usuarios pueden subir los archivos PDF que contienen las actas de evaluación. Este módulo incluye una vista que maneja tanto las solicitudes GET como POST. Cuando un usuario sube un archivo PDF, la vista guarda el archivo en el servidor y luego lo convierte a imágenes JPG utilizando la función **convert_pdf_to_jpg()** vista anteriormente. Las imágenes generadas se listan y se preparan para el procesamiento posterior.

Figura 6.
Módulo de Carga

4.2.9.3. Procesamiento de Imágenes y OCR

Una vez que las imágenes JPG se han generado, se aplican las técnicas de procesamiento de imágenes y OCR descritas anteriormente. Este procesamiento se realiza en segundo plano para asegurar que la interfaz de usuario permanezca responsiva. Django puede integrarse con herramientas como Celery para manejar tareas en segundo plano, lo que permite procesar grandes volúmenes de datos de manera eficiente.

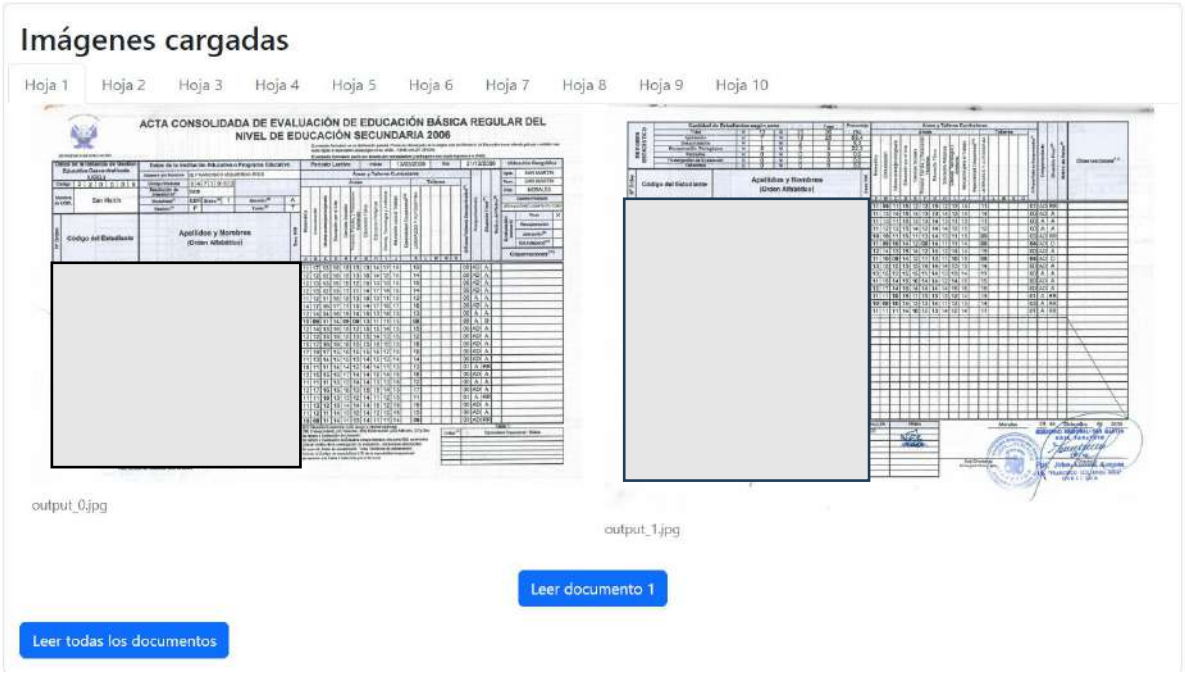

Figura 7.
Imágenes cargadas

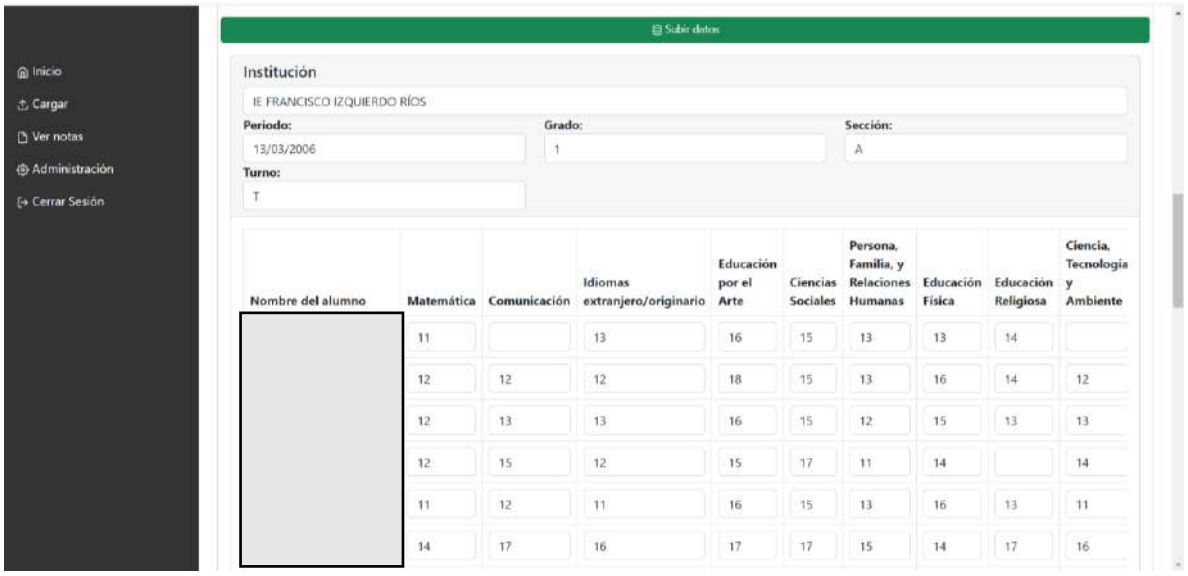

Figura 8.
Interfaz de carga de imagenes

4.9.2.4. Almacenamiento en Base de Datos

Los datos extraídos mediante OCR se estructuran en diccionarios y se almacenan en una base de datos MySQL 8.0.30. Django proporciona un ORM (Object-Relational Mapping) que facilita la interacción con la base de datos, permitiendo crear, leer, actualizar y eliminar registros de manera sencilla. Las tablas de la base de datos incluyen las siguientes tablas:

- **instituciones_educativas:** Como indica su nombre, almacena información sobre las instituciones educativas. Incluye campos como:
  - **id**: Identificador único de la institución.
  - **descripcion**: Nombre de la institución educativa.
  - **estado**: Estado del registro.
- **periodos_lectivos:** Contiene los periodos lectivos durante los cuales se recopilan las notas. Sus campos son:
  - **id**: Identificador único del periodo.
  - **inicio**: Fecha de inicio del periodo.
  - **fin:** Fecha de finalización del periodo.
  - **estado:** Estado del registro de periodo.
- **secciones:** La tabla de secciones almacena información sobre las diferentes secciones dentro de las instituciones educativas. Incluye:
  - **id**: Identificador único de la sección.
  - **descripcion**: Descripción o nombre de la sección.
  - **id_institucion_educativa**: Identificador de la institución educativa a la que pertenece la sección.
  - **estado**: Estado del registro de la sección.
- **estudiantes:** Esta tabla almacena información sobre los estudiantes. Sus campos son:
  - **id**: Identificador único del estudiante.
  - **codigo**: Código del estudiante.
  - **nombres**: Nombres del estudiante.
  - **apellidos**: Apellidos del estudiante.
  - **estado**: Estado del registro del estudiante.
- **secciones_estudiantes:** Esta tabla relaciona a los estudiantes con las secciones en las que están inscritos por lo que funciona como una tabla detalle. Tiene los siguientes atributos:
  - **id**: Identificador único del detalle.
  - **id_seccion**: Identificador de la sección.
  - **id_estudiante**: Identificador del estudiante.
  - **estado**: Estado del registro detalle.
- **notas:** La tabla de notas almacena las calificaciones de los estudiantes en las diferentes asignaturas. Sus campos son:
  - **id**: Identificador único de la nota.
  - **id_asignatura**: Identificador de la asignatura.
  - **id_estudiante**: Identificador del estudiante.
  - **nota**: Calificación obtenida.
  - **id_periodo**: Identificador del periodo lectivo.
  - **estado**: Estado del registro de la nota.
- **asignaturas:** Esta tabla almacena información sobre las asignaturas. Los campos con los que cuenta son:
  - **id**: Identificador único de la asignatura.
  - **descripcion**: Descripción o nombre de la asignatura.
  - **estado**: Estado del registro de la asignatura.

