



ISSN: 1646-9895

Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

J u n h o 2 6 • J u n e 2 6



©ITMA 2026 <http://www.risti.xyz>

Nº 62

Edição / Edición

Nº 62, 06/2026

Tiragem / Tirage: 1000

Preço por número / Precio por número: 17,5€

Subscrição anual / Suscripción anual: 30€ (2 números)

ISSN: 1646-9895

Depósito legal:

Indexação / Indexación

Academic Journals Database, CiteFactor, Dialnet, DOAJ, DOI, EBSCO, GALE, IndexCopernicus, Index of Information Systems Journals, Web of Science, Latindex, ProQuest, QUALIS, SciELO, SIS, Ulrich's.

Publicação / Publicación

ITMA – Information and Technology Management Association

Pc. 9 de Abril, 26, 4200-422 Porto, Portugal

E-mail: aistic@gmail.com

Web: <http://www.risti.xyz>

Director

Álvaro Rocha, Universidade Lusófona, PT

Coordenadores da Edição / Coordinadores de la Edición

Álvaro Rocha, Universidade Lusófona, PT

Conselho Editorial / Consejo Editorial

Abel Méndez, Technological Institute of Costa Rica, CR

Alejandro Peña, EAFIT University, CO

Alma Gomez-Rodríguez, University of Vigo, ES

Ana Rita Calvao, ESTGA, University of Aveiro, PT

Ania Cravero, Universidad de La Frontera, CL

António Abreu Silva, ISCAP, Polytechnic Institute of Porto, PT

Antonio Garcia, University of Santiago de Compostela, ES

António Godinho, ISEC, Polytechnic Institute of Coimbra, PT

Antonio Jiménez-Martín, Polytechnic University of Madrid, ES

Arturo J. Méndez, University of Vigo, ES

August Climent, Ramon LLull University, ES

Beatriz Rodríguez, Universidad de la Republica, UY

Borja Bordel, Polytechnic University of Madrid, ES

Brenda L. Flores-Rios, Universidad Autónoma de Baja California, MX

Bruno Miguel Ferreira Gonçalves, Polytechnic Institute of Braganca, PT

Carlos Alexandre Silva, Federal Institute of Minas Gerais, BR

Carlos Carreto, Polytechnic Institute of Guarda, PT

Carlos Morais, Polytechnic Institute of Braganza, PT

Carlos Rompante Cunha, UNIAG & CeDRI &

Polytechnic Institute Bragança, PT

Carlos Vaz de Carvalho, Polytechnic Institute of Porto, PT

Célio Marques, Polytechnic Institute of Tomar, PT

Círo Martins, University of Algarve, PT

Concepción Damián-Hernández, Instituto

Tecnológico Superior de Misantla, MX

Cristina Caridade, Polytechnic Institute of Coimbra, PT

Daniel Poland, University of Aveiro, PT
Dante Carrizo, University of Atacama, CL
Edwin Cedeño Herrera, Universidad de Panamá, PA
Fabio Marques, ESTGA, University of Aveiro, PT
Felipe Vasquez, Universidad de La Frontera, CL
Fernando Moreira, REMIT, IJP, Portucalense University & IEETA,
University of Aveiro, PT
Fernando Paulo Belfo, Coimbra Business School – ISCAC, Polytechnic
Institute of Coimbra, PT
Fernando Ribeiro, Polytechnic Institute of Castelo Branco, PT
Fernando Bandeira, PT
Filipe Cardoso, Polytechnic Institute of Viseu, PT
Flor Gomes de María Sánchez Aguirre, Universidad César Vallejo, PE
Francisco Javier Lena-Acebo, University of Cantabria, ES
Gabriel Guerrero-Contreras, University of Cádiz, ES
Gerardo González Filgueira, University of A Coruña, ES
Gloria Valencia, Universidad de las Fuerzas Armadas, EC
Helder Gomes, University of Aveiro, PT
Helder Zagalo, University of Aveiro, PT
Hélia Guerra, University of the Azores, PT
Henrique Gil, Polytechnic Institute of Castelo Branco, PT
Henrique S. Mamede, Open University, PT
Isaias Bianchi, Federal University of Santa Catarina, BR
Ismael Etxeberria-Agiriano, Universidad del País Vasco, ES
Ivan Garcia, Universidad Tecnológica de la Mixteca, MX
Jeimy Cano, Universidad de los Andes, CO
John Emilio Almeida, ISTECH - Porto, PT
João Paulo Ferreira, Polytechnic Institute of Coimbra, PT
Roberto de Toledo Quadros, CEFET/RJ, BR
Joao Tavares, University of Porto, PT
Joaquim Reis, ISCTE - University Institute of Lisbon, PT
Jorge Eduardo Ibarra Esque, Universidad
Autónoma de Baja California, MX

Jorge Hochstetter, Universidad de La Frontera, CL
José Álvarez-García, University of Extremadura, ES
Jose Felipe Cocon Juarez, Universidad Autónoma del Carmen, MX
José Lousado, Polytechnic Institute of Viseu, PT
José Luis Pastrana Brincones, University of Malaga, ES
Jose M. Molina, Carlos III University of Madrid, ES
José Ribeiro, Polytechnic of Leiria, PT
Jose Silvestre Silva, Military Academy, PT
Juan Angel Contreras Vas, University of Extremadura, ES
Juan M. Santos-Gago, University of Vigo, ES
Juan Pablo DAmato, National University of the Center of the
Province of Buenos Aires, AR
Leila Weitzel, Fluminense Federal University, BR
Leonardo Bermón Angarita, National University of Colombia, CO
Lilia Muñoz, Universidad Tecnológica de Panamá, PA
Lucila Romero, Universidad Nacional del Litoral, AR
Luis Alvarez Sabucedo, University of Vigo, ES
Luis Enrique Sánchez Crespo, University of Castilla-la Mancha, ES
Luisa María Romero-Moreno, Universidad Sevilla, ES
Luz María Hernández Cruz, Universidad Autónoma de Campeche, MX
Luz Sussy Bayona Oré, Universidad Nacional Mayor de San Carlos, PE
Marcelo Marciszack, National Technological University, AR
Marcelo Zambrano Vizuite, Universidad Tecnica Del Norte
Marco Painho, NOVA IMS, PT
Margarita Ramirez Ramirez, Universidad Autónoma de Baja California, MX
Maria Cristina Marcelino Bento, UNIFATEA, BR
María de la Cruz del Río-Rama, University of Vigo, ES
Maria de los Milagros Gutierrez, National Technological University, AR
Maria do Rosario Bernardo, Open University, PT
Maria João Ferreira, Portucalense University, PT
Maria João Gomes, University of Minho, PT
Maria Sousa, ISCTE - University Institute of Lisbon, PT
Marisol B. Correia, ESGHT - University of Algarve & CiTUR, PT

Maristela Holanda, University of Brasilia, BR
Martin Llamas Nistal, University of Vigo, ES
Miguel Casquilho, University of Lisbon, PT
Miguel Ramón González Castro, Aimen Technological Center, ES
Mirna Ariadna Muñoz Mata, CIMAT, MX
Nelson Rocha, University of Aveiro, PT
Nuno Melão, Polytechnic Institute of Viseu, PT
Nuno Ribeiro, Fernando Pessoa University, Portugal
Patricia Dias, State University of Minas Gerais, BR
Paula Prata, University of Beira Interior, PT
Paulo Martins, University of Trás-os-Montes and Alto Douro, PT
Paulo Pinto, New University of Lisbon, PT
Paulo Rurato, Fernando Pessoa University, Portugal
Pedro Araújo, University of Beira Interior, PT
Pedro R. Palos-Sanchez, Universidad de Sevilla, ES
Pedro Sanz-Angulo, University of Valladolid, ES
Pedro Sobral, Fernando Pessoa University, Portugal
Pedro Sousa, University of Minho, Portugal
René Faruk Garzozi-Pincay, Santa Elena Peninsula State University, EC
Ruben Pereira, ISCTE – University Institute of Lisbon, PT
Rui Pedro Marques, University of Aveiro, PT
Rui Silva Moreira, Fernando Pessoa University, PT
Samuel Sepúlveda, Universidad de La Frontera, CL
Sandro Carvalho, Polytechnic of Cávado and Ave, PT
Santiago Raul Gonzales Sanches, Universidad Cesar Vallejo, PE
Sara Balderas-Díaz, University of Cádiz, ES
Sergio Araya Guzmán, Universidad del Bío-Bío, CL
Sergio F. Lopes, University of Minho, PT
Solange N Alves-Souza, University of Sao Paulo, BR
Telmo Silva, DigiMedia, University of Aveiro, PT
Thiago Dias, Federal Center for Technological Education of Minas Gerais, BR
Veronica Vasconcelos, Polytechnic Institute of Coimbra, PT
Vítor Carvalho, Polytechnic Institute of Cávado and Ave, PT

Vitor Santos, New University of Lisbon - NOVA IMS, PT

Wagner Tanaka Botelho, Federal University of ABC, BR

Yamid Hernández Julio, Universidad del Sinú, CO

Índice / Index

EDITORIAL

Inteligência Artificial Responsável e Ciberresiliência nos Sistemas de Informação	1
<i>Álvaro Rocha</i>	

ARTIGOS / ARTICULOS / ARTICLES

Detección de la ceguera en pacientes con Retinopatía Diabética utilizando enfoques de aprendizaje profundo y redes neuronales convolucionales: Una Revisión Sistemática	5
---	---

Julio Cárdenas, José Herrera, Henry Tornero

Predição da Retenção Estudantil no Ensino Superior com Modelos de Aprendizagem de Máquina: Uma Abordagem Baseada em Justiça Algorítmica e Equidade	23
--	----

Tatiana Beltrão Alves da Costa, Luciano de Souza Cabral, Rafael Ferreira Mello

Modelos de Machine Learning para la Predicción de la Deserción en una Institución de Educación Superior en Colombia.....	38
---	----

Ludwing Bernart Carreño Patiño

Diseño e implementación de un prototipo para la estimación del tiempo de espera en la cafetería del edificio #3 de la Universidad Tecnológica de Panamá basado en visión artificial.....	47
---	----

Lucia Cárdenas, Andrea Pipino, Paulo Picota

Riesgos de ciberseguridad y estrategias de mitigación en el Gobierno Digital: una revisión sistemática	63
---	----

Néstor Giankeiler Burga Vásquez

Propuesta de un modelo de madurez para evaluar la ciberresiliencia de las pymes frente al ransomware	86
---	----

Domingo Ruiz Arroyo

Inteligência Artificial Responsável e Ciberresiliência nos Sistemas de Informação

Responsible Artificial Intelligence and Cyber Resilience in Information Systems

Álvaro Rocha¹

alvaro.rocha@ulusofona.pt

¹ FCNET, Universidade Lusófona, Porto, Portugal

DOI: 10.17013/risti.62.1-3

A RISTI – Revista Ibérica de Sistemas e Tecnologias de Informação reafirma, na sua edição número 62, o compromisso com a publicação de investigação científica rigorosa, inovadora e relevante para a comunidade académica e profissional da área dos Sistemas e Tecnologias de Informação. Para esta edição foram recebidas 59 submissões, oriundas de 13 países cujo idioma oficial é o Português ou o Espanhol, tendo sido aceites apenas 6 artigos. A taxa de aceitação de 10,2% traduz o elevado nível de exigência do processo de revisão científica e, simultaneamente, a amplitude geográfica e a vitalidade da comunidade ibero-americana que a revista serve.

Os seis artigos publicados nesta edição partilham um eixo comum: a passagem da demonstração tecnológica para a responsabilidade sobre os efeitos da tecnologia. Seja na saúde, na educação, nos serviços de proximidade ou na proteção de organizações e do Estado, os trabalhos aqui reunidos não se limitam a mostrar que os sistemas inteligentes funcionam — interrogam em que condições devem funcionar, com que garantias de equidade, privacidade, continuidade e confiança.

O primeiro artigo apresenta uma revisão sistemática sobre a deteção da cegueira em pacientes com retinopatia diabética através de abordagens de aprendizagem profunda e redes neuronais convolucionais. Combinando análise bibliométrica e análise qualitativa sobre fontes indexadas na IEEE Digital Library, na ScienceDirect e na Web of Science, o estudo sistematiza técnicas de segmentação e pré-processamento, compara o

desempenho de diferentes algoritmos e caracteriza os conjuntos de dados de imagens do fundo do olho mais utilizados. A contribuição é particularmente pertinente para contextos com escassez de especialistas em retinologia, onde o diagnóstico assistido por computador pode alargar significativamente o acesso ao rastreio precoce.

O segundo artigo aborda a predição da retenção estudantil no ensino superior a partir de uma perspectiva de justiça algorítmica. Articulando técnicas de balanceamento de classes (SMOTE e ADASYN), estratégias de otimização de hiperparâmetros (GridSearch e Optuna) e a avaliação comparada de seis algoritmos supervisionados, o estudo acrescenta às métricas convencionais de desempenho as métricas de equidade Demographic Parity Difference e Equalized Odds Difference. O resultado mais relevante é também o mais incómodo: os modelos com melhor desempenho preditivo não são necessariamente os mais equitativos, o que evidencia um compromisso que os sistemas de alerta precoce não podem continuar a ignorar.

O terceiro artigo prossegue no domínio da análise educacional, propondo modelos preditivos de abandono escolar numa instituição de ensino superior da Colômbia. Seguindo a metodologia CRISP-DM e recorrendo a técnicas de balanceamento e a validação por AutoML e validação cruzada, o trabalho identifica como preditores mais robustos o rendimento académico acumulado, o desempenho nas unidades curriculares iniciais e a participação num programa público de apoio ao acesso ao ensino superior. A leitura conjunta com o artigo anterior é sugestiva: a evidência empírica sobre os fatores de permanência tem tanto valor quanto o cuidado com que é traduzida em intervenção institucional.

No quarto artigo, a visão por computador é aplicada a um problema de serviço quotidiano — a estimativa do tempo de espera numa cafetaria universitária. A arquitetura proposta integra câmaras IP através do protocolo RTSP, o modelo YOLOv8 para deteção e seguimento de pessoas numa zona de interesse delimitada por um polígono, e uma implementação multithread que permite o processamento concorrente e a notificação em tempo real através de uma aplicação móvel institucional. Merece destaque a opção, assumida desde a fase de conceção, por um sistema que não armazena nem transmite imagens identificáveis, extraindo apenas métricas agregadas — uma demonstração prática de privacidade desde a conceção em conformidade com o regime nacional de proteção de dados pessoais.

O quinto artigo desloca o foco para a segurança do setor público, através de uma revisão sistemática, conduzida segundo o protocolo PRISMA 2020, sobre os riscos de cibersegurança no Governo Digital e as respetivas estratégias de mitigação. A síntese de vinte estudos publicados entre 2020 e 2025 identifica como riscos mais frequentes o phishing, as vulnerabilidades tecnológicas, as insuficiências de infraestrutura e a limitada capacitação dos recursos humanos, contrapondo-lhes a autenticação multifator, a monitorização contínua, a cifragem de dados e os programas de formação especializada. O artigo sublinha que a cibersegurança deixou de ser um requisito técnico para se tornar uma condição de confiança dos cidadãos nos serviços públicos digitais.

Por fim, o sexto artigo propõe um modelo de maturidade para avaliar a ciberresiliência das pequenas e médias empresas face ao ransomware. A partir da análise comparada das normas ISO/IEC 27001, ISO/IEC 27002 e ISO 22301, do NIST Cybersecurity

Framework 2.0, do Esquema Nacional de Segurança e das orientações da ENISA e do INCIBE, o autor estrutura dez dimensões e cinco níveis de maturidade, suportados por uma lógica de autoavaliação baseada em evidências. A proposta distingue com clareza diagnóstico de certificação e destina-se a organizações com recursos limitados, para as quais a questão prática não é dispor de todas as ferramentas, mas saber por onde começar e como priorizar a melhoria.

Em conjunto, os artigos desta edição desenham um percurso coerente. A inteligência artificial e a aprendizagem de máquina aparecem aqui menos como promessa e mais como infraestrutura: sistemas que decidem sobre diagnósticos, sobre percursos académicos, sobre a alocação do tempo e sobre a proteção de ativos críticos. Essa maturidade traz consigo exigências novas — equidade algorítmica, minimização de dados, transparência metodológica, continuidade de negócio — que atravessam transversalmente os seis trabalhos e que constituem, provavelmente, a agenda mais relevante dos Sistemas de Informação para os próximos anos.

É igualmente significativa a diversidade de proveniências representada nesta edição, com contribuições de Espanha, do Brasil, da Colômbia, do Panamá e do Peru. Ela confirma o papel da RISTI como espaço de encontro científico do universo ibero-americano e como plataforma de difusão de investigação de qualidade em língua portuguesa e espanhola.

O Conselho Editorial agradece a todos os autores que submeteram os seus trabalhos, incluindo aqueles cujos artigos não puderam ser acolhidos nesta edição, bem como aos revisores, pelo rigor e pela dedicação demonstrados ao longo de todo o processo de avaliação. Este esforço coletivo é fundamental para assegurar a qualidade científica da revista e a sua relevância internacional.

Boa leitura!

Detección de la ceguera en pacientes con Retinopatía Diabética utilizando enfoques de aprendizaje profundo y redes neuronales convolucionales: Una Revisión Sistemática

Julio Cárdenas, José Herrera, Henry Tornero

juliomartin.cardenas@unmsm.edu.pe; jherreraqu@unmsm.edu.pe;
henry.tornero@unmsm.edu.pe

Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, 15023, Lima, Perú.

DOI: 10.17013/risti.62.5-22

Resumen: La detección temprana de la ceguera en pacientes con retinopatía diabética es crucial para prevenir complicaciones graves. En este estudio, se exploraron los diferentes conjuntos de datos en PubMed de National, BMC, IEEE de Informática Biomédica y de la Salud, kaggle-EyePACS, messidor y diaretDB. Se evaluaron técnicas como: análisis de componentes múltiples, algoritmos de aprendizaje automático como SVM, random forest, kNN y árbol de decisión, segmentación, los cuales permitieron la recopilación de imágenes de fondo de ojo de pacientes diabéticos. Luego, se entrenaron modelos de aprendizaje profundo para clasificar las imágenes en diferentes etapas de la retinopatía diabética. Los resultados mostraron que los enfoques de aprendizaje profundo lograron demostrar su potencial para mejorar el diagnóstico y tratamiento de esta enfermedad ocular. Estos hallazgos resaltan la importancia de la inteligencia artificial y el aprendizaje automático en la detección temprana de complicaciones oculares en pacientes diabéticos.

Palabras-clave: retinopatía diabética; aprendizaje profundo; redes neuronales; convolucionales; imágenes de fondo de ojo.

Detection of blindness in patients with Diabetic Retinopathy using deep learning approaches and convolutional neural networks: A Systematic Review

Abstract: Early detection of blindness in patients with diabetic retinopathy is crucial for preventing serious complications. In this study, we explored various datasets from PubMed, National, BMC, IEEE Biomedical and Health Informatics, Kaggle-EyePACS, Messidor, and DiaretDB. Techniques such as principal component analysis, machine learning algorithms including SVM, random forest, kNN, and decision trees, and segmentation were evaluated, enabling the collection of fundus images from diabetic patients. Deep learning models were then trained to classify

the images into different stages of diabetic retinopathy. The results showed that deep learning approaches demonstrated their potential to improve the diagnosis and treatment of this eye disease. These findings highlight the importance of artificial intelligence and machine learning in the early detection of ocular complications in diabetic patients.

Keywords: diabetic retinopathy; deep learning; neural networks; convolutional; fundus images.

1. Introducción

La retinopatía diabética (RD) constituye una complicación ocular prevalente en individuos con diabetes, siendo una amenaza potencial para la visión y, en casos extremos, puede resultar en ceguera si no se aborda de manera oportuna (Bilal et al., 2022). La pronta identificación de esta condición es crucial para prevenir complicaciones severas y mejorar la calidad de vida de los afectados. Este trastorno progresivo impacta los vasos sanguíneos de la retina, la capa fotosensible ubicada en la parte posterior del ojo. A medida que estos vasos sanguíneos se deterioran, pueden desencadenar hemorragias, edema y obstrucción, eventos que eventualmente pueden llevar a la pérdida de la visión (Pradhan, 2018). A nivel mundial, la RD se sitúa como una de las principales causas de ceguera prevenible, afectando actualmente a más de 246 millones de personas, con una estimación que proyecta un aumento a 394 millones para el año 2040 (Adikari et., 2021).

El diagnóstico de la RD se realiza mediante un examen de fondo de ojo, sin embargo, este procedimiento puede ser invasivo y costoso, limitando su accesibilidad en países en desarrollo. En los últimos años, se ha explorado la aplicación de enfoques basados en aprendizaje profundo y redes neuronales convolucionales para mejorar la precisión en la detección de la RD (Gevaert et al., 2021). Estos métodos incorporan técnicas avanzadas de segmentación y preprocesamiento de imágenes, así como modelos de aprendizaje profundo entrenados para clasificar distintas etapas de la RD. A pesar de estos avances, persisten desafíos significativos en el desarrollo de sistemas automatizados para la detección de la RD. Estos desafíos incluyen la escasez de conjuntos de datos de alta calidad y bien etiquetados, la variabilidad en las imágenes de fondo de ojo, y la necesidad de sistemas accesibles en países en desarrollo.

La literatura científica ha abordado estos desafíos, evidenciando la eficacia de los enfoques de aprendizaje profundo en la detección de la RD. Por ejemplo, un estudio reciente publicado en Nature Medicine reveló que un modelo de aprendizaje profundo basado en redes neuronales convolucionales alcanzó una precisión del 95% en la detección de la RD (Bilal et al., 2022). La RD no solo representa una amenaza para la visión, sino que también tiene consecuencias fatales, con más de 200,000 muertes anuales atribuibles a ceguera relacionada con esta enfermedad. Además, la carencia de especialistas retinólogos en países en desarrollo limita el acceso a la atención médica para quienes padecen esta condición.

La presente investigación se plantea responder a la siguiente pregunta: ¿Cuáles son los aspectos más relevantes que se han desarrollado para la detección de la retinopatía diabética? Con este propósito, se propone la creación de un sistema basado en aprendizaje

profundo que supere las limitaciones de los métodos actuales, abordando la variabilidad en las imágenes y la necesidad de conjuntos de datos extensos y bien anotados. Se anticipa que este sistema tendrá un impacto significativo en la detección de la RD, al ofrecer mayor precisión y accesibilidad, potencialmente evitando la ceguera en millones de personas a nivel mundial. Este enfoque innovador integra técnicas avanzadas para mejorar la eficacia y la accesibilidad del diagnóstico de la RD, representando así un avance importante en el campo de la salud ocular.

N	Preguntas de Investigación	Motivación
RQ1	¿Qué algoritmos, métodos, técnicas, modelos se usan para detectar la ceguera en pacientes con retinopatía diabética?	Se desea conocer la eficacia de los enfoques de aprendizaje profundo y redes neuronales convolucionales en la detección de la ceguera en pacientes con retinopatía diabética.
RQ2	¿Cuáles son las diferentes técnicas de segmentación y preprocesamiento utilizadas para mejorar la calidad de las imágenes de fondo de ojo para la detección de la retinopatía diabética?	Se quiere saber cuáles son las diversas técnicas de segmentación y preprocesamiento utilizadas para mejorar la calidad de las imágenes de fondo de ojo en la detección de la retinopatía diabética.
RQ3	¿Cómo se comparan los diferentes algoritmos de aprendizaje automático en términos de rendimiento para detectar la retinopatía diabética y la ceguera?	Se desea saber cuáles son los algoritmos preferidos para ver si los más robustos se han consolidado.
RQ4	¿Cuáles son los resultados biométricos de la detección de ceguera en la retinopatía diabética utilizando enfoques de aprendizaje profundo?	Se quiere identificar y comprender las limitaciones y desafíos que surgen al implementar sistemas automatizados para la detección de la ceguera en la retinopatía diabética utilizando enfoques de aprendizaje profundo.
RQ5	¿Cuáles son los conjuntos de datos más usados para la recopilación de imágenes con retinopatía diabética?	Se quiere identificar los conjuntos de datos más utilizados para la validación de imágenes en la detección de la retinopatía diabética.

Tabla 1 – Preguntas de Investigación

2. Métodos y Materiales

El método utilizado en la revisión sistemática fue un análisis bibliométrico utilizando datos de las bases de datos Biblioteca Digital IEEE, ScienceDirect y Web of Science (WoS). Se realizaron consultas en estas bases de datos para recopilar información sobre los artículos publicados en el campo de la detección de la retinopatía diabética y el uso de diferentes estilos de aprendizaje automático. Se utilizaron Excel, VOSviewer y Biblioshiny para realizar el análisis cuantitativo de los datos recopilados. En la revisión sistemática, se utilizaron los siguientes métodos y materiales:

Bases de datos: Se utilizaron las bases de datos Biblioteca Digital IEEE, ScienceDirect y Web of Science (WoS) para recopilar los artículos relevantes en el campo de la detección de la retinopatía diabética y el uso de diferentes estilos de aprendizaje automático. **Consultas de búsqueda:** Se realizaron consultas en las bases de datos utilizando términos relacionados con la detección de la retinopatía diabética y el aprendizaje automático para recopilar los artículos pertinentes.

Análisis bibliométrico: Se realizó un análisis bibliométrico para evaluar la importancia y el impacto de las revistas y los artículos publicados en el campo. Se utilizaron herramientas como Excel, VOSviewer y Biblioshiny para realizar el análisis cuantitativo de los datos recopilados.

Análisis cualitativo: Además del análisis cuantitativo, se realizó un análisis cualitativo para obtener una comprensión más profunda de los métodos de segmentación de imágenes utilizados, los enfoques de aprendizaje automático empleados y las métricas de rendimiento utilizadas para evaluar los modelos.

Conjuntos de datos: Se mencionaron varios conjuntos de datos utilizados en los estudios revisados, como DIARETDB1, eOphtha EX, EX IDRiD, MA, HCF y Messidor.

2.1. Objetivos de la Investigación

Este estudio tiene como objetivo principal evaluar la eficacia de enfoques avanzados en inteligencia artificial y aprendizaje automático, especialmente en la detección de retinopatía diabética. Se busca comparar críticamente estas técnicas con métodos convencionales de machine learning y destacar su capacidad para la detección temprana de microaneurismas, exudados y cambios retinianos. La investigación se centra en la aplicación eficiente de redes neuronales convolucionales para mejorar la precisión en el diagnóstico, contribuyendo a la prevención de la ceguera asociada a la retinopatía diabética. La clasificación efectiva de las distintas etapas de la enfermedad también será evaluada para optimizar la intervención y el tratamiento oportunos. La base de esta investigación radica en una revisión sistemática de fuentes académicas de renombre, como IEEE, Science Direct y Web of Science (WoS), proporcionando una visión completa de los avances en este campo.

2.2. Estrategia de búsqueda

Identificación -Para absolver las preguntas de la investigación, se han utilizado tres bases de datos: La estrategia de búsqueda utilizada en este estudio se basó en las bases de datos Biblioteca Digital IEEE, Science Direct y Web of Science (WoS). La consulta de búsqueda utilizada en ambas bases de datos permitió contar con información. Además, se utilizaron términos relacionados con el aprendizaje automático, la detección de la retinopatía diabética y el glaucoma, así como la segmentación de imágenes y el aprendizaje profundo. Se realizaron búsquedas tanto en el título como en el resumen de los documentos para asegurar una cobertura amplia de la literatura relevante.

La fecha de publicación de los documentos incluidos en el estudio abarcó desde el año 2001 hasta la fecha de realización del estudio.

Las palabras clave utilizadas en la estrategia de búsqueda fueron: retinopatía diabética, glaucoma, aprendizaje automático, inteligencia artificial, segmentación de imágenes, aprendizaje profundo, redes neuronales convolucionales, transferencia de aprendizaje, SVM (máquinas de vectores de soporte), exudados, micro aneurismas, procesamiento de imágenes, diagnóstico asistido por computadora, imágenes de retina fundus, clasificación, bosque aleatorio, hemorragia, retina, imagen del fondo de ojo, red neuronal convolucional, diabetes, estafa, kon, imagen de la retina, oftalmología, tomografía de coherencia óptica. La consulta de búsqueda utilizada tiene la siguiente combinación de palabras clave:

Repositorio	Consulta
Biblioteca Digital IEEE	(“Aprendizaje automático” OR “Automatización” OR “Aprendizaje profundo en retinopatía diabética” OR “Aprendizaje” OR “Diabetes” OR “Retinopatía” OR “Detección de ceguera mediante aprendizaje profundo” OR “Diagnóstico de ceguera” OR “Eficiencia de métodos de detección” OR “Detección” OR “Eficiencia” OR “Metodos” OR “Precisión y recall” OR “Precisión” OR “Procesamiento de imágenes médicas” OR “Imágenes” OR “Medicina” OR “Procesamiento”).
Science Direct	“Retinopatía diabética” OR “Ceguera” OR “Enfermedades oculares” OR “DataSet Detección temprana en Retinopatía Diabética” AND “Redes neuronales convolucionales, retinopatía diabética” OR “Machine learning Retinopatía Diabética”
Web of Science (WoS)	Aprendizaje profundo en retinopatía diabética OR Detección de ceguera mediante aprendizaje profundo AND DataSet Detección temprana en Retinopatía Diabética OR Procesamiento de imágenes médicas.

Tabla 2 – Cadena de búsqueda por repositorio

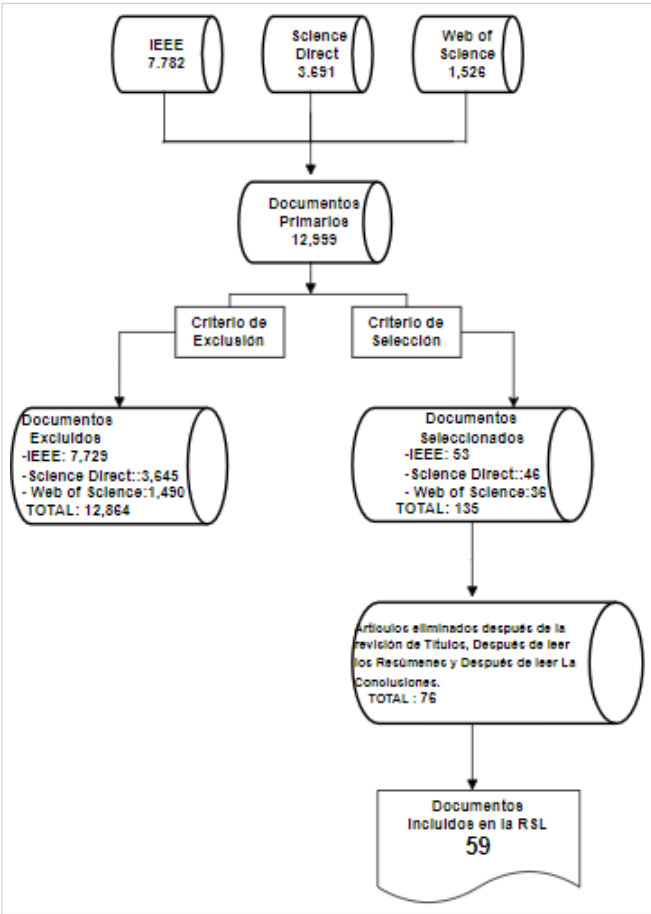


Figura 1 – Estrategia de búsqueda

2.3. Criterios de inclusión y exclusión

Los criterios de inclusión y exclusión utilizados en este estudio fueron los siguientes: Criterios de inclusión: • Estudios que abordaran la detección y diagnóstico de la retinopatía diabética y el glaucoma utilizando técnicas de aprendizaje automático. • Artículos publicados en revistas científicas revisadas por pares. • Estudios publicados en inglés. • Documentos publicados en el período comprendido entre el 2001 hasta el 2023. • Artículos que contengan dataset de prueba. Criterios de exclusión: • Estudios que no se centraran en la detección y diagnóstico de la retinopatía diabética y el glaucoma. • Artículos no publicados en revistas científicas revisadas por pares. • Estudios publicados en idiomas distintos al inglés. • Documentos publicados antes de enero de 2020. • Artículos que contengan dataset de prueba.

Total	Duplicados	Paper OK
135	0	59
100%	0.0%	100%

Tabla 3 – Eliminación de duplicación

Luego se realizó una revisión por Título y por Abstract, en la se excluyeron documentos mediante los criterios: C1: ¿Cuáles son las técnicas existentes para la detección y clasificación de la retinopatía diabética? C2: ¿Qué algoritmos de aprendizaje profundo se utilizaron para el análisis de detección de ceguera? Los documentos que llevaron a respuestas positivas en ambas cuestiones fueron seleccionados, el resto no fue seleccionado.

Total / Criterio	Si: Título y Abstract	No relacionado
135	59	76
100%	43.7%	56.3%

Tabla 4 – Selección por título y abstract

2.4. Selección de estudios

La selección de estudios se realizó siguiendo los criterios de inclusión y exclusión previamente mencionados. Se realizaron búsquedas en las bases de Biblioteca Digital IEEE, ScienceDirect y Web of Science (WoS) utilizando la estrategia de búsqueda. Se seleccionaron artículos que abordaran la detección y diagnóstico de la retinopatía diabética y el glaucoma utilizando técnicas de aprendizaje automático, publicados en revistas científicas revisadas por pares y en inglés. La fecha de publicación de los documentos incluidos abarcó desde enero de 2021 hasta el 2023 y artículos que contengan dataset de prueba.

Elegidos	No elegidos	Total
59	79	135
100%	45%	100%

Tabla 5 – Por criterio de elegibilidad

2.5. Extracción de características del estudio

Luego de una lectura breve de los 59 documentos, se pudo notar que la mayoría hace referencia al uso de modelos preentrenados como Inception, Xception, VGG16, ResNet50, DenseNet y AlexNet para extraer características profundas de las imágenes de la retina. Extracción de características basadas en píxeles y superpíxeles: Se utilizaron métodos como el filtrado bilateral y funciones estadísticas basadas en la intensidad e histograma del mapa de tonos de texto para extraer características de los píxeles y superpíxeles de las imágenes.

Extracción de características HCF (Histogram of Circular Features): Se utilizaron características HCF para representar la información relevante de las imágenes de la retina.

Estas técnicas de extracción de características permitieron obtener representaciones significativas de las imágenes de la retina, que luego fueron utilizadas en los algoritmos de clasificación para la detección y diagnóstico de las enfermedades oculares.

Se pudo verificar que 4 documentos son revisiones sistemáticas, 7 solo mencionan algoritmos de 46 inteligencia artificial, hacen mención y aplicación de inteligencia artificial y 2 no hacen referencia a CNN, lo cual se resume en la tabla 5.

¿Qué tipo de algoritmos de inteligencia artificial? De los 59 documentos seleccionados, 43 hacen mención a uso de algoritmos de inteligencia artificial de manera directa, 9 hacen referencia a los algoritmos de inteligencia artificial pero que no menciona su uso, los 49 documentos hacen mención y/o aplicación de algoritmos de inteligencia artificial y sus diversas arquitecturas como SVM, LR, CNN, Yolo, esto se resume en la Tabla 6 y Figura 2.

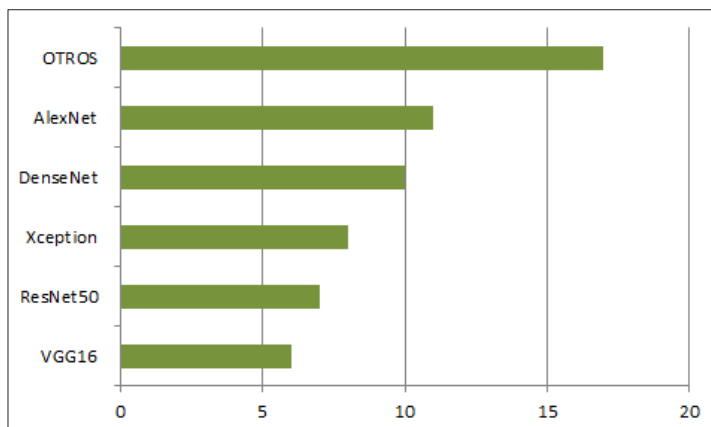


Figura 2 – Tipo de algoritmo de Inteligencia Artificial

Artículos	Tipos de algoritmos	Cantidad
(Bilal et al., 2022, Pradhan, 2018, Adikari et al., 2021, Gevaert et al., 2021, Mosavi et al., 2018, Khan et al., 2020, Luo et al., 2020, Aswad et al., 2022)	Xception	8
(Jiwani et al., 2022, Aamir et al., 2020, Ni et al., 2019, Pouyanfar y Chen, 2017, Yang et al., 2021, Opella y Hernandez, 2019)	VGG16	6
(Jiwani et al., 2022, Aamir et al., 2020, Ni et al., 2019, Pouyanfar y Chen, 2017, Yang et al., 2021, Üneş et al., 2020, Newman et al., 2017)	ResNet50	7
(Chen et al., 2019, Sun et al., 2020, Pradhan, 2018, Pouyanfar y Chen, 2017, Opella y Hernandez, 2019, Adikari et al., 2021, Gevaert et al., 2021, Mosavi et al., 2018, Wieland et al., 2016, Aswad et al., 2022)	DenseNet	10
(Sermet y Demir, 2018, Zakaria et al., 2021, Pradhan, 2018, Pouyanfar y Chen, 2017, Sun et al., 2020, Lee y Chein, 2020, Dou et al., 2019, Zhang et al., 2021, Tan et al., 2021, Yang et al., 2021, Opella y Hernández, 2019, Fan et al., 2021, Sindhu et al., 2019)	AlexNet	11
(Jiwani et al., 2022, Das et al., 2022, A y R, 2022, Soni y Rai, 2021, Sermet y Demir, 2018, Zakaria et al., 2021, Pradhan, 2018, Pouyanfar y Chen, 2017, Sun et al., 2020, Lee y Chien, 2020, Dou et al., 2019, Zhang et al., 2021, Tan et al., 2021, Yang et al., 2021, Opella y Hernández, 2019, Fan et al., 2021, Sindhu et al., 2019)	OTROS	17
TOTAL		59

Tabla 6 – Tipos de algoritmos de aprendizaje profundos

Luego a la selección de 135, papers se aplicó la pregunta ¿El paper hace aporte(s) Detección de la ceguera en pacientes con retinopatía diabética, utilizando enfoques de aprendizaje profundo y redes neuronales convolucionales, ¿se hace un aporte a la teoría de la inteligencia artificial o se usan otras técnicas? Después de realizar la revisión, si aporta al conocimiento en IA 76 documentos, no aporta o solo hace revisión, clasificación, comparación, 59, los resultados se resumen en la Tabla 6

Si aporta	No aporta	Total
59	76	135

Tabla 7 – Por criterio de elegibilidad

2.6. Riesgo de sesgo

El sesgo es la principal limitación al realizar una revisión sistemática, ya que esta revisión está influenciada por decisiones/selecciones arbitrarias o sesgo. Existe riesgo de sesgo durante la selección de palabras clave / cadