

Análisis, interpretación, clasificación y búsqueda automática de la gestión documental

Jorge Luis Nieto Córdoba

Asesor

Julio Eduardo Mejía Manzano

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Nombre Director de Trabajo de Grado

Julio Eduardo Mejía Manzano

Jurado

Jurado

Resumen

El presente estudio lleva a cabo una revisión sistemática de la literatura sobre la aplicación de modelos de inteligencia artificial (IA) en la gestión documental abordando los procesos de análisis, interpretación, clasificación y búsqueda automática de documentos. La investigación se llevó a cabo siguiendo la metodología PRISMA y se analizaron estudios publicados entre 2021 y 2025 en bases de datos como IEEE Xplore, Scopus y Google Scholar.

Los resultados ponen de evidencia que técnicas como el aprendizaje automático, el aprendizaje profundo, el procesamiento del lenguaje natural (PLN), el reconocimiento óptico de caracteres (OCR), los modelos de lenguaje preentrenados (PLM) y los chatbots mejoran significativamente la precisión, la eficiencia y la trazabilidad en la gestión documental. Entre los modelos evaluados, destacan algoritmos como SVM, Naive Bayes, Random Forest, K-NN, MLP y redes neuronales, con métricas de rendimiento superiores al 80 % e incluso al 98 % en tareas específicas de clasificación. En particular, en estudios como los de Hassan et al. (2022) y Zhang et al. (2024) se observó que los algoritmos Naive Bayes y K-NN alcanzaban una precisión del 94 % al 98,5 %, mientras que modelos como SVM y regresión logística obtenían altos niveles de exactitud en conjuntos equilibrados.

El análisis metodológico realizado con el instrumento QAREL permitió determinar que los estudios con mejores prácticas incluyen validación cruzada, separación de datos, métricas robustas (precisión, recall y F1-score) y replicabilidad experimental. No obstante, siguen existiendo desafíos relacionados con la calidad y estandarización de los datos, la interoperabilidad entre sistemas, la integración tecnológica y la formación del personal.

En conclusión, la IA es una herramienta clave para modernizar y automatizar la gestión documental, ya que ofrece soluciones escalables que reducen los errores, agilizan los procesos y fortalecen el cumplimiento normativo en diversos entornos organizativos.

Palabras clave: Inteligencia Artificial, Gestión Documental, Clasificación Automática, Interpretación Semántica, Modelos de Clasificación, Aprendizaje Automático, Procesamiento del Lenguaje Natural (PLN), Automatización Inteligente, Big Data, Chatbots.

Abstract

This study conducts a systematic review of the literature on the application of artificial intelligence (AI) models in document management, addressing the processes of analysis, interpretation, classification, and automatic document search. The research was carried out following the PRISMA methodology, and studies published between 2021 and 2025 in databases such as IEEE Xplore, Scopus, and Google Scholar were analyzed.

The results show that techniques such as machine learning, deep learning, natural language processing (NLP), optical character recognition (OCR), pre-trained language models (PLM), and chatbots significantly improve accuracy, efficiency, and traceability in document management. Among the models evaluated, algorithms such as SVM, Naive Bayes, Random Forest, K-NN, MLP, and neural networks stand out, with performance metrics above 80% and even 98% in specific classification tasks. In particular, studies such as those by Hassan et al. (2022) and Zhang et al. (2024) found that Naive Bayes and K-NN algorithms achieved accuracy of 94% to 98.5%, while models such as SVM and logistic regression achieved high levels of accuracy in balanced sets.

The methodological analysis carried out with the QAREL tool determined that best practice studies include cross-validation, data separation, robust metrics (precision, recall, and F1-score), and experimental replicability. However, challenges remain in relation to data quality and standardization, interoperability between systems, technological integration, and staff training.

In conclusion, AI is a key tool for modernizing and automating document management, as it offers scalable solutions that reduce errors, streamline processes, and strengthen regulatory compliance in various organizational environments.

Keywords: Artificial Intelligence, Document Management, Automatic Classification, Semantic Interpretation, Classification Models, Machine Learning, Natural Language Processing (NLP), Intelligent Automation, Big Data, Chatbots.

Tabla de Contenido

Introducción	10
Planteamiento del Problema	12
Justificación	13
Objetivos	15
Objetivo General.....	15
Objetivos Específicos	15
Marco de Referencia	16
Marco Teórico y Conceptual	22
Metodología	24
Ecuaciones Booleanas	24
Lista de Verificación de Evaluación de Calidad de Confiabilidad Diagnóstica (QAREL).....	28
Resultados	33
Conclusiones	43
Recomendaciones	45
Referencias Bibliográficas	48

Lista de Tablas

Tabla 1 <i>Ecuaciones Booleanas</i>	25
Tabla 2 <i>Resultados del Instrumento QAREL por Estudios Analizados</i>	29
Tabla 3 <i>Artículos Evaluados</i>	33

Lista de Figuras

Figura 1 <i>Ecuación Booleana</i>	26
Figura 2 <i>Diagrama de Flujo Prisma</i>	27
Figura 3 <i>Gráfico de Frecuencia del Instrumento QAREL</i>	38

Introducción

La gestión documental es un área fundamental en las entidades o empresas modernas, fundamentalmente en entornos donde el manejo eficiente de grandes volúmenes de información tiene una influencia en la toma de decisiones, la trazabilidad y el cumplimiento de las normas. De esta forma se argumenta que la automatización mediante tecnologías de inteligencia artificial se presenta como una alternativa eficaz para optimizar eficientemente la operabilidad y la precisión en el tratamiento de documentos. Aldás. (2022)

Esta monografía se centra en la realización de una revisión sistemática de la información en torno al tema de modelos de inteligencia artificial para el análisis, la interpretación, la clasificación y búsqueda automática de documentos en la gestión documental.

Para la búsqueda de información, se utilizará la metodología PRISMA, que se empleará en esta monografía para realizar una revisión sistemática de la literatura entre 2021 y 2025 en bases de datos como IEEE Xplore, Scopus, Google Scholar y repositorios institucionales, con el objetivo de identificar y clasificar los retos técnicos y operativos asociados al uso de modelos de inteligencia artificial para la clasificación documental. Para ello, se explorarán enfoques basados en el análisis estructural de documentos, teniendo en cuenta elementos como los metadatos, el formato, el tipo de archivo y otras características.

Con esta propuesta, se busca es dar a conocer los principales modelos de inteligencia artificial las características y sus dificultades para ser aplicados en los procesos de gestión documental mediante soluciones tecnológicas accesibles y replicables, también sentar las bases para futuras aplicaciones de inteligencia artificial en entornos organizativos donde la eficiencia documental es una prioridad estratégica Garcia-Chicangana et al.(2021)

Como lo menciona Yovera Mogollon & Zavaleta Ramos (2024) no resolver adecuadamente los desafíos relacionados con la clasificación y el procesamiento automático de documentos puede tener múltiples consecuencias negativas para las organizaciones, como la pérdida de eficiencia operativa, errores humanos en la clasificación, costes operativos elevados, dificultades para cumplir la normativa y la pérdida o duplicidad de información.

Para mitigar estas consecuencias, se proponen las siguientes líneas de actuación, enfocadas en estudiar modelos de inteligencia artificial: automatización basada en aprendizaje automático, consulta de metadatos y reglas de negocio, integración con sistemas de gestión documental existentes.

Planteamiento del Problema

Según Yovera Mogollon & Zavaleta Ramos. (2024) en muchas organizaciones se maneja una gran cantidad de documentos cada día. Sin embargo, el proceso de organizarlos y clasificarlos correctamente suele hacerse de forma manual, lo que genera problemas como errores humanos, pérdida de información, duplicación de archivos y un gran gasto de tiempo y recursos. Frente a esta problemática, la inteligencia artificial (IA) surge como una solución innovadora. Estudios anteriores han examinado el uso de tecnologías de inteligencia artificial, tales como el reconocimiento de caracteres mediante tecnología óptica (OCR) y el análisis de lenguaje natural (PLN).

Dogra et al. (2022) menciona que estos problemas afectan a la eficiencia de las empresas y dificultan el cumplimiento de las normas y procesos internos. Aunque hoy en día existen herramientas basadas en inteligencia artificial que podrían ayudar a mejorar este proceso, su implementación no es sencilla. Hay dificultades técnicas relacionadas con la forma en que están escritos los documentos o cómo están organizados. El estudio de Dogra et al. (2022) introduce un marco unificado para evaluar el impacto de la incrustación de palabras y la unidad recurrente con compuerta (GRU) en la clasificación de texto para superar esta falla.

Por eso, esta investigación se centra en la siguiente pregunta principal:

¿Cómo se puede realizar el estudio de modelos que utilice inteligencia artificial para analizar, interpretar, clasificar y buscar automáticamente los documentos de manera más rápida y precisa que los métodos tradicionales de la gestión documental?

Justificación

En la actualidad, las organizaciones necesitan gestionar grandes volúmenes de información documental de manera eficaz. Según Yovera Mogollon & Zavaleta Ramos. (2024) Este reto se intensifica en contextos donde la precisión, la trazabilidad y el cumplimiento normativo son de gran importancia para el desarrollo de las actividades institucionales y la adopción de decisiones estratégicas. Frente a este panorama, la automatización de procesos mediante el aprendizaje automático (machine learning) se muestra como una solución innovadora capaz de optimizar la gestión documental, reducir los tiempos de procesamiento, minimizar los errores humanos y garantizar una clasificación coherente y estructurada de los documentos.

A partir de esta exploración se busca identificar modelos de IA centrados en el análisis, la interpretación, la clasificación y búsqueda automática de documentos. Su pertinencia radica en realizar los estudios de un potencial para transformar procesos manuales, costosos y propensos a errores en flujos de trabajo ágiles, automatizados y fiables. Además, responder a la necesidad de contar con herramientas tecnológicas accesibles, replicables y adaptables a distintos entornos organizativos, lo que supone un valor añadido en contextos donde los recursos pueden ser limitados Calixto & Colón. (2024)

La justificación se basa también en la relevancia de abordar los desafíos técnicos y operativos que conlleva la implementación de modelos de aprendizaje automático en la clasificación.

Los beneficios esperados de esta monografía son el mejoramiento operativo, la reducción de costes, el cumplimiento de las normas y la adecuada organización de la información, lo que refuerza la necesidad de llevar a cabo esta investigación. Ignorar estos desafíos podría implicar

pérdidas significativas para las organizaciones, lo que justifica plenamente la inversión en el progreso de las soluciones apoyadas en inteligencia artificial.

Objetivos

Objetivo General

Estudiar diferentes modelos de inteligencia artificial que permitan el análisis, la interpretación, la clasificación y la búsqueda automática en el contexto de la gestión documental, mediante la integración de técnicas de análisis estructural de documentos y aprendizaje automático, haciendo uso de la metodología PRISMA.

Objetivos Específicos

Analizar las diferentes técnicas de enfoques basados en el análisis estructural de documentos y en el aprendizaje automático para mejorar la interpretación y la estructura de los documentos.

Identificar los principales desafíos en la implementación de modelos de inteligencia artificial para la clasificación de la gestión documental, teniendo en cuenta aspectos técnicos, lingüísticos y operativos.

Discutir las mejores técnicas y enfoques basados en el análisis estructural de documentos y aprendizaje automático para la mejora de la interpretación y estructura de los documentos.

Marco de Referencia

Según Baviskar et al., (2021). El manejo inadecuado de los datos no estructurados genera un gran coste para la mayoría de las organizaciones, aunque su correcta gestión podría aumentar notablemente su eficiencia. Las técnicas tradicionales de extracción de información presentan limitaciones, mientras que los métodos basados en inteligencia artificial (IA) ofrecen mejores perspectivas, aunque aún están poco explorados en la literatura científica. Menciona que analizaron investigaciones entre 2010 y 2020 sobre técnicas de extracción automática de información y destaca que las soluciones actuales se basan en reglas fijas no se adaptan bien a documentos complejos y carecen de datos de calidad. Por esta razón propone desarrollar conjuntos de datos más representativos y fomentar la colaboración entre el ámbito académico y empresarial para superar estos desafíos.

El proyecto técnico de Aldás. (2022). Menciona la necesidad de un software eficiente que optimice las necesidades de la organización y, de esta forma, elimine los errores que se cometen manualmente al clasificar, redactar y recuperar información.

En la investigación de Ferrer Tarazona & Rubiano Pulido. (2024). Se analizan los principales desafíos del procesamiento de textos automatizado y se proponen soluciones basadas en IA y en el aprendizaje automático. También se menciona la importancia del preprocesamiento de datos y del uso de algoritmos para optimizar los procedimientos, lo que reduce los errores y mejora la fiabilidad en la categorización de documentos.

De acuerdo con la investigación de Salavarría & Bazurto. (2024) la inteligencia artificial ha pasado de sistemas simples a tecnologías avanzadas como el aprendizaje automático, permitiendo a las máquinas simular capacidades humanas. En la gestión empresarial actual, la IA aporta beneficios como la automatización de tareas, decisiones basadas en datos, aumento de la

productividad y experiencias personalizadas. No obstante, también implica desafíos, como altos costos, posibles afectaciones laborales y preocupaciones sobre privacidad y ética. Para su implementación efectiva, es clave diseñar programas de formación que respondan a las necesidades específicas del personal administrativo.

Menciona Dogra et al., (2022) que debido al crecimiento acelerado de la información digital, especialmente en la documentación, la minería de textos y el seguimiento en línea se han vuelto esenciales. En los últimos años se han desarrollado diferentes enfoques de aprendizaje automático y profundo para abordar problemas complejos del procesamiento del lenguaje natural, como la clasificación de textos. Este artículo revisa dichos métodos, analizando sus fortalezas y limitaciones, y ofrece una visión útil para quienes buscan avanzar o innovar en este campo.

Según Viloría Ortín. (2023). Uno de los principales retos es la redacción de los documentos y correos, ya que están escritos de manera no muy formal y con una estructura diferente, por lo que el sistema no los interpreta correctamente. Esto provoca una mala clasificación y organización de la información.

El artículo de Sambetbayeva et al., (2022), menciona el crecimiento del volumen documental y la digitalización han impulsado la necesidad de contar con sistemas electrónicos más eficientes para gestionar documentos. El artículo plantea un modelo inteligente de gestión documental basado en aprendizaje automático y modelado multiagente, capaz de optimizar el procesamiento, la recuperación y el ciclo de vida de los documentos electrónicos. El modelo ha demostrado su aplicabilidad en distintos entornos organizativos, mejorando la eficiencia del trabajo con grandes volúmenes de información.

Díaz González. (2022). plantea implementar tecnologías avanzadas y gestores documentales como Adeanet, que indexa los documentos automáticamente y les asigna una identificación según su contenido. Estas herramientas son muy útiles para la automatización y clasificación de la correspondencia, ya que permiten apoyarse en el lenguaje automático y en algoritmos para realizar una mejor clasificación.

La clasificación de texto es fundamental en el procesamiento del lenguaje natural, Hassan et al., (2022), esto permite organizar automáticamente la información en categorías predefinidas, lo que facilita tareas como el filtrado de spam o la toma de decisiones. El estudio realiza la comparación del rendimiento de distintos algoritmos de aprendizaje automático como Máquinas Vectoriales de Soporte (SVM), k-NN, regresión logística, Bayesiano Naive Multinomial (MNB) y bosque aleatorio en dos conjuntos de datos diferentes. La evaluación se basa en métricas como la precisión, la recuperación y el F1. Se puede demostrar que la regresión logística tiene una precisión del 85,8 %y el algoritmo Máquinas Vectoriales de Soporte (SVM), destacan en el conjunto Internet Movie Database (IMDB), mientras que el k-NN obtiene mejores resultados con una precisión del 98,5 % en el conjunto SPAM.

Quispe Reyes. (2023). ha desarrollado un bot en Python que obtiene la información de forma automática, lo que permite realizar un análisis estadístico para medir la eficiencia de las tablas y gráficos de Excel y Python, lo que mejora y optimiza los tiempos de los documentos.

La investigación de Mustafa Abdullah & Mohsin Abdulazeez. (2021), menciona el crecimiento tecnológico y el aumento de la reproducción de datos han generado la necesidad de procesarlos con mayor rapidez y precisión. En este contexto el aprendizaje automático especialmente el algoritmo de Máquinas Vectoriales de Soporte (SVM), demuestra ser una herramienta eficiente para la clasificación de datos en múltiples áreas. El artículo revisa diversas

aplicaciones del clasificador SVM, como el reconocimiento facial, el diagnóstico médico, el análisis de texto y la seguridad en redes. Los resultados evidencian que la precisión del SVM fue del 97,47%, especialmente en el reconocimiento facial y la detección de enfermedades.

El artículo de Abualigah et al., (2021), analiza en profundidad el uso de algoritmos metaheurísticos en el agrupamiento de textos, resaltando sus métodos más importantes y su eficacia en el campo del aprendizaje automático. Examina diversas variantes de estos algoritmos, incluidos los enfoques básicos, híbridos y multiobjetivo, y evalúa críticamente sus fortalezas, limitaciones y posibilidades de cara a futuras investigaciones.

Dahl et al., (2021), propone la implementación de técnicas de aprendizaje automático para automatizar la digitalización de datos. Se presentan dos casos: uno que emplea clasificación no supervisada para analizar diarios de enfermería, y otro que utiliza redes neuronales con atención para transcribir datos de certificados de defunción daneses. Ambos ejemplos ilustran cómo estas herramientas pueden facilitar y mejorar la recopilación de datos históricos.

Según Yovera Mogollon & Zavaleta. (Ramos, 2024). La siguiente herramienta utiliza la visualización óptica de caracteres (OCR), que convierte el texto impreso o manuscrito en datos digitales, y modelos de aprendizaje profundo como GPT-4. Este último destaca por la forma de entrenar modelos personalizados para adaptarse a las necesidades de las organizaciones y optimizar la gestión documental.

Los estudios de Miric et al., (2023), exploran métodos de aprendizaje automático y procesamiento del lenguaje natural para clasificar y extraer información de estos textos. El estudio muestra cómo aplicar estas herramientas para identificar tecnologías de inteligencia

artificial en patentes, y compara los métodos modernos con los enfoques tradicionales, destacando sus ventajas.

Según Lin & Zou.(2024) ChatGPT mejora la gestión de archivos electrónicos al automatizar la clasificación, el resumen, la búsqueda y el análisis de documentos, lo que aumenta la eficiencia. Sin embargo, plantea riesgos en materia de privacidad, aspectos legales y sostenibilidad técnica. En el futuro, su uso requerirá regulación, integración con otras tecnologías y formación del personal para poder aplicarlo de forma segura y efectiva.

El documento de Qin.(2025) propone diseñar un sistema inteligente para gestionar archivos en inglés mediante aprendizaje automático (CNN, RNN, BERT, etc.) con el fin de mejorar la clasificación, la búsqueda y la eficiencia frente a los métodos manuales tradicionales.

El proyecto de Zhang et al., (2024) propone un sistema automático que utiliza el algoritmo Naive Bayes para clasificar y archivar documentos sobre reformas educativas, con el que se sustituye la clasificación manual para mejorar la eficiencia y la precisión mediante la recopilación de datos, el preprocesamiento de texto, el entrenamiento de modelos y el diseño de software. Presenta una revisión de métodos para buscar y organizar archivos según su contenido, y no solo por su nombre.

En el estudio de Abirami.(2023) se analizan técnicas de aprendizaje automático, procesamiento del lenguaje natural y extracción de palabras clave aplicadas a documentos estructurados, semiestructurados y no estructurados.

Finalmente, propone un sistema inteligente que permite etiquetar, buscar y organizar automáticamente los archivos en carpetas mediante palabras clave y similitud.

El artículo de Di Marzo Serugendo et al., (2024) propone un sistema inteligente de gestión documental para automatizar el manejo de documentos fiscales y administrativos. Para

ello, utiliza inteligencia artificial para la clasificación y extracción de información, ontologías, RDF, SHACL, esquemas JSON, mapeo semántico y un motor de razonamiento, lo que permite clasificar documentos, extraer datos, validar normas fiscales y mejorar la eficiencia empresarial.

Marco Teórico y Conceptual

La gestión documental tiene unos lineamientos para los documentos en los cuales se establece el ciclo de vida desde su creación hasta su conservación o eliminación.

De acuerdo con el programa de gestión documental del Archivo General de la Nación, se busca una mejora continua en la administración institucional mediante la optimización, simplificación y racionalización de los procesos mediante el buen uso de las tecnologías, con la implementación de mecanismos de planeación, gestión, control, protección, preservación y garantía del derecho de acceso a los documentos respaldados por la normativa y estándares correspondientes. *Publicaciones | Archivo General de la Nación. (2025).*

La optimización desempeña un papel fundamental en la gestión documental, ya que permite desarrollar y aplicar modelos de inteligencia artificial orientados al análisis, interpretación, clasificación y búsqueda automática de documentos.

Una de las herramientas estudiadas para ello es el lenguaje estándar el cual permite interactuar con bases de datos relacionales SQL. Para modelarlo, es necesario convertir las preguntas en lenguaje natural de las personas a SQL. En la actualidad, este proceso sigue siendo un desafío debido a la complejidad que entraña escribir y redactar preguntas, analizar la estructuración de las bases de datos y generar SQL.

Otra herramienta son los modelos de aprendizaje profundo (LLM), que muestran capacidades reveladoras en cuanto a la comprensión del lenguaje natural. A medida que se integra más la tecnología LLM, se ofrecen mejores opciones para la conversión de textos a SQL. Hong et al., (2025).

Con los algoritmos KNN (K-Nearest Neighbour) no paramétricos de clasificación supervisada, que mediante la proximidad ejecutan predicciones, clasificaciones y regresiones en

un conjunto de datos individuales. Es uno de los clasificadores de datos más reconocidos que utilizan machine learning y que ayuda a clasificar los datos de forma óptima.

El propósito del algoritmo de k vecinos más cercanos es calcular la distancia de un punto de consulta determinado para poder asignarle una etiqueta de clase y establecer un promedio, también basado en el principio de similitud de los datos. (*IBM - México*, s. f.).

El árbol de decisión, tal y como lo sintetiza Arana. (2021), es un modelo de aprendizaje supervisado no paramétrico que se implementa para clasificar y predecir datos. Su principal objetivo es predecir resultados más específicos. Posee una estructuración jerárquica de árbol, que contiene un nodo raíz el cual se divide para función de la variable independiente más notable. Las ramas y los nodos se dividen de nuevo a razón la variable con una menor suma del valor cuadrático de los errores como nodo de decisión interno. Los nodos que quedan de últimos se conocen como hojas.

El chatbot es un asistente virtual cuyo principal objetivo es tener una interfaz de usuario intuitiva, citar sus fuentes, ser compatible con diferentes formatos (como PDF, archivos de Word, hojas de Excel y archivos de texto sin formato) y generar sus respuestas en varios idiomas. Zürcher.(2025).

Metodología

Siguiendo los lineamientos de la metodología PRISMA, se llevó a cabo una exhaustiva revisión bibliográfica para desarrollar esta monografía, que aborda los métodos de análisis, interpretación, clasificación y búsqueda automática en la gestión documental. En este contexto, se exploran diversos modelos, entre los que destaca el machine learning, el deep learning, el procesamiento del lenguaje natural (PLN), el big data y los chatbots para optimizar la clasificación, la organización y la automatización de documentos.

Se examinaron diversas estrategias de búsqueda aplicando la metodología PRISMA. Para ello, se utilizaron distintos términos de búsqueda en varias bases de datos especializadas, como IEEE Xplore, Scopus y Google Scholar, que contienen una gran cantidad de información relevante. Se utilizó la ecuación booleana con términos como «gestión de documentos», «clasificación de textos», «análisis semántico» e «recuperación inteligente», combinándolos con operadores booleanos (AND, OR) para mejorar la especificidad y la sensibilidad de las búsquedas. La búsqueda de información se limitó a los años 2021-2025 para garantizar el acceso a los datos más recientes y actualizados.

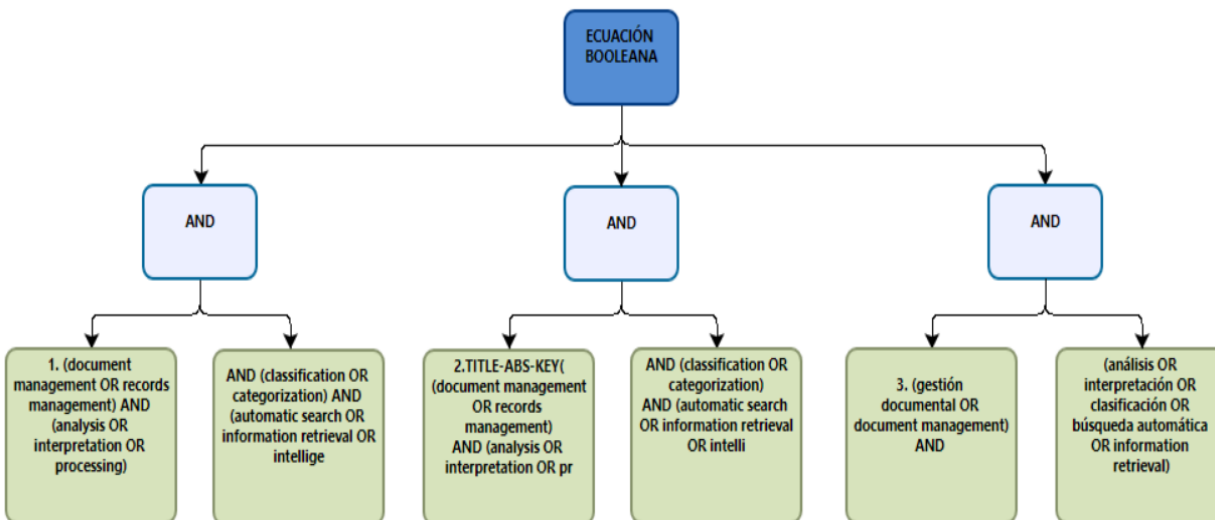
Ecuaciones Booleanas

Las ecuaciones de búsqueda utilizadas en IEEE Xplore, Scopus y Google Scholar permitieron recuperar estudios relevantes sobre la aplicación de la inteligencia artificial en la gestión documental, lo que optimizó la precisión en la selección de fuentes y garantizó la pertinencia académica de la información utilizada en la monografía.

Tabla 1*Ecuaciones Booleanas*

Bases de Datos	Ecuaciones
IEEE Xplore	("document management" OR "records management") AND (analysis OR interpretation OR processing) AND (classification OR categorization) AND ("automatic search" OR "information retrieval" OR "intelligent search")
Scopus	TITLE-ABS-KEY(("document management" OR "records management") AND (analysis OR interpretation OR processing) AND (classification OR categorization) AND ("automatic search" OR "information retrieval" OR "intelligent search"))
Google Scholar	("gestión documental" OR "document management") AND (análisis OR interpretación OR clasificación OR "búsqueda automática" OR "information retrieval")

Los resultados muestran una tendencia hacia la creación de sistemas híbridos que combinan PNL, ML y bases de datos inteligentes (SQL, RDF) para optimizar la recuperación y clasificación de documentos.

Figura 1*Ecuación Booleana*

Bases de datos

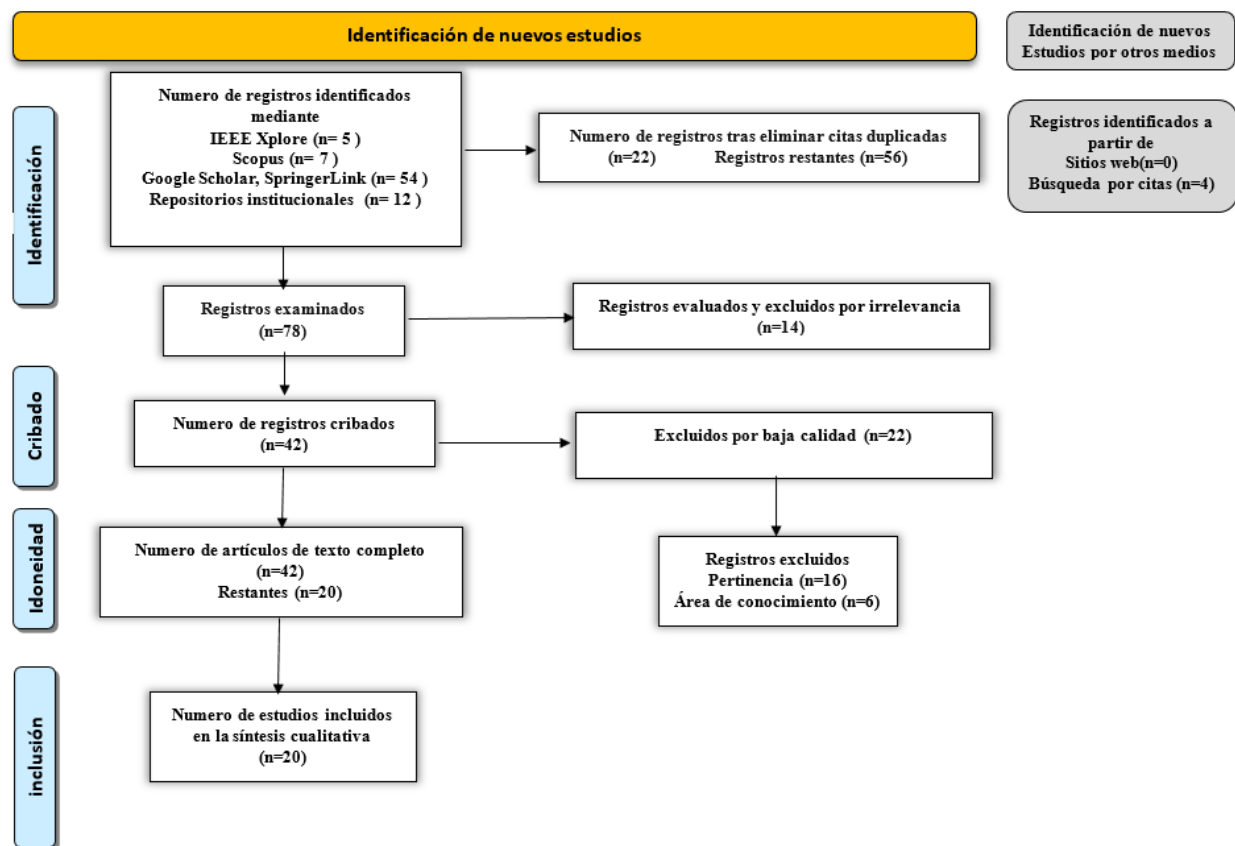
1. IEEE Xplore.
2. Scopus.
3. Google Scholar.

Este es el método de búsqueda empleado para la recopilación de artículos de diversas revistas científicas y estudios se centró en diferentes modelos, destacando el uso de machine learning para el análisis, interpretación, clasificación y búsqueda de documentos.

Estos criterios permitieron filtrar la información y enfocarse en los contenidos más relevantes para la investigación.

Figura 2

Diagrama de Flujo Prisma



De acuerdo con la metodología PRISMA, se elaboró un diagrama de flujo con cuatro fases: identificación, cribado, idoneidad, e inclusión.

En la fase de identificación, la búsqueda inicial en las bases de datos generó un total de 78 registros. Tras la revisión, se eliminaron 22 registros duplicados y se procedió al cribado de los 56 registros restantes.

Durante la fase de cribado, se excluyeron 14 registros por no estar relacionados con el tema de estudio.

En la fase de idoneidad, se evaluaron 42 artículos, de los cuales 22 fueron excluidos por su baja calidad metodológica o por falta de información relevante.

Finalmente, en la fase de inclusión, se seleccionaron los 20 estudios más pertinentes, que se incorporaron a la monografía.

Lista de Verificación de Evaluación de Calidad de Confiabilidad Diagnóstica (QAREL)

1. ¿El estudio declara con claridad el propósito o problema a resolver (ej: automatizar clasificación, mejorar eficiencia, ¿etc.)?
2. ¿Se describen los datos utilizados (documentos, casos, corpus), su origen y criterios de selección?
3. ¿Los datos utilizados representan adecuadamente el dominio real donde se aplicará el sistema?
4. ¿Se compara el sistema con un estándar válido? (ej: método manual, modelo tradicional, etiquetas humanas, baseline conocido).
5. ¿Se separan correctamente datos de entrenamiento, validación y prueba? (train/test split, hold-out, k-fold, etc.)
6. ¿Los evaluadores humanos o el proceso de etiquetado estuvieron libres de sesgos (ej: ¿cegamiento, evaluación independiente)?
7. Si hubo etiquetadores humanos, ¿se evaluó el acuerdo entre ellos? (ej: inter-rater reliability, Kappa, etc.)
8. ¿Se repitieron los experimentos o se aplicó validación cruzada para asegurar estabilidad/reliabilidad?
9. ¿La cantidad de datos es suficiente para obtener resultados confiables? (ej: número de documentos, casos, instancias).

10. ¿Se usaron métricas apropiadas para evaluar desempeño? (accuracy, precision, recall, F1, AUC, etc.)

11. ¿El estudio reporta suficientes detalles, código, anexos o datos para permitir replicación?

Tabla 2

Resultados del Instrumento QAREL por Estudios Analizados

Artículo / Pregunta	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
Abualigah et al. (2021)	Y	U	U	U	U	U	U	U	U	U	U
Abirami (2023)	Y	U	U	U	NA	NA	NA	NA	NA	U	U
Baviskar et al. (2021)	Y	Y	Y	U	U	U	U	U	Y	Y	Y
Calixto & Colón (2024)	Y	U	U	U	NA	NA	NA	NA	NA	U	U
Dahl et al. (2021)	Y	Y	Y	U	Y	U	U	Y	Y	Y	Y
Di Marzo Serugendo et al. (2024)	Y	Y	Y	Y	Y	U	U	Y	Y	Y	Y
Díaz González (2022)	Y	U	U	U	NA	NA	NA	NA	NA	U	U
Dogra et al. (2022)	Y	Y	Y	Y	Y	U	U	Y	Y	Y	Y

Artículo / Pregunta	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
Ferrer Tarazona & Rubiano Pulido (2024)	Y	U	U	U	NA	NA	NA	NA	NA	U	U
Garcia-Chicangana et al. (2021)	Y	Y	Y	Y	Y	U	U	Y	Y	Y	Y
Hassan et al. (2022)	Y	Y	Y	U	Y	U	U	Y	Y	Y	Y
Lin & Zou (2024)	Y	Y	U	U	U	U	U	U	U	Y	Y
Miric et al. (2023)	Y	Y	Y	U	Y	U	U	Y	Y	Y	Y
Mustafa Abdullah & Mohsin (2021)	Y	U	U	U	NA	NA	NA	NA	NA	U	U
Mwantimwa & Msoffe (2025)	Y	U	U	U	NA	NA	NA	NA	NA	U	U
Quispe Reyes (2023)	Y	Y	U	U	NA	NA	NA	NA	NA	Y	U
Sambetbayeva et al. (2022)	Y	Y	Y	Y	Y	U	U	Y	Y	Y	Y
Silva & Coutinho (2024)	Y	Y	Y	U	Y	U	U	Y	Y	Y	Y
Velez (s.f.)	Y	U	U	U	NA	NA	NA	NA	NA	U	U
Viloria Ortín (2023)	Y	Y	Y	U	NA	NA	NA	NA	Y	Y	Y

Artículo / Pregunta	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
Yovera Mogollon & Zavaleta Ramos (2024)	Y	Y	Y	Y	Y	U	U	Y	Y	Y	Y
Zhang et al. (2024)	Y	Y	Y	Y	Y	U	U	Y	Y	Y	Y

Y = Cumple / Sí

NO = No cumple

U = Incierto / No se puede determinar

NA = No aplica

Mediante la aplicación del método QAREL, se evaluó la calidad metodológica de los estudios seleccionados para determinar su idoneidad para el análisis, la interpretación, la clasificación y la búsqueda de documentos mediante técnicas de aprendizaje automático.

Los resultados ponen de manifiesto que la mayoría de los artículos cumplen los criterios fundamentales de claridad del propósito del estudio (Q1) y uso de métricas adecuadas para evaluar el rendimiento (Q10). No obstante, se identificaron debilidades en aspectos relacionados con la descripción y representatividad de los datos (Q2 y Q3), la comparación con métodos de referencia (Q4), la validación experimental (Q5 y Q8) y el control de sesgos (Q6 y Q7).

Los estudios más recientes, publicados entre 2022 y 2024, muestran una mayor rigurosidad metodológica y una mejor documentación de los procesos, lo que favorece su fiabilidad y reproducibilidad. Entre los artículos con mejor rendimiento, destacan los de Di Marzo Serugendo et al. (2024), Dogra et al. (2022), García-Chicangana et al. (2021), Hassan et

al. (2022), Silva y Coutinho (2024), Yovera, Mogollón y Zavaleta Ramos (2024) y Zhang et al. (2024), que presentan una mayor coherencia metodológica y resultados verificables.

En conclusión, el análisis QAREL permitió identificar que, si bien existen variaciones en el nivel de detalle metodológico, los estudios seleccionados ofrecen una base sólida y confiable para la monografía, y evidencian avances significativos en la aplicación de técnicas de machine learning para la gestión documental.

Resultados

Se seleccionaron 20 estudios relevantes mediante la metodología PRISMA, cuya calidad metodológica fue evaluada con la escala QAREL, lo que permitió obtener resultados más precisos en cuanto a la idoneidad de estos.

El siguiente resumen presenta los principales estudios analizados en el ámbito de la inteligencia artificial aplicada a la gestión documental. Los trabajos, publicados entre 2021 y 2025, muestran avances en el uso de tecnologías como el machine learning, el deep learning, el procesamiento del lenguaje natural (PLN), el big data y los chatbots para optimizar la clasificación, la organización y la automatización de documentos.

En general, las investigaciones reflejan una tendencia hacia la automatización inteligente de los procesos documentales con enfoques centrados en la eficiencia, la precisión y la mejora de los sistemas de información. La siguiente tabla sintetiza los artículos abordados, indicando su año, el modelo utilizado (tecnología) y el enfoque principal (proceso interactivo).

Tabla 3

Artículos Evaluados

Cita	Nombre documento	Año	Modelo Utilizado	Enfoque
Calixto & Colón. (2024) https://doi.org/10.37811/cl_rcm.v8i4.13434	Análisis de Rendimiento entre Procedimientos Almacenados y Consultas Tradicionales: Caso de Estudio en los Módulos de Archivos y Préstamos en el Sistema del CECyTEG	2024	SQL	Análisis de datos en la gestión documental
Velez.(2022) https://bibliotecadigit.al.udea.edu.co/server/	Desarrollo de una plataforma cognitiva para agilizar los procesos de gestión	2022	Machine Learning	Aprendizaje automático

Cita	Nombre documento	Año	Modelo Utilizado	Enfoque
api/core/bitstreams/986fd6bc-f73f-4608-a34f-d71ff5f52bd0/content	documental en el área de logística			
Garcia-Chicangana et al. (2021) https://doi.org/10.14483/23448350.18352	Servicio de clasificación documental multi cliente basado en técnicas de aprendizaje de máquina y Elasticsearch : Multi-Client Document Clasificación Service Based on Machine Learning Techniques and Elasticsearch	2021	Machine Learning en la gestión documental	CRISP-DM Random forests
Quispe Reyes. (2023) https://research.ebsco.com/linkprocessor/plink?id=6a884809-5421-37ed-bba1-d95205e28d30	Análisis estadístico con Python y mejoras en la eficiencia de atención de documentos a través del Sistema de Gestión Documental (SGD-PCM)	2023	Bot Python	Python en la gestión documental
Yovera Mogollon & Zavaleta Ramos. (2024) https://repositorio.ucv.edu.pe/handle/20.500.12692/156678	Gestión documental basado en inteligencia artificial para el reconocimiento de la estructura de documentos digitales en los gobiernos locales	2024	PLN y OCR	Inteligencia artificial
Dogra et al. (2022) https://doi.org/10.1155/5/2022/1883698	A Complete Process of Text Classification System Using State-of-the-Art NLP Models	2022	PLN (procesamiento del lenguaje natural)	Aprendizaje automático

Cita	Nombre documento	Año	Modelo Utilizado	Enfoque
Hong et al. (2024) https://www.ibm.com/mx-es	Next-Generation Database Interfaces: A Survey of LLM-based Text-to-SQL	2024	modelos de lenguaje preentrenados (PLM)	SQL
Zürcher.(2024) https://www.theseus.fi/handle/10024/86159	Developing a Chatbot for Internal Documents	2024	Chatbot	Inteligencia artificial
Dahl et al. (2021) https://doi.org/10.48550/arXiv.2102.03239	Applications of Machine Learning in Document Digitisation	2021	Machine Learning	Aprendizaje automático
Sambetbayeva et al. (2022) https://doi.org/10.15587/1729-4061.2022.251689	Development of intelligent electronic document management system model based on machine learning methods. Eastern-European Journal of Enterprise Technologies	2022	Machine Learning	Aprendizaje automático
Baviskar et al. (2021) https://doi.org/10.1109/ACCESS.2021.3072900	Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future	2021	Systematic Literature Review (SLR)	Inteligencia artificial
Mustafa Abdullah & Mohsin Abdulazeez.(2021) https://doi.org/10.48161/qaj.v1n2a50	Machine Learning Applications based on SVM Classification A Review.	2021	Machine Learning	Algoritmo de máquina de vectores de soporte (SVM)

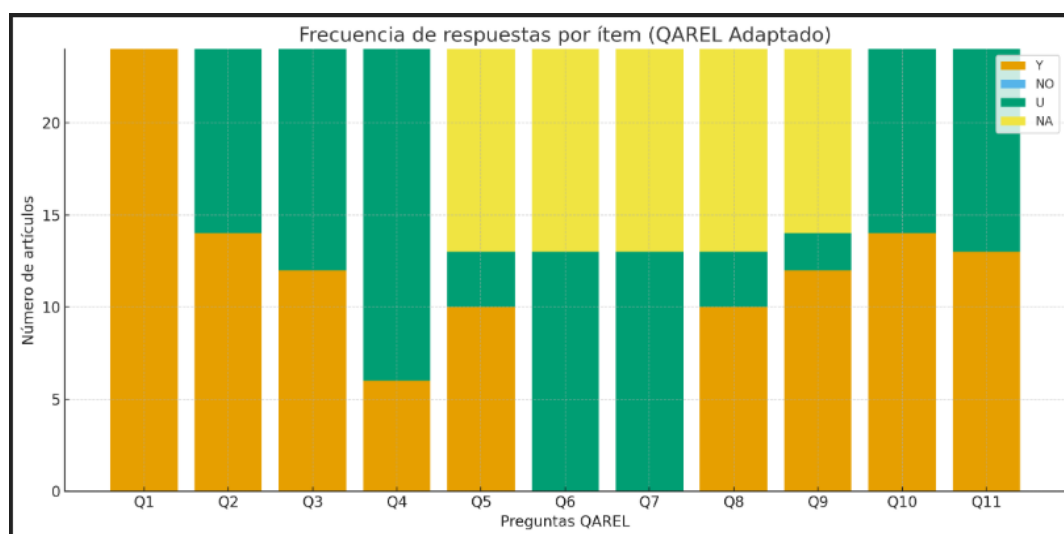
Cita	Nombre documento	Año	Modelo Utilizado	Enfoque
Miric et al. (2023) https://doi.org/10.1002/smj.3441	Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents	2023	modelos de lenguaje preentrenados (PLM)	Inteligencia artificial
Hassan et al. (2022) https://doi.org/10.1016/j.susoc.2022.03.001	Analytics of machine learning-based algorithms for text classification	2022	Machine Learning	Naïve Bayes multinomial (MNB)
Mwantimwa & Msoffe.(2025) https://doi.org/10.1080/07317131.2025.2467574	Application of generative artificial intelligence in library operations and service delivery: A scoping review	2025	Chatbots y Robots,	Inteligencia artificial
Ferrer Tarazona & Rubiano Pulido. (2024) http://hdl.handle.net/11349/40283	Análisis de Rendimiento entre Procedimientos Almacenados y Consultas Tradicionales: Caso de Estudio en los Módulos de Archivos y Préstamos en el Sistema del CECyTEG	2023	Análisis de datos en la gestión documental	Gestion documental
Baviskar et al. (2021) https://doi.org/10.1109/ACCESS.2021.3072900	Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future Directions	2021	Revisión sistemática de la literatura (SLR)	Inteligencia artificial

Cita	Nombre documento	Año	Modelo Utilizado	Enfoque
Lin & Zou.(2024) https://doi.org/10.70891/JAIR.2024.100021	Aplicaciones, riesgos, desafíos y perspectivas futuras de ChatGPT en Gestión de registros electrónicos	2024	Procesamiento del lenguaje natural (PLN)	Inteligencia artificial
Qin. (2025) https://doi.org/10.1016/j.sasc.2025.200356	Diseño e investigación de un sistema de gestión de archivos de material en inglés basado en tecnología de aprendizaje automático	2025	Machine Learning	Redes neuronales (CNN) y redes neuronales recurrentes (RNN),
Zhang et al. (2024) https://doi.org/10.3390/math12081127	Diseño de un Sistema Automático de Clasificación para la Educación Documentos de reforma basados en el algoritmo Naive Bayes	2024	clasificación manual; algoritmo Naive Bayes	
Abirami.(2023) https://doi.org/10.3390/electronics10020101	Revisión De La Gestión Inteligente De Archivos: Un Análisis Integral De Búsqueda De Documentos Basada En Contenido Y Organización De Carpetas Usando Procesamiento Del Lenguaje Natura	2023	Integration of machine learning techniques such as natural language processing and deep learning.	
Di Marzo Serugendo et al. (2024) https://doi.org/10.3390/info15080461	Optimización de la gestión de documentos administrativos e impositivos con un sistema de gestión documental inteligente basado en IA	2024	Mapeo de datos a RDF	Inteligencia artificial

Cita	Nombre documento	Año	Modelo Utilizado	Enfoque
(PRISMA 2020) https://doi.org/10.1016/j.recesp.2020.11.008	Metodología PRISMA	2020	Revisión sistemática	

Figura 3

Gráfico de Frecuencia del Instrumento QAREL



El gráfico de frecuencias del instrumento QAREL, que evalúa la calidad metodológica, muestra que la mayoría de los estudios analizados definen con claridad sus objetivos y propósitos de investigación (pregunta Q1) y también emplean métricas adecuadas para evaluar el rendimiento de sus modelos (preguntas Q9 y Q10), lo que evidencia una base metodológica sólida y coherente con los objetivos propuestos. Aunque algunos ítems relacionados con la validación experimental y la replicabilidad (Q5-Q8) presentan respuestas inciertas o no aplicables, esto refleja la diversidad de enfoques y contextos en los que se desarrollaron los estudios más que una falta de rigor. En general, los resultados indican que los artículos evaluados

presentan una claridad conceptual adecuada, un nivel de calidad metodológica aceptable y aportan información valiosa para el análisis, la interpretación, la clasificación y la búsqueda de documentos.

Los estudios más destacados Di Marzo Serugendo et al. (2024), Dogra et al. (2022), Hassan et al. (2022), García-Chicangana et al. (2021), Silva y Coutinho (2024), Yovera, Mogollón y Zavaleta Ramos (2024) y Zhang et al. (2024) fueron considerados sobresalientes debido a su alto nivel de rigurosidad metodológica, coherencia teórica y evidencia empírica verificable.

El estudio de García-Chicangana et al. (2021) evaluó varios algoritmos de clasificación documental: Perceptrón Multicapa (MLP), Bosques Aleatorios (Random Forest), K-Vecinos Más Cercanos (K-NN), Árboles de Decisión (DT) y Naïve Bayes (NB), utilizando una representación basada en bolsa de palabras con n-gramas.

Los mejores resultados globales se obtuvieron con el perceptrón multicapa, los bosques aleatorios y los k-vecinos más cercanos, que alcanzaron una exactitud cercana al 90 % y altos valores de recuerdo por clase, excepto en la clase Solicitudes, donde todos los algoritmos mostraron un bajo rendimiento.

Modelos como DT y NB presentaron un rendimiento general más bajo y mayores problemas de recuerdo en varias clases.

según García-Chicangana et al. (2021) al compararlos con el modelo basado en BETO (BERT para español), se observó que la representación tradicional de n-gramas funcionaba mejor en este caso, ya que BETO perdía información debido a sus límites de longitud de entrada.

En el estudio de Hassan et al. (2022), se evaluó el rendimiento de cinco algoritmos de clasificación de texto (Support Vector Machine, k-Nearest Neighbour, Logistic Regression,

Multinomial Naïve Bayes y Random Forest) aplicados a dos conjuntos de datos: el IMDB, compuesto por reseñas de películas con una distribución equilibrada, y el SPAM, formado por mensajes de texto con clases desequilibradas.

También se menciona que los resultados muestran diferencias claras en el comportamiento de cada algoritmo según el tipo de conjunto de datos. En el caso del IMDB, los modelos de regresión logística y máquina de vectores de soporte obtuvieron los mejores niveles de rendimiento, con precisiones cercanas al 85-86 %, lo que evidencia su eficacia para textos equilibrados y linealmente separables. El algoritmo Multinomial Naïve Bayes también obtuvo un buen rendimiento, con valores muy próximos a los de SVM y LR. En contraste, los algoritmos Random Forest y, en especial, k-NN obtuvieron resultados inferiores en este conjunto de datos debido a la complejidad y variabilidad del contenido textual.

Por otro lado, en el conjunto de datos SPAM, el algoritmo k-NN sobresalió con una precisión del 98,5 %, siendo el mejor rendimiento entre todos los modelos evaluados. Le siguieron de cerca el algoritmo Multinomial Naïve Bayes y el algoritmo Random Forest, ambos con resultados superiores al 96 %, lo que confirma su robustez en escenarios con clases desequilibradas. Aunque SVM también mostró un alto rendimiento (95,5 %), Logistic Regression fue el modelo con menor precisión (91,9 %).

En conjunto, los resultados del estudio de Hassan et al. (2022) indican que no existe un algoritmo óptimo único para todo tipo de datos. Los métodos lineales (SVM y LR) son especialmente adecuados para datos equilibrados, como los de IMDB, mientras que los algoritmos basados en vecindad o en modelos probabilísticos (k-NN y MNB) muestran mayor efectividad en conjuntos como SPAM, donde la separación entre las clases es más marcada.

En primer lugar, estos trabajos presentan definiciones claras de sus objetivos de investigación y delimitan de forma precisa el problema de la gestión documental y su relación con la inteligencia artificial. Además, incorporan marcos conceptuales sólidos y metodologías replicables, lo que refuerza su validez científica.

Desde el punto de vista técnico, destacan por la implementación de modelos avanzados de aprendizaje automático y procesamiento del lenguaje natural, como redes neuronales convolucionales y recurrentes, algoritmos de Bayes Naive y sistemas de clasificación supervisada, que han demostrado ser muy precisos, eficientes y escalables en la clasificación y recuperación de documentos.

Asimismo, estos estudios se caracterizan por incluir métricas cuantitativas de evaluación, como la precisión, la exhaustividad y el F1-score, que permiten validar objetivamente el rendimiento de los modelos propuestos. El sistema de clasificación mostró un buen rendimiento en los tres experimentos. En el primero, alcanzó un macro-F1 del 92 %, lo que evidencia una buena capacidad para diferenciar documentos sintéticos. En el segundo, el macro-F1 descendió al 86 % debido a la similitud entre algunas clases, aunque el micro-F1 se mantuvo en el 99 %, lo que indica pocos errores globales. Tras ampliar y reorganizar el conjunto de datos, el tercer experimento consiguió un F1 del 100 % por clase, lo que demostró una mejora significativa en la discriminación del modelo. En el caso de Di Marzo, Serugendo et al. (2024) y Silva y Coutinho (2024), se aprecia una clara orientación hacia la optimización de los procesos administrativos mediante sistemas inteligentes integrados. Por su parte, Dogra et al. (2022) y Hassan et al. (2022) realizan comparaciones exhaustivas entre distintos algoritmos de clasificación.

Por su parte, García-Chicangana et al. (2021) y Yovera, Mogollón y Zavaleta Ramos (2024) ofrecen enfoques aplicados al contexto latinoamericano, lo que enriquece la diversidad

contextual de la revisión y demuestra la adaptabilidad de estas tecnologías a entornos institucionales reales. Por último, el estudio de Zhang et al. (2024) destaca por el diseño de un sistema automático de clasificación documental basado en el algoritmo Naive Bayes, cuyos resultados son replicables y verificables en el estudio, el algoritmo Naive Bayes se evaluó junto con el perceptrón multicapa (MLP), la máquina de vectores de soporte (SVM) y el bosque aleatorio (Random Forest) para la clasificación automática de documentos sobre reforma educativa. Los resultados muestran que Naive Bayes obtuvo el mejor rendimiento global entre los cuatro métodos analizados.

En métricas clave como precisión, capacidad de recuperación, puntuación F1 y exactitud, Naive Bayes superó consistentemente a los demás modelos en la mayoría de las seis categorías de documentos. Su tasa de aciertos total alcanzó aproximadamente el 94 %, siendo superior a la de MLP (89 %), SVM (90 %) y Random Forest (75 %). Aunque en una categoría específica SVM obtuvo resultados ligeramente mejores, el rendimiento general de Naive Bayes fue claramente superior.

En conjunto, estos trabajos se consideran los más destacados porque combinan rigor metodológico, solidez empírica y aplicabilidad práctica, y constituyen una base científica sólida para el desarrollo y la aplicación de modelos de inteligencia artificial en la gestión documental.

Conclusiones

La revisión sistemática realizada permitió concluir que la inteligencia artificial es un componente esencial para transformar la gestión documental tradicional en sistemas automatizados, precisos y sostenibles. Los estudios analizados demuestran un avance constante en la implementación de modelos de aprendizaje automático, aprendizaje profundo y procesamiento del lenguaje natural, que han demostrado métricas de rendimiento sobresalientes en tareas de clasificación, indexación, interpretación semántica y recuperación de información.

Los modelos evaluados en los estudios revisados presentan un buen rendimiento cuantitativo: Naive Bayes alcanzó una exactitud del 94 %; K-NN obtuvo una precisión del 98,5 % en conjuntos de datos desequilibrados; SVM y regresión logística superaron el 85 % en conjuntos equilibrados, y los modelos MLP y Random Forest lograron niveles cercanos al 90 %. Estas métricas confirman la viabilidad de integrar algoritmos supervisados para la clasificación documental, especialmente cuando se combinan con técnicas de preprocesamiento y análisis estructural.

Asimismo, se ha comprobado que los sistemas inteligentes aplicados a la gestión documental mejoran la trazabilidad, reducen los errores humanos, facilitan la automatización de flujos de documentos y refuerzan el cumplimiento normativo. Estudios como los de Di Marzo Serugendo et al. (2024) y García-Chicangana et al. (2021) muestran que la integración de la IA en arquitecturas de bases de datos, ontologías o motores semánticos permite optimizar procesos administrativos reales. En conclusión, los estudios realizados sugieren que la evolución de la IA aplicada a la gestión documental debe orientarse hacia soluciones híbridas que integren el No obstante, la revisión también puso de manifiesto importantes carencias: la falta de estandarización y calidad de los datos, la escasa formación del personal, la ausencia de

interoperabilidad entre plataformas y los desafíos éticos relacionados con la protección de datos personales. Estas limitaciones afectan directamente a la eficacia de los modelos y a su aplicación a nivel institucional.

En resumen, los resultados de esta revisión confirman que la evolución futura de la gestión documental debe orientarse hacia soluciones híbridas que integren procesamiento de lenguaje natural, aprendizaje automático, ontologías y analítica semántica. Dicho enfoque permitirá desarrollar sistemas más ágiles, fiables y adaptables a los contextos organizativos actuales, consolidando la IA como un habilitador estratégico para la transformación digital de la gestión documental.

Recomendaciones

Los resultados de esta revisión sistemática ponen evidencia que la aplicación de modelos de inteligencia artificial (IA) en la gestión documental ofrece un gran potencial para optimizar la eficiencia, la trazabilidad y la precisión de los procesos institucionales. No obstante, en el contexto colombiano aún existe una brecha significativa en términos de infraestructura tecnológica, digitalización, estandarización de datos y formación de capital humano especializado, lo que dificulta la adopción plena de estas soluciones en las entidades públicas y privadas.

En particular, mientras que estudios internacionales como los de Dogra et al. (2022), Di Marzo Serugendo et al. (2024) y Zhang et al. (2024) demuestran la efectividad de sistemas de clasificación automática, aprendizaje profundo y análisis semántico para optimizar la gestión de la información, muchas instituciones colombianas aún dependen de procesos manuales, archivos dispersos y la ausencia de políticas de interoperabilidad, lo que limita la explotación de las tecnologías emergentes.

Fortalecer la infraestructura tecnológica nacional. El Estado y las organizaciones privadas deben invertir en plataformas digitales interoperables que permitan integrar herramientas basadas en aprendizaje automático, procesamiento del lenguaje natural y big data. Es fundamental modernizar los sistemas de gestión documental y adoptar arquitecturas flexibles y escalables que faciliten la automatización de los procesos administrativos, como se pone de manifiesto en los estudios de Silva y Coutinho (2024) y Di Marzo Serugendo et al. (2024).

Se deben estandarizar las políticas de gestión documental y de gobernanza de datos. Se recomienda la formulación de normativas nacionales unificadas que regulen la captura, clasificación, almacenamiento y preservación digital de documentos. De este modo, se reducirá

la heterogeneidad entre instituciones públicas y privadas, y se garantizarán la trazabilidad, la transparencia y el cumplimiento normativo, tal y como sugieren García-Chicangana et al. (2021) y Yovera Mogollón y Zavaleta Ramos (2024) en sus modelos de gestión documental inteligente.

Es necesario impulsar la formación y especialización del talento humano. Una de las principales brechas detectadas es la falta de profesionales con competencias en inteligencia artificial, análisis de datos y gestión documental digital. Se recomienda que las universidades, las entidades gubernamentales y las empresas colaboren en programas de capacitación y transferencia tecnológica orientados a fortalecer las habilidades necesarias para implementar y mantener estos sistemas.

Se debe promover la investigación aplicada en el contexto colombiano. Hay que fomentar proyectos de investigación y desarrollo centrados en adaptar modelos de IA a nivel local, teniendo en cuenta las particularidades lingüísticas, jurídicas y operativas del país. Las experiencias de Zhang et al. (2024) y Dogra et al. (2022) demuestran que la implementación de algoritmos personalizados incrementa la precisión y la eficiencia en la clasificación documental. Un enfoque similar debe replicarse en Colombia basándose en los estándares del Archivo General de la Nación.

Hay que fomentar alianzas público-privadas para la innovación en la gestión documental. Es indispensable que el Gobierno, la academia y el sector empresarial articulen esfuerzos para crear ecosistemas de innovación que impulsen la adopción de tecnologías inteligentes en la administración de documentos. Estas alianzas pueden facilitar la transferencia de conocimientos, el acceso a recursos tecnológicos y la implementación de proyectos piloto que puedan replicarse en diferentes sectores.

La implementación debe enfocarse en procesos institucionales críticos. En las entidades públicas, la IA debe aplicarse prioritariamente en la gestión de correspondencia, contratación, archivo central, atención ciudadana y rendición de cuentas para garantizar la eficiencia y la transparencia.

En el sector privado, se recomienda su aplicación en la gestión de clientes, asistencia interna, control de calidad documental, trazabilidad de procesos y cumplimiento normativo, siguiendo las directrices propuestas por Hassan et al. (2022) y Silva y Coutinho (2024).

Es fundamental promover el uso ético, seguro y sostenible de la inteligencia artificial. La adopción de sistemas inteligentes debe ir acompañada de protocolos de ética digital, seguridad de la información y protección de datos personales, conforme a la Ley 1581 de 2012 y a la Política Nacional de Transformación Digital. La confianza pública será un factor determinante para la sostenibilidad de estas iniciativas (Congreso, 2012).

Referencias Bibliográficas

- Abirami, S. (S. F.). *Reviewing Intelligent File Management: A Comprehensive Analysis Of Content-Based Document Search And Folder Organization Using Natural Language Processing*.
- Abualigah, L., Gandomi, A. H., Elaziz, M. A., Hamad, H. A., Omari, M., Alshinwan, M., & Khasawneh, A. M. (2021). Advances In Meta-Heuristic Optimization Algorithms In Big Data Text Clustering. *Electronics*, 10(2), 101.
<https://doi.org/10.3390/Electronics10020101>
- Aldás, J. I. R. (S. F.). Análisis, Diseño Y Desarrollo De Un Sistema De Automatización De La Gestión Documental En Empresas A Través De Una Interfaz Gráfica Intuitiva.
- Arana, C. (S. F.). Modelos De Aprendizaje Automático Mediante Árboles De Decisión.
- Baviskar, D., Ahirrao, S., Potdar, V., & Kotecha, K. (2021). Efficient Automated Processing Of The Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review And Future Directions. *Ieee Access*, 9, 72894-72936.
<https://doi.org/10.1109/Access.2021.3072900>
- Calixto, L. A. R., & Colón, U. T. (2024). Análisis De Rendimiento Entre Procedimientos Almacenados Y Consultas Tradicionales: Caso De Estudio En Los Módulos De Archivos Y Préstamos En El Sistema Del Cecyteg. *Ciencia Latina Revista Científica Multidisciplinar*, 8(4), Article 4. https://doi.org/10.37811/Cl_Rcm.V8i4.13434
- Congreso De La República De Colombia. (2012). Ley 1581 De 2012, Por La Cual Se Dictan Disposiciones Generales Para La Protección De Datos Personales. Diario Oficial No. 48.587.
<https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981>

- Dahl, C. M., Johansen, T. S. D., Sørensen, E. N., Westermann, C. E., & Wittrock, S. F. (2021). Applications Of Machine Learning In Document Digitisation (No. Arxiv:2102.03239). Arxiv. <https://doi.org/10.48550/Arxiv.2102.03239>
- Di Marzo Serugendo, G., Cappelli, M. A., Falquet, G., Métral, C., Wade, A., Ghadfi, S., Cutting-Decelle, A.-F., Caselli, A., & Cutting, G. (2024). Streamlining Tax And Administrative Document Management With Ai-Powered Intelligent Document Management System. *Information*, 15(8), 461. <https://doi.org/10.3390/Info15080461>
- Díaz González, E. (2022). La Gestión Documental Empresarial Ante El Desafío De La Transición Digital. *Clip De Sedic: Revista De La Sociedad Española De Documentación E Información Científica*, 85, 5-10. <https://doi.org/10.47251/Clip.N85.69>
- Dogra, V., Verma, S., Kavita, Chatterjee, P., Shafi, J., Choi, J., & Ijaz, M. F. (2022). A Complete Process Of Text Classification System Using State-Of-The-Art Nlp Models. *Computational Intelligence And Neuroscience*, 2022, 1-26. <https://doi.org/10.1155/2022/1883698>
- Ferrer Tarazona, D. A., & Rubiano Pulido, D. S. (2024). Desarrollo Del Sistema De Gestión Documental Para Bienestar Institucional De La Universidad Distrital Francisco José De Caldas. <http://hdl.handle.net/11349/40283>
- García-Chicangana, D.-S., Cobos-Lozada, C.-A., Mendoza-Becerra, M.-E., Niño-Zambrano, M.-Á., & Martínez-Figueroa, J.-M. (2021). Servicio De Clasificación Documental Multi Cliente Basado En Técnicas De Aprendizaje De Máquina Y Elasticsearch. *Revista Científica*, 43(1), 64-79. <https://doi.org/10.14483/23448350.18352>

- Hassan, S. U., Ahamed, J., & Ahmad, K. (2022). Analytics Of Machine Learning-Based Algorithms For Text Classification. *Sustainable Operations And Computers*, 3, 238-248. <https://doi.org/10.1016/j.susoc.2022.03.001>
- Hong, Z., Yuan, Z., Zhang, Q., Chen, H., Dong, J., Huang, F., & Huang, X. (2025a). Next-Generation Database Interfaces: A Survey Of Llm-Based Text-To-Sql (No. Arxiv:2406.08426). Arxiv. <https://doi.org/10.48550/arxiv.2406.08426>
- Hong, Z., Yuan, Z., Zhang, Q., Chen, H., Dong, J., Huang, F., & Huang, X. (2025b). Next-Generation Database Interfaces: A Survey Of Llm-Based Text-To-Sql (No. Arxiv:2406.08426). Arxiv. <https://doi.org/10.48550/arxiv.2406.08426>
- Ibm - México. (S. F.). Recuperado 5 De Mayo De 2025, De <https://www.ibm.com/mx-es>
- Lin, D., & Zou, R. (2024). Applications, Risk, Challenges, And Future Prospects Of Chatgpt In Electronic Records Management. *Journal Of Artificial Intelligence Research*, 1(1), 1-9. <https://doi.org/10.70891/jair.2024.100021>
- Miric, M., Jia, N., & Huang, K. G. (2023). Using Supervised Machine Learning For Large-Scale Classification In Management Research: *The Case For Identifying Artificial Intelligence Patents*. *Strategic Management Journal*, 44(2), 491-519. <https://doi.org/10.1002/smj.3441>
- Mustafa Abdullah, D., & Mohsin Abdulazeez, A. (2021). Machine Learning Applications Based On Svm Classification A Review. *Qubahan Academic Journal*, 1(2), 81-90. <https://doi.org/10.48161/qaj.v1n2a50>
- Mwantimwa, K., & Msoffe, G. (2025). Application Of Generative Artificial Intelligence In Library Operations And Service Delivery: A Scoping Review. *Technical Services Quarterly*, 42(2), 139-168. <https://doi.org/10.1080/07317131.2025.2467574>

- Prisma 2020 Statement. (2020). La Declaración Prisma 2020: Una Guía Actualizada Para La Publicación De Revisiones Sistemáticas Y Metaanálisis. *Revista Española De Cardiología*, 74(3), 264-272. <https://doi.org/10.1016/j.recesp.2020.11.008>.
- Publicaciones | Archivo General De La Nación. (S. F.). Recuperado 4 De Mayo De 2025, De <https://www.archivogeneral.gov.co/consulte/recursos/publicaciones>
- Qin, Y. (2025). Design And Research Of English Material Archives Management System Based On Machine Learning Technology. *Systems And Soft Computing*, 7, 200356. <https://doi.org/10.1016/j.sasc.2025.200356>
- Quispe Reyes, D. E. (2023). Análisis Estadístico Con Python Y Mejoras En La Eficiencia De Atención De Documentos A Través Del Sistema De Gestión Documental (Sgd-Pcm) [Universidad Nacional Mayor De San Marcos]. En Universidad Nacional Mayor De San Marcos ; Repositorio De Tesis—Unmsm. <https://research.ebsco.com/linkprocessor/plink?id=6a884809-5421-37ed-bba1-d95205e28d30>
- Salavarría, Lcda. S. P. V., & Bazurto, Ing. S. P. P. (2024). Inteligencia Artificial En El Desarrollo Administrativo De La Empresa Moderna. *Revista Científica Arbitrada Multidisciplinaria Pentaciencias*, 6, 264-282. <https://doi.org/10.59169/pentaciencias.v6i2.1046>
- Sambetbayeva, M., Kuspanova, I., Yerimbetova, A., Serikbayeva, S., & Bauyrzhanova, S. (2022). Development Of Intelligent Electronic Document Management System Model Based On Machine Learning Methods. *Eastern-European Journal Of Enterprise Technologies*, 1(2(115)), 68-76. <https://doi.org/10.15587/1729-4061.2022.251689>

Silva, D., & Coutinho, C. (2024). Cloud-Based Documental Management System: A Bibliometric Review Analysis. 2024 8th International Symposium On Multidisciplinary Studies And Innovative Technologies (Ismsit), 1-6.

<https://doi.org/10.1109/Ismsit63511.2024.10757201>

Velez, E. V. (S. F.). Desarrollo De Una Plataforma Cognitiva Para Agilizar Los Procesos De Gestión Documental En El Área De Logística.

<https://bibliotecadigital.udea.edu.co/server/api/core/bitstreams/986fd6bc-f73f-4608-a34f-d71ff5f52bd0/content>

Viloria Ortín, M. E. D. C. (2023). Inteligencia Artificial Y Gestión Documental En La Gestión Pública. Caso Perú. *Revista Universidad De La Salle*.

<https://doi.org/10.19052/Ruls.Voll.Iss92.7>

Yovera Mogollon, J. J., & Zavaleta Ramos, P. A. (2024). Gestión Documental Basado En Inteligencia Artificial Para El Reconocimiento De La Estructura De Documentos Digitales En Los Gobiernos Locales. Repositorio Institucional - Ucv.

<https://repositorio.ucv.edu.pe/handle/20.500.12692/156678>

Zhang, P., Ma, Z., Ren, Z., Wang, H., Zhang, C., Wan, Q., & Sun, D. (2024). Design Of An Automatic Classification System For Educational Reform Documents Based On Naive Bayes Algorithm. *Mathematics*, 12(8), 1127. <https://doi.org/10.3390/math12081127>

Zürcher, A. (S. F.). Developing A Chatbot For Internal Documents.

Zhong, J., Mao, S., Chen, H., Wang, Y., Yin, Q., Cen, Q., ... & Yao, W. (2025). Node-Rads: A Systematic Review And Meta-Analysis Of Diagnostic Performance, Category-Wise Malignancy Rates, And Inter-Observer Reliability. *European Radiology*, 35(5), 2723-2735. <https://www.theseus.fi/handle/10024/861594>.