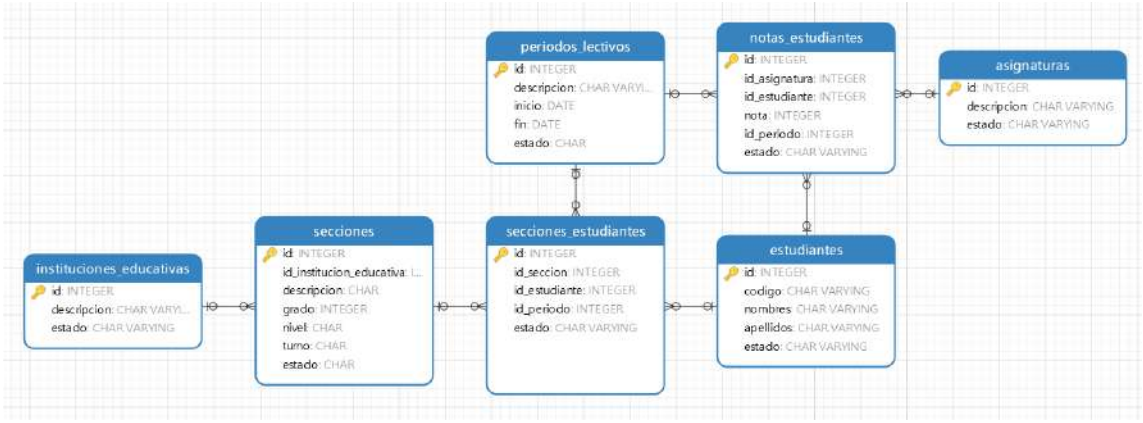

Figura 9.
Modelo Entidad-Relación de la base de datos

4.2.9.5. Módulo de Visualización de Notas

El módulo de visualización de notas permite a los usuarios consultar y administrar las notas de los estudiantes. Este módulo incluye vistas que recuperan los datos de la base de datos y los presentan en una interfaz amigable. Los usuarios pueden buscar y filtrar las notas por diferentes criterios, facilitando la gestión de la información académica.

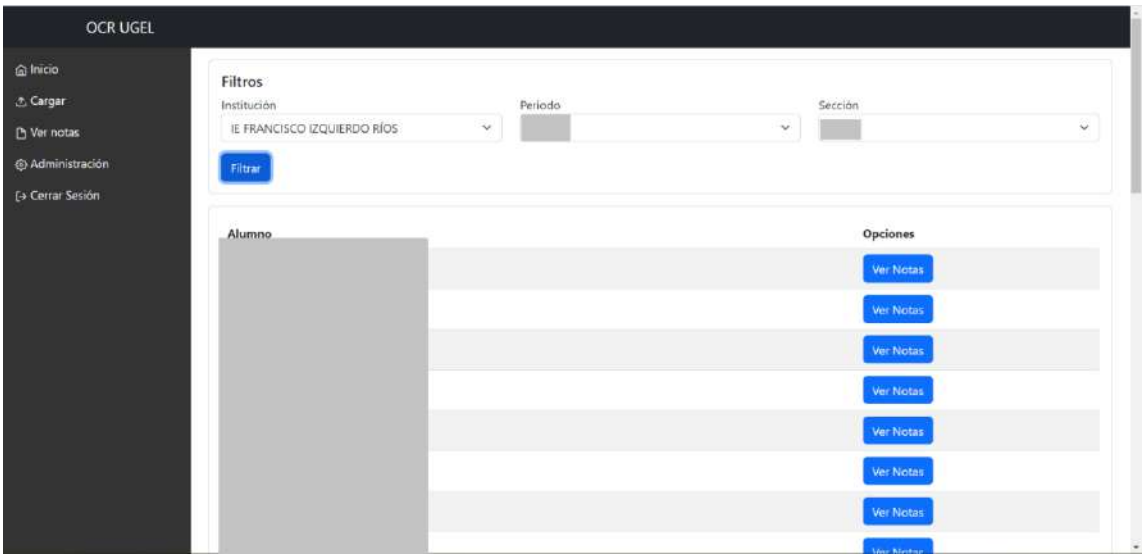

Figura 10.
Módulo de Visualización de Notas

4.2.9.6. Módulo de Administración

El módulo de administración permite gestionar las tablas de mantenimiento, como asignaturas, estudiantes por sección, instituciones educativas, periodos lectivos y secciones. Django incluye un panel de administración integrado que facilita la gestión de estos datos. Los administradores pueden agregar, editar y eliminar registros

directamente desde el panel de administración, asegurando que la información esté siempre actualizada.

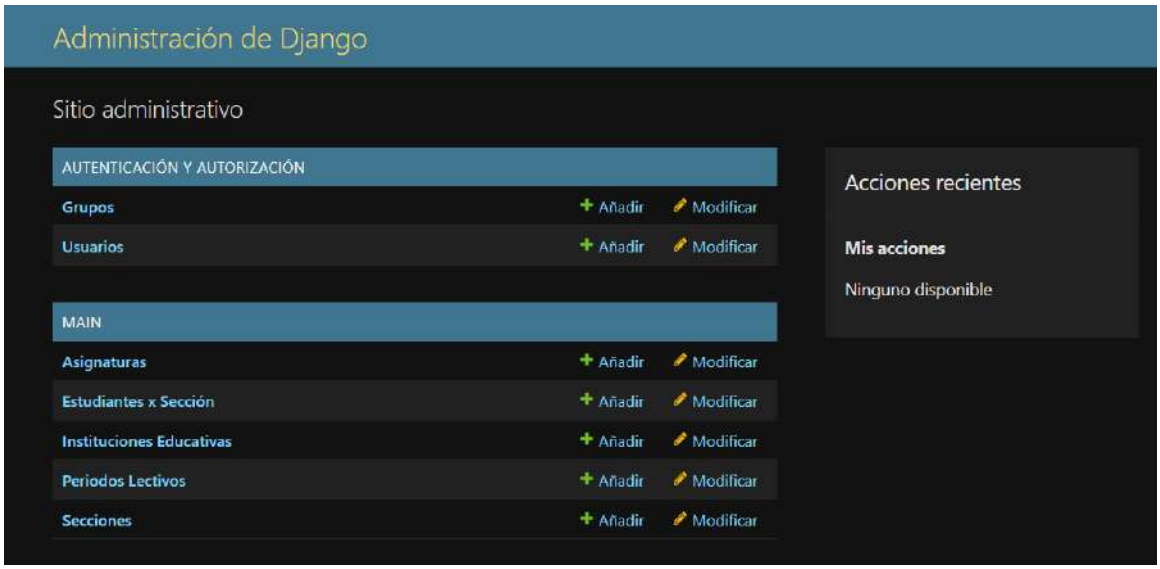

Figura 11.
Módulo de Administración

4.2.10. Prueba de calidad de digitalización

Para evaluar la calidad de la digitalización de las actas de evaluación en la UGEL San Martín mediante técnicas de OCR, se tuvo en cuenta tres indicadores:

- Tiempo de digitalización: Se midió el tiempo en segundos que toma digitalizar un acta de evaluación con la herramienta tecnológica desarrollada, utilizando un cronómetro para registrar la duración del proceso.
- Tasa de error de digitalización: Se calculó dividiendo los campos erróneos por el total de campos extraídos mediante la técnica de OCR, y luego se multiplicó por 100 para obtener la tasa en porcentaje.

$$Tasa\ de\ error = \frac{Campos\ erróneos}{Total\ campos} x100$$

- Tasa de precisión: Se calculó dividiendo los campos correctos por el total de campos extraídos mediante la técnica de OCR, y luego se multiplicó por 100 para obtener la tasa en porcentaje.

$$Tasa\ de\ precisión = \frac{Campos\ correctos}{Total\ campos} x100$$

Tabla 3.
Indicadores de desempeño del proceso de digitalización de actas

Indicadores	N	Mínimo	Máximo	Media	Desv. Estándar
Tiempo de digitalización	30	74	154	112,30	20,28
Tasa de error	30	2,16	40,28	11,33	8,53
Tasa de precisión	30	59,72	98,51	91,13	8,87

De acuerdo a la Tabla 3, el tiempo promedio para digitalizar un acta de evaluación con el sistema OCR implementado en la UGEL San Martin fue de 112,30 seg, con un rango que va desde un mínimo de 74 seg hasta un máximo de 154 seg. La desviación estándar de 20,28 seg indica una variabilidad moderada en los tiempos de procesamiento, lo que sugiere que, aunque la mayoría de las actas se digitalizan en un tiempo relativamente similar, algunos casos pueden tardar más.

La tasa de error promedio fue del 11,33%, con valores mínimos de 2,16% y máximos de 40,28%. La desviación estándar de 8,53% refleja una variabilidad en la exactitud del proceso de digitalización, lo que indica que, si bien en la mayoría de los casos los errores son relativamente bajos, existen ciertos documentos en los que la precisión de la extracción de texto mediante OCR es considerablemente menor, lo cual sugiere que se deben implementar mejoras en el reconocimiento de caracteres para minimizar estos errores y lograr mayor consistencia.

Por otro lado, la tasa de precisión fue alta, con un promedio de 91,13%, lo que demuestra la eficacia general de la digitalización a través de OCR. El rango de precisión varía entre 59,72% y 98,51%, con una desviación estándar de 8,87%, indicando que la mayoría de las actas alcanzan una alta precisión en la extracción de texto, aunque algunos documentos presentan menor exactitud. Estos resultados confirman que la herramienta utilizada es efectiva, pero es necesario optimizar el proceso para reducir la variabilidad y alcanzar una mayor uniformidad en la calidad de digitalización.

4.3 Resultado específico 3: Medir la disponibilidad de actas de evaluación en la UGEL San Martín después de su digitalización basada en reconocimiento óptico de caracteres.

Tras implementar el acceso digital a las actas de evaluación mediante tecnologías OCR, se midieron nuevamente las dimensiones de accesibilidad y disponibilidad utilizando los mismos instrumentos del pretest. En cuanto a la accesibilidad, el tiempo promedio de localización de las actas de evaluación se redujo a 118,07 seg de 903,53 seg en promedio, con una desviación estándar de 35,33 seg (Tabla 4). Esto evidencia

una mejora en la optimización del tiempo para el visado de certificados de estudios solicitados en la UGEL San Martín, debido a la facilidad de localización de las actas de evaluación que previamente se realizaba de forma manual.

Tabla 4.
Postest de la accesibilidad a las actas de evaluación

Indicador	N	Mínimo	Máximo	Media	Desv. Estándar
Tiempo de localización de actas de evaluación	30	61	176	118,07	35,33

Con respecto a la dimensión de disponibilidad, después de implementar la digitalización de las actas de evaluación en la UGEL San Martín mediante técnicas de OCR y tras explicar su funcionalidad a los funcionarios, la mayoría (82%) percibió que la disponibilidad es alta (Figura 12), en contraste con el pretest, donde el 55% la calificó como regular. Esto evidencia mejoras en la disponibilidad de las actas, destacando la relevancia de la solución tecnológica planteada.

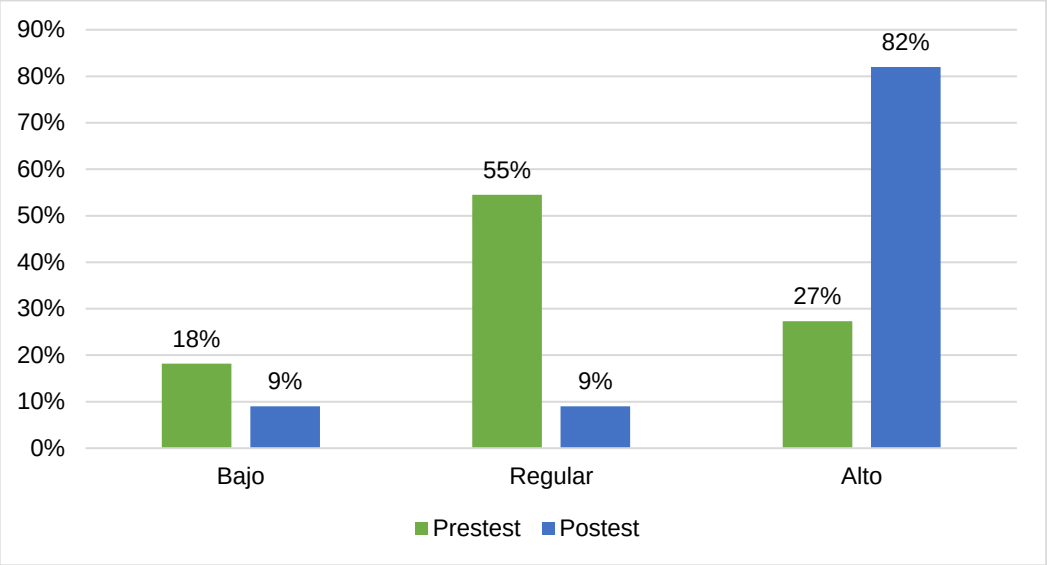

Figura 12.
Comparación del pre y postest de la disponibilidad de las actas de evaluación

Prueba de hipótesis

H_0: La digitalización basada en reconocimiento óptico de caracteres no mejora la disponibilidad de actas de evaluación en la UGEL San Martín.

H_1: La digitalización basada en reconocimiento óptico de caracteres mejora significativamente la disponibilidad de actas de evaluación en la UGEL San Martín.

- **Dimensión accesibilidad**

Nivel de significancia: 5% o 0,05

Prueba de normalidad de los datos: Se tuvo en cuenta la prueba de normalidad de Shapiro-Wilk para muestras menores o iguales a 50. Ya que ambos resultados del pre y postest fueron mayores que 0,05 (nivel de significancia), se determina que los datos provienen de una distribución normal (Tabla 5).

Tabla 5.
Prueba de normalidad de los datos

	Shapiro-Wilk		
	Estadístico	**gl**	**Sig.**
Pretest	0,962	30	0,353
Postest	0,955	30	0,233

Elección de la prueba estadística: Dado que los datos siguen una distribución normal, son cuantitativos y provienen de muestras independientes recopiladas en diferentes procesos, se optó por utilizar la prueba paramétrica T de Student (Flores-Ruiz et al., 2017).

Estimación del p-valor: De acuerdo a la Tabla 6, la prueba T Student de muestras independientes evidencia que hay una diferencia significativa en la accesibilidad de las actas de evaluación antes y después de la digitalización mediante OCR, ya que el valor de significancia (p = 0,000) es menor a 0,05. La prueba de Levene indica que no se asumen varianzas iguales (F = 41,926; p = 0,000), por lo que se interpreta la fila correspondiente. La diferencia de medias es de 785,467 segundos, con un intervalo de confianza del 95% entre 723,995 y 846,939, lo que confirma una mejora significativa en la accesibilidad.

Toma de decisión: Dado que el valor de significancia (p = 0,000) es menor al nivel de significancia típico (0,05), se rechaza la hipótesis nula (H0). Por lo tanto, se acepta la hipótesis alternativa (H1), concluyendo que la digitalización basada en reconocimiento óptico de caracteres mejora significativamente la accesibilidad de las actas de evaluación en la UGEL San Martín.

Tabla 6.
Prueba T Student

		Prueba de Levene		Prueba t para la igualdad de medias						
		F	**Sig.**	**t**	**gl**	**Sig. (bilateral)**	**Dif. de medias**	**Dif. de error estándar**	**95% de intervalo de confianza**	
									Inf	**Sup**
Accesibilidad	Se asumen varianzas iguales	41,926	0,000	26,035	58	0,000	785,67	30,17	725,07	845,85
	No se asumen varianzas iguales			26,035	31,77	0,000	785,47	30,17	723,99	846,94

- **Dimensión disponibilidad**

Nivel de significancia: 5% o 0,05

Prueba de normalidad de los datos: No corresponde porque los datos son cualitativos ordinales, por lo tanto, se debe aplicar una prueba no paramétrica, como señalan Juárez García et al. (2002).

Elección de la prueba estadística: Se seleccionó la prueba de Wilcoxon, ya que se cumple los supuesto de que los datos son categóricos y la muestra es relacional (aplicado a una sola muestra antes y después) (Juárez García et al., 2002).

Estimación del p-valor: La Tabla 7 muestra los resultados de la prueba de rangos con signo de Wilcoxon para comparar la disponibilidad de las actas de evaluación antes y después de la digitalización basada en OCR, donde se reportó que el valor de Z es -2,111 y el valor de significancia asintótica bilateral es 0,035.

Toma de decisión: Dado que el valor p (0,035) es menor que 0,05, se rechaza la hipótesis nula (H0) y se acepta la hipótesis alternativa (H1), es decir, la digitalización basada en reconocimiento óptico de caracteres mejora significativamente la disponibilidad de las actas de evaluación en la UGEL San Martín.

Tabla 7.
Prueba de Wilcoxon

	Pos_Disponibilidad - Pre_Disponibilidad
Z	-2,111[b]
Sig. asintótica(bilateral)	**0,035**

a. Prueba de rangos con signo de Wilcoxon
b. Se basa en rangos negativos.

Los hallazgos sobre la mejora en la disponibilidad de documentos a través de la digitalización con OCR se alinean con el estudio de Chumwatana y Rattana-umnuaychai (2021), quienes demostraron que, al aplicar técnicas de OCR para convertir documentos físicos a formato digital, se obtuvo un rendimiento de precisión promedio del 75,38% en la extracción de atributos y del 66,92% en la extracción de valores de documentos impresos. Este estudio concluye que la tecnología OCR aporta importantes beneficios a las organizaciones, facilitando la búsqueda de información y reduciendo la acumulación de papel en oficinas.

Asimismo, coincide con la investigación de Ayala-Churo et al. (2022), que propuso un modelo de detección basado en el algoritmo YOLOv5 y la librería Tesseract-OCR para extraer y contextualizar campos en documentos de identificación. Los resultados de esta investigación reflejaron una alta precisión del 93,7%, un recall del 99,2% y una precisión media de entrenamiento (mAP) de 100%, evidenciando la efectividad de estas tecnologías para el procesamiento de documentos. Por tanto, es evidente que

las técnicas de OCR ofrecen soluciones eficientes para mejorar la disponibilidad y accesibilidad de documentos en diferentes contextos.

Además, la aplicación de OCR en la digitalización de actas de evaluación de la UGEL San Martín demuestra que esta tecnología no solo facilita el acceso a la información, sino que también contribuye a mejorar la eficiencia administrativa al reducir el tiempo de búsqueda y manipulación de documentos. Este enfoque puede ser replicado en otros procesos administrativos para optimizar la gestión documental, mejorar la trazabilidad y asegurar el fácil acceso a datos relevantes. Así, la digitalización con OCR no solo aporta a la disponibilidad de documentos, sino que también apoya la transformación digital de las instituciones (Zaqueu, 2024), permitiendo un manejo más efectivo de la información y el aprovechamiento de recursos tecnológicos para la toma de decisiones efectivas.

CONCLUSIONES

1. Antes de la digitalización, el tiempo promedio para localizar las actas de evaluación en la UGEL San Martín fue de 903,53 segundos (aproximadamente 15 minutos), y la mayoría de los funcionarios (55%) calificaba la disponibilidad como "regular" debido a problemas con la búsqueda, organización y protección de las actas.

2. Se logró digitalizar las actas de evaluación en la UGEL San Martín mediante técnicas de OCR de manera eficiente utilizando tecnologías como el software Tesseract para la extracción de texto y bases de datos SQL para el almacenamiento seguro de la información digitalizada. El proceso incluyó el escaneo de las actas en alta resolución, el preprocesamiento de las imágenes (conversión a escala de grises, binarización y corrección de perspectiva) para optimizar la calidad del reconocimiento y la segmentación de áreas de interés.

3. Después de digitalizar las actas de evaluación a través de tecnologías OCR en la UGEL San Martín el tiempo promedio de localización de las actas se redujo de 903,53 seg a 118,07 seg, evidenciando una optimización considerable en el proceso de visado de certificados de estudios. Asimismo, la percepción de disponibilidad entre los funcionarios pasó de ser considerada "regular" por el 55% en el pretest, a "alta" por el 82% en el postest, reflejando el impacto de la solución implementada.

4. A un nivel de confianza del 95%, se concluye que la digitalización basada en reconocimiento óptico de caracteres mejoró significativamente la accesibilidad y disponibilidad de las actas de evaluación en la UGEL San Martín, debido a que el p-valor (0,000 y 0,035, respectivamente) fue menor al nivel de significancia (0,05).

RECOMENDACIONES

1. Para maximizar los beneficios de la digitalización de actas, es fundamental brindar capacitación regular a los funcionarios de la UGEL San Martín sobre el manejo adecuado de las herramientas tecnológicas y su actualización, asegurando un uso eficiente del sistema y el aprovechamiento completo de sus funcionalidades.

2. Debido que la digitalización aumenta la accesibilidad y disponibilidad de las actas, es importante implementar medidas de respaldo automático y protocolos de seguridad para proteger la información, garantizando que los documentos digitalizados estén seguros y disponibles en caso de fallos tecnológicos.

3. Es recomendable realizar revisiones periódicas del sistema de OCR y su impacto en los procesos administrativos, con el fin de identificar posibles mejoras o ajustes necesarios. Este monitoreo permitirá mantener la eficiencia lograda e incluso mejorarla con el tiempo, atendiendo a cambios en las necesidades de los usuarios.

4. Finalmente, se sugiere ampliar la implementación de tecnologías OCR a otros procesos administrativos de la UGEL San Martín para contribuir a agilizar trámites adicionales, optimizar recursos y mejorar el servicio en otras áreas de gestión.

REFERENCIAS BIBLIOGRÁFICAS

Abbasova, V. S. (2020). Main concepts of the document management system required for its implementation in enterprises. *ScienceRise*, *1*, 32-37. https://doi.org/10.21303/sr.v0i1.1149

Abdul Hamid, M. S. R., Mohd Arzaman, A. F., Razali, M. A., Masrom, N. R., & Margono, M. (2023). The Development of Smart Document Management System With Mobile Application Technology In Agricul-Tural Sector (Malaysia Sustainability Palm Oil). *Journal of Technology Management and Technopreneurship*, *11*(1), 28–43.

Apaza Flores, D. D. E. (2022). *Implementación de lectura de documentos digitales con machine learning en la empresa Cli Gestiones Aduaneras S.A.* [Universidad Nacional de Moquegua]. http://181.176.3.22/handle/UNAM/452

Ayala-Churo, E. A., Cuenca, J. P., & Quevedo-Sacoto, A. S. (2022). Detección y digitalización de datos de interés en documentos de identificación. *Domino De Las Ciencias*, *8*(3), 2167-2185. https://dominiodelasciencias.com/ojs/index.php/es/article/view/2995

Azzam, F., Jaber, M., Saies, A., Kirresh, T., Awadallah, R., Karakra, A., Barghouthi, H., & Amarneh, S. (2023). The Use of Blockchain Technology and OCR in E-Government for Document Management: Inbound Invoice Management as an Example. *Applied Sciences*, *13*(14), 8463. https://doi.org/10.3390/app13148463

Berchmans, D., & Kumar, S. S. (2014). Optical character recognition: An overview and an insight. *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, 1361-1365. https://doi.org/10.1109/ICCICCT.2014.6993174

Buctuanon, M. M., Gadiane, J. L. A., Margallo, F. A., & Lucero, P. R. (2021). Incorporating Rule-based Pattern Recognition Approach for Document Structure Classification on Cloud-based Document Management System. *Mindanao Journal of Science and Technology*, *19*(2), 17-39. https://mjst.ustp.edu.ph/index.php/mjst/article/view/999

Camilo Momblanc, L., & Castro Milán, H. Y. (2020). La gestión documental y el control interno: un binomio indispensable. *Revista Del Archivo Nacional*, *84*(1-12), 9–26.

Cañarte-Aizprua, K. J., Romero-Fernández, A. J., Cañizares-Galarza, F. P., & Machuca-Vivar, S. A. (2022). Sistema informático para la digitalización y gestión

del archivo histórico de una registraduría, Chone-Ecuador. *CIENCIAMATRIA*, *8*(4), 676-684. https://doi.org/10.35381/cm.v8i4.879

Carrillo Fuertes, T. I. (2020). *Desarrollo de un aplicativo movil para la extracción automática de información del documento de indentificación mediante visión computacional* [Pontificia Universidad Católica del Perú]. https://tesis.pucp.edu.pe/repositorio/handle/20.500.12404/16608

Carrillo Morales, E. L., & Chávez Chiquillan, J. (2022). *La digitalización de documentos y la gestión administrativa en la Municipalidad Provincial de Chincheros, 2022* [Universidad Tecnológica del Perú]. https://repositorio.utp.edu.pe/handle/20.500.12867/6834

Chaudhuri, A., Mandaviya, K., Badelia, P., & Ghosh, S. K. (2017). Optical Character Recognition Systems. En *Optical Character Recognition Systems for Different Languages with Soft Computing.* (pp. 9-41). Studies in Fuzziness and Soft Computing, vol 352. Springer, Cham. https://doi.org/10.1007/978-3-319-50252-6_2

Chavez, L., & Salinas, J. (2021). Segmentación de los alumnos ingresantes a una universidad pública aplicando el algoritmo K-prototype. *Tierra nuestra*, *15*(2), 10-21. https://doi.org/10.21704/rtn.v15i2.1825

Cheung, A., Bennamoun, M., & Bergmann, N. W. (2001). An Arabic optical character recognition system using recognition-based segmentation. *Pattern Recognition*, *34*(2), 215-233. https://doi.org/10.1016/S0031-3203(99)00227-7

Chumwatana, T., & Rattana-umnuaychai, W. (2021). Using OCR Framework and Information Extraction for Thai Documents Digitization. *2021 9th International Electrical Engineering Congress (iEECON)*, 440-443. https://doi.org/10.1109/iEECON51072.2021.9440300

Das, M., Tao, X., Liu, Y., & Cheng, J. C. P. (2022). A blockchain-based integrated document management framework for construction applications. *Automation in Construction*, *133*, 104001. https://doi.org/10.1016/j.autcon.2021.104001

Díaz Jiménez, A., & Mena Mujica, M. M. (2022). Política de gestión documental para la Universidad Central “Marta Abreu” de Las Villas. *Investigación Bibliotecológica: archivonomía, bibliotecología e información*, *36*(92), 153. https://doi.org/10.22201/iibi.24488321xe.2022.92.58565

Flores-Ruiz, E., Miranda-Novales, M. G., & Villasís-Keever, M. Á. (2017). El protocolo

de investigación VI: cómo elegir la prueba estadística adecuada. Estadística inferencial. *Revista Alergia México*, *64*(3), 364-370. https://doi.org/10.29262/ram.v64i3.304

García Peñalvo, F. J. (2021). Transformación digital en las universidades: Implicaciones de la pandemia de la COVID-19. *Education in the Knowledge Society (EKS)*, *22*, e25465. https://doi.org/10.14201/eks.25465

Golesworthy, T., Pepper, J., Takkenberg, J. J., & Treasure, T. (2022). Report of Evaluation of Personalised External Aortic Root Support (PEARS) Using the Ideal Framework after 18 Years: Have Questions Merged that are Amenable to a Randomised Controlled Clinical Trial? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4207913

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, *23*(1), 40-55. https://doi.org/10.1038/s41580-021-00407-0

Jain, P., Taneja, D. K., & Taneja, D. H. (2021). Which OCR toolset is good and why? A comparative study. *Kuwait Journal of Science*, *48*(2). https://doi.org/10.48129/kjs.v48i2.9589

Jayoma, J. M., Moyon, E. S., & Morales, E. M. O. (2020). OCR Based Document Archiving and Indexing Using PyTesseract: A Record Management System for DSWD Caraga, Philippines. *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, 1-6. https://doi.org/10.1109/HNICEM51456.2020.9400000

Juárez García, F., Villatoro Velázquez, J. A., & López Lugo, E. K. (2002). *Apuntes de Estadística Inferencial*. Instituto Nacional de Psiquiatría Ramón de la Fuente.

Khalil, M., & Fakhratov, M. (2023). *Improvement of document management system in construction*. 050029. https://doi.org/10.1063/5.0143799

Kruchinin, S. V., & Bagrova, E. V. (2019). Systems of Electronic Document Management in Russian Education. Pros and Cons. *2019 International Conference «Quality Management, Transport and Information Security, Information Technologies» (IT&QM&IS)*, 628-630. https://doi.org/10.1109/ITQMIS.2019.8928315

Makhnevich, D. (2023). Theoretical overview of the implementation of electronic

document management in institutions of general secondary education. *Pedagogical Sciences*, *1(112)*, 177-189. https://doi.org/10.35433/pedagogy.1(112).2023.177-189

Martínez Cano, H. (2021). Sistema informático para la digitalización del expediente académico del archivo histórico de la secretaría docente. *Serie Científica de la Universidad de las Ciencias Informáticas*, *14*(9), 57-74. https://dialnet.unirioja.es/servlet/articulo?codigo=8590640

Mazon-Fierro, M., Molina-Granja, F., Mendoza, X. P. L., Jara, A. P., & Swaminathan, J. N. (2023). *Towards a Model of Information Audit in the Document Management of Public Institutions* (pp. 797-807). https://doi.org/10.1007/978-981-19-4960-9_60

Mejías, M., Guarate Coronado, Y. C., & Jiménez Peralta, A. L. (2022). Inteligencia artificial en el campo de la enfermería. Implicaciones en la asistencia, administración y educación. *Salud, Ciencia y Tecnología*, *2*, 88. https://doi.org/10.56294/saludcyt202288

Millán Naveas, G., & Vargas Guzmán, M. (2020). Un algoritmo de control de flujo para redes de computadoras de alta velocidad. *Ingeniare. Revista chilena de ingeniería*, *28*(1), 24-30. https://doi.org/10.4067/S0718-33052020000100024

Morze, N. V, & Strutynska, O. V. (2021). Digital transformation in society: key aspects for model development. *Journal of Physics: Conference Series*, *1946*(1), 012021. https://doi.org/10.1088/1742-6596/1946/1/012021

Moudgil, A., Singh, S., & Gautam, V. (2022). An Overview of Recent Trends in OCR Systems for Manuscripts. En J. M. R. S. Tavares, P. Dutta, S. Dutta, & D. Samanta (Eds.), *Cyber Intelligence and Information Retrieval. Lecture Notes in Networks and Systems* (pp. 525-533). Springer, Singapore. https://doi.org/10.1007/978-981-16-4284-5_46

Muñoz Soro, J. F., & Nogueras Iso, J. (2014). La digitalización de documentos en la Administración de Justicia. *Ibersid: revista de sistemas de información y documentación*, *8*, 49-53. https://doi.org/10.54886/ibersid.v8i0.4179

Nakaura, T., Higaki, T., Awai, K., Ikeda, O., & Yamashita, Y. (2020). A primer for understanding radiology articles about machine learning and deep learning. *Diagnostic and Interventional Imaging*, *101*(12), 765-770. https://doi.org/10.1016/j.diii.2020.10.001

Namysl, M., & Konya, I. (2019). Efficient, Lexicon-Free OCR using Deep Learning.

2019 International Conference on Document Analysis and Recognition (ICDAR), 295-301. https://doi.org/10.1109/ICDAR.2019.00055

Obukhov, A., Krasnyanskiy, M., & Nikolyukin, M. (2020). Algorithm of adaptation of electronic document management system based on machine learning technology. *Progress in Artificial Intelligence*, *9*(4), 287-303. https://doi.org/10.1007/s13748-020-00214-2

Pimienta, J., & de la Orden, A. (2017). *Metodología de la investigación (3ra ed.)*. Pearson Educación.

Promin, E., & Suriyachai, P. (2019). Improvement of Scanned Medical Document Management System. *2019 11th International Conference on Knowledge and Smart Technology (KST)*, 126-131. https://doi.org/10.1109/KST.2019.8687398

Sai Rakesh Kamisetty, V. N., Sohan Chidvilas, B., Revathy, S., Jeyanthi, P., Anu, V. M., & Mary Gladence, L. (2022). Digitization of Data from Invoice using OCR. *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, 1-10. https://doi.org/10.1109/ICCMC53470.2022.9754117

Salas-Tanchiva, C. C. (2022). Repercusión e importancia de la automatización del trámite documentario en las instituciones públicas. *Revista científica de sistemas e informática*, *2*(1), e266. https://doi.org/10.51252/rcsi.v2i1.266

Sosa del Angel, C. O., Caballero Rico, F. C., Guzmán García, J. C., & Perales Garza, C. Y. (2022). Gestión documental a través del Sistema Institucional de Archivos. Una aproximación desde el orden normativo mexicano. *Revista General de Información y Documentación*, *32*(1), 243-265. https://doi.org/10.5209/rgid.82947

Srivastava, S., Verma, A., & Sharma, S. (2022). Optical Character Recognition Techniques: A Review. *2022 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 1-6. https://doi.org/10.1109/SCEECS54111.2022.9740911

Sucari León, R., Aroquipa Durán, Y., Quina Quina, L. D., Quispe Yapo, E., Sucari León, A., & Huanca Torres, F. A. (2020). Visión artificial en reconocimiento de patrones para clasificación de frutas en agronegocios. *Puriq*, *2*(2), 109-118. https://doi.org/10.37073/puriq.2.2.76

Thorat, C., Bhat, A., Sawant, P., Bartakke, I., & Shirsath, S. (2022). A Detailed Review on Text Extraction Using Optical Character Recognition. En *Lecture Notes in*

Networks and Systems, vol 314 (pp. 719-728). Springer, Singapore. https://doi.org/10.1007/978-981-16-5655-2_69

Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Qi Dong, J., Fabian, N., & Haenlein, M. (2021). Digital transformation: A multidisciplinary reflection and research agenda. *Journal of Business Research*, *122*, 889-901. https://doi.org/10.1016/j.jbusres.2019.09.022

Viet Anh, P., Tung Khanh, N. D., Manh Dat, T., & Van Dan, P. (2022). Improved ocr quality for smart scanned document management system. *Journal of Science and Technique*, *9*(1), 51-67. https://doi.org/10.56651/lqdtu.jst.v9.n01.60.ict

Vrana, J., & Singh, R. (2021). Digitization, Digitalization, and Digital Transformation. En *Handbook of Nondestructive Evaluation 4.0* (pp. 1-17). Springer International Publishing. https://doi.org/10.1007/978-3-030-48200-8_39-1

Zaqueu, L. (2024). Challenges and opportunities for digital transformation in Mozambique's higher education institutions. *Revista Científica de Sistemas e Informática*, *4*(2), e690. https://doi.org/10.51252/rcsi.v4i2.690

ANEXOS

Anexo 1. Matriz de consistencia

<table>
<tr><td colspan="6">Título: Digitalización basada en reconocimiento óptico de caracteres para mejorar la disponibilidad de actas de evaluación en la UGEL San Martín</td></tr>
<tr><th>Problema</th><th>Objetivos</th><th>Hipótesis</th><th>Variable abstracta</th><th>Variable concreta</th><th>Escala</th></tr>
<tr><td rowspan="2">¿En qué medida la digitalización basada en reconocimiento óptico de caracteres mejora la disponibilidad de actas de evaluación en la UGEL San Martín?</td><td rowspan="2">General
Determinar en qué medida la digitalización basada en reconocimiento óptico de caracteres mejora la disponibilidad de actas de evaluación en la UGEL San Martín
Específicos
1. Evaluar la disponibilidad de actas de evaluación en la UGEL San Martín.
2. Digitalizar las actas de evaluación en la UGEL San Martín mediante técnicas de reconocimiento óptico de caracteres.
3. Medir la disponibilidad de actas de evaluación en la UGEL San Martín después de su digitalización basada en reconocimiento óptico de caracteres.</td><td rowspan="2">H₀: La digitalización basada en reconocimiento óptico de caracteres no mejora la disponibilidad de actas de evaluación en la UGEL San Martín.
H₁: La digitalización basada en reconocimiento óptico de caracteres mejora la disponibilidad de actas de evaluación en la UGEL San Martín.</td><td>Digitalización basada en reconocimiento óptico de caracteres</td><td>Tiempo de digitalización
Tasa de error de digitalización
Tasas de precisión</td><td>Cuantitativo-Discreto</td></tr>
<tr><td>Disponibilidad de actas de evaluación</td><td>Accesibilidad:
Tiempo de localización
Disponibilidad:
Facilidad de búsqueda
Seguridad de información
Almacenamiento
Eficiencia de labor administrativa
Eficiencia de atención de solicitudes</td><td>Cuantitativo-Discreto
Cualitativo-Ordinal</td></tr>
<tr><th>Tipo y diseño de</th><th>Población y muestra</th><th>Técnicas e instrumentos</th><th colspan="3">Estadística a utilizar</th></tr>
</table>

investigación			
Tipo: Aplicada **Enfoque:** Cuantitativo **Nivel:** Explicativo **Diseño:** Pre-experimental	**Población:** Conformada por dos grupos de análisis: i) las solicitudes de certificados de estudios, que representan el proceso para medir la dimensión de accesibilidad a las actas de evaluación en la UGEL San Martín; y ii) los miembros del personal de la institución, quienes aportarán su percepción para evaluar la dimensión de disponibilidad de las actas de evaluación. **Muestra:** La muestra se seleccionará por conveniencia (no probabilístico), tomando en consideración el total de solicitudes presentadas en un mes. Asimismo, la selección de los participantes se limitará a aquellos directamente involucrados en la gestión documentaria para la emisión de certificados de estudios. Este grupo de trabajadores está compuesto por un total de 11 individuos, distribuidos entre 4 en el área de Secretaría, 2 en Mesa de Partes, 2 en Dirección y 3 en el departamento de Informática.	**Técnicas:** - Observación - Encuesta **Instrumentos:** - Ficha de registro de datos - Cuestionario	**Estadística descriptiva:** Análisis descriptivo (media, desviación estándar, etc.) y análisis de frecuencias relativas y absolutas **Estadística inferencial:** Prueba de normalidad, Prueba de hipótesis, Prueba de Wilcoxon

Anexo 2. Instrumentos de recolección de datos

A. Ficha de registro de datos para medir la accesibilidad a actas de evaluación

Eventos*	Tiempo de localización de acta de evaluación
Solicitud N°1	
Solicitud N° 2	
Solicitud N° 3	
…	
…	
N solicitudes	

Nota: * Corresponde a las solicitudes de certificado de estudio en el que el personal de secretaría debe localizar las actas de evaluación archivados en los libros para comparar las notas y proceder con el visado respectivo.

B. Encuesta para medir la disponibilidad de actas de evaluación

Estimado colaborador de la Unidad de Gestión Educativa Local San Martín (UGEL),

Nos complace invitarte a participar en esta encuesta que tiene como objetivo conocer tus percepciones acerca de la disponibilidad de las actas de evaluación en la UGEL San Martín. Esta encuesta forma parte de un estudio de investigación que busca determinar el impacto de la digitalización de las actas de evaluación mediante técnicas de Reconocimiento Óptico de Caracteres (OCR) en la mejora de su disponibilidad y acceso. Tu opinión es de vital importancia para entender cómo las nuevas tecnologías pueden contribuir a agilizar los procesos internos y la eficiencia en la gestión de los documentos en nuestra institución. Esta encuesta se aplicará en dos momentos: antes de la implementación de la digitalización mediante OCR y después de su implementación, permitiéndonos comparar cualquier cambio en tus percepciones y/o experiencias.

Agradecemos de antemano tu participación. Tu colaboración nos ayudará a tomar decisiones informadas y a garantizar que las mejoras implementadas sean beneficiosas para todos los miembros de la UGEL San Martín. Asimismo, aclaramos que la encuesta es anónima y la información proporcionada se utilizará únicamente con fines de investigación. No hay respuestas correctas ni incorrectas, y tus opiniones son fundamentales para el éxito de este estudio. El tiempo estimado para completar la encuesta es de aproximadamente 10 minutos.

Por favor, lee los enunciados que se presentan a continuación y marque con una X el casillero que considere como respuesta.

Indicadores	Escala				
	Totalmente en desacuerdo	En desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo
1. Encontrar actas de evaluación es sencillo					
2. La ubicación de las actas de					

evaluación es organizada					
3. Las actas de evaluación están protegidas de daños (deterioro, plagas, etc.)					
4. Las actas de evaluación cuenta con respaldos (copias)					
5. Las actas se mantienen en condiciones óptima de conservación					
6. El espacio de almacenamiento de las actas de evaluación es óptimo					
7. Las solicitudes de certificación se atienden de manera rápida					

C. Ficha de registro de datos para medir la calidad de la digitalización

Digitalización de actas de evaluación	Tiempo de digitalización	Tasa de error de digitalización	Tasa de precisión
Digitalización N° 1			
Digitalización N° 2			
Digitalización N° 3			
…			
…			
N digitalizaciones			

Anexo 3. Base de datos

Evento	Tiempo de localización	
	Pretest	Postet
Solictud_1	793	140
Solictud_2	1019	140
Solictud_3	602	170
Solictud_4	747	63
Solictud_5	1091	66
Solictud_6	806	82
Solictud_7	916	152
Solictud_8	643	172
Solictud_9	1101	61
Solictud_10	1046	122
Solictud_11	989	145
Solictud_12	959	66
Solictud_13	779	92
Solictud_14	802	112
Solictud_15	858	82
Solictud_16	972	98
Solictud_17	1047	176
Solictud_18	905	125
Solictud_19	1153	126
Solictud_20	905	151
Solictud_21	1103	70
Solictud_22	667	120
Solictud_23	856	117
Solictud_24	809	127
Solictud_25	612	149
Solictud_26	1007	102
Solictud_27	1086	99
Solictud_28	1167	106
Solictud_29	862	176
Solictud_30	804	135

Pretest Disponibilidad								
Encuestados	Item_1	Item_2	Item_3	Item_4	Item_5	Item_6	Item_7	Suma
E1	4	3	4	4	4	3	3	25
E2	2	3	4	1	2	3	2	17
E3	4	4	3	4	4	4	3	26
E4	2	2	2	2	2	2	2	14
E5	2	3	4	4	3	3	2	21
E6	4	5	1	2	4	2	2	20
E7	5	5	2	2	2	2	3	21

E8	4	3	1	2	3	1	4	18
E9	5	5	1	1	2	2	4	20
E10	5	4	2	1	2	3	3	20
E11	4	4	4	1	4	2	3	22

Postest Disponibilidad								
Encuestados	**Item_1**	**Item_2**	**Item_3**	**Item_4**	**Item_5**	**Item_6**	**Item_7**	**Suma**
E1	4	5	4	4	5	4	5	31
E2	4	4	4	5	4	5	5	31
E3	5	4	5	5	4	4	4	31
E4	4	4	4	4	4	4	5	29
E5	4	4	5	4	5	4	5	31
E6	4	5	4	5	5	4	4	31
E7	5	4	5	4	4	5	4	31
E8	5	4	4	4	4	5	5	31
E9	4	4	4	4	4	4	3	27
E10	4	5	4	5	4	4	5	31
E11	4	4	5	4	5	5	4	31

Digitalización de actas de evaluación	**Tiempo de digitalización**	**Tasa de error de digitalización**	**Tasa de precisión**
Digitalización_1	104	6.09	93.91
Digitalización_2	110	3.14	96.86
Digitalización_3	128	13.75	86.25
Digitalización_4	92	5.50	94.50
Digitalización_5	74	40.28	59.72
Digitalización_6	104	2.16	97.84
Digitalización_7	141	27.83	72.17
Digitalización_8	112	14.18	85.82
Digitalización_9	130	16.82	83.18
Digitalización_10	154	27.46	72.54
Digitalización_11	140	4.41	96.89
Digitalización_12	88	7.23	94.35
Digitalización_13	92	9.10	91.51
Digitalización_14	138	8.82	96.08
Digitalización_15	132	7.53	97.84
Digitalización_16	101	8.10	93.32
Digitalización_17	117	19.77	96.51
Digitalización_18	90	8.60	91.05

Digitalización_19	125	5.70	98.51
Digitalización_20	119	17.30	97.64
Digitalización_21	122	8.20	91.20
Digitalización_22	81	9.50	96.15
Digitalización_23	89	6.70	95.78
Digitalización_24	104	16.91	90.16
Digitalización_25	125	4.62	93.26
Digitalización_26	85	4.40	93.26
Digitalización_27	134	5.60	94.25
Digitalización_28	117	6.77	90.01
Digitalización_29	107	14.98	97.89
Digitalización_30	114	8.46	95.56

Anexo 4. Solicitud de permiso y autorización para la ejecución del estudio

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

"Año del Bicentenario, de la consolidación de nuestra Independencia, y de la conmemoración de las heroicas batallas de Junin y Ayacucho"

CARTA N° 01-2024/ALMV-CDPT

Para: Dr. Wildoro Pinchi Daza
Director UGEL-San Martín

De: Bach. Aranza Luccia Marcelo Vasquez
Bach. César David Paredes Torres

Asunto: Solicita autorización para realizar trabajo de investigación en la UGEL-San Martín

Grato es saludarlo cordialmente y al mismo sirva la presente para solicitarle autorización para ejecutar el proyecto de investigación denominado "Digitalización basada en reconocimiento óptico de caracteres para mejorar la disponibilidad de actas de evaluación en la UGEL San Martín" presentados por los autores que suscriben esta carta con el objetivo de obtener el grado de Ingeniero de Sistemas e Informática en la Universidad Nacional de San Martín (Tarapoto).

Cabe recalcar, que nuestra intervención respetará los principios éticos de la investigación científica, teniendo en cuenta el consentimiento informado de los participantes y el anonimato de los datos recopilados que serán utilizados solo para los fines del estudio. Asimismo, la intervención consiste en aplicar técnicas de inteligencia artificial como visión por computadora y procesamiento de lenguaje natural para garantizar la disponibilidad y accesibilidad de las actas de evaluación que actualmente se encuentren expuestos a pérdidas o pueden perder su calidad.

En este sentido, adjuntamos la resolución de aprobación del proyecto de investigación y el proyecto como tal.

A la espera de vuestra atención oportuna y deseándole éxitos en su gestión.

Nos despedimos,

Atentamente.

Bach. Aranza Luccia Marcelo Vasquez
DNI N°: 71597049

Bach. César David Paredes Torres
DNI N°: 70194642

UNIDAD EJECUTORA 301 - BAJO MAYO
UNIDAD DE GESTIÓN EDUCATIVA LOCAL
SAN MARTÍN - TARAPOTO
Secretaría General - Mesa de Partes
RECEPCIÓN
Reg: N° 008 - 2024 - 457235
Fecha: 22 FEB. 2024 Firma

GOBIERNO REGIONAL SAN MARTÍN

DIRECCIÓN REGIONAL DE EDUCACIÓN

UGEL SAN MARTIN

"Año del Bicentenario, de la consolidación de nuestra Independencia, y de la conmemoración de las heroicas batallas de Junín y Ayacucho"

Tarapoto, 27 de febrero de 2024.

CARTA N° 0007-2024-GRSM-DRE/UGELSM-JOA

Señorita: **ARANZA LUCCIA MARCELO VÁSQUEZ**

CIUDAD

Asunto: AUTORIZA REALIZAR TRABAJO

REF.: EXPEDIENTE (GS 008-2024 **457235** DE FECHA 22-02-2024)

Por la presente me dirijo a usted expresándole el saludo cordial en representación de la Oficina de Administración, al mismo tiempo y en atención a lo solicitado con el documento de la referencia, se le **AUTORIZA realizar el trabajo de investigación** en la Unidad de Gestión Educativa Local San Martín, para ejecutar el proyecto de investigación denominado: **"Digitalización basada en reconocimiento óptico de caracteres para mejorar la disponibilidad de actas de evaluación".**

Sin otro particular, me suscribo de usted.

Atentamente,

GOBIERNO REGIONAL DE SAN MARTÍN
UNIDAD DE GESTIÓN EDUCATIVA LOCAL SAN MARTÍN
Unidad Ejecutora N° 301 - SAN MARTÍN

OFICINA DE ADMINISTRACIÓN U.E. 301 - SAN MARTÍN

.....
CPCC. Edwin Antonio Alvarado Romero (e)
JEFE DE LA OFICINA DE ADMINISTRACIÓN

EAAR/JOA
Crd/sec

Anexo 5. Evidencias fotográficas de la recolección de datos e implementación de la solución tecnológica

Anexo 6. Estructura de ficheros del sistema Django

La estructura de ficheros del sistema Django sigue una estructura predeterminada. A continuación, se describe la estructura de ficheros del proyecto:

- **main**: Contiene los módulos específicos de la aplicación, como las vistas, modelos y funciones personalizadas.
    - **migrations**: Almacena los archivos de migración de la base de datos.
    - **templates**: Contiene las plantillas HTML utilizadas en la interfaz de usuario.
    - **__pycache__**: Carpeta generada automáticamente que almacena archivos compilados de Python.
    - **admin.py**: Archivo para la configuración del panel de administración de Django.
    - **apps.py**: Archivo de configuración de la aplicación.
    - **extraccion.py**: Contiene funciones específicas para la extracción de datos.
    - **funciones.py**: Contiene funciones auxiliares utilizadas en el procesamiento de imágenes y OCR.
    - **models.py**: Define los modelos de datos utilizados en la aplicación.
    - **tests.py**: Contiene pruebas automatizadas para la aplicación.
    - **views.py**: Define las vistas que manejan las solicitudes HTTP y renderizan las respuestas.
- **ocr_project**: Contiene la configuración global del proyecto Django.
    - **__pycache__**: Carpeta generada automáticamente que almacena archivos compilados de Python.
    - **asgi.py**: Archivo de configuración para el servidor ASGI.
    - **settings.py**: Archivo de configuración global del proyecto.
    - **urls.py**: Define las rutas del proyecto.
    - **wsgi.py**: Archivo de configuración para el servidor WSGI.
- **static_files**: Carpeta para almacenar archivos estáticos como CSS, JavaScript e imágenes.
- **db.sqlite3**: Archivo de la base de datos SQLite utilizado durante el desarrollo.
- **manage.py**: Script de línea de comandos para interactuar con el proyecto Django.

Digitalización basada en reconocimiento óptico de caracteres para mejorar la disponibilidad de actas de evaluación en la UGEL San Martín

por CÉSAR DAVID PAREDES TORRES

Fecha de entrega: 21-feb-2025 08:09a.m. (UTC-0500)
Identificador de la entrega: 2594633040
Nombre del archivo: TESIS_CESAR_PAREDES_ARANZA_LUCIA_20.02.2025_5_.docx (5.21M)
Total de palabras: 17021
Total de caracteres: 101328

Digitalización basada en reconocimiento óptico de caracteres para mejorar la disponibilidad de actas de evaluación en la UGEL San Martín

INFORME DE ORIGINALIDAD

17%	16%	5%	9%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	tesis.unsm.edu.pe Fuente de Internet	4%
2	repositorio.unsm.edu.pe Fuente de Internet	2%
3	repositorio.ucv.edu.pe Fuente de Internet	1%
4	Submitted to Universidad Nacional de San Martín Trabajo del estudiante	1%
5	www.coursehero.com Fuente de Internet	1%
6	hdl.handle.net Fuente de Internet	1%
7	alicia.concytec.gob.pe Fuente de Internet	<1%
8	dominiodelasciencias.com Fuente de Internet	<1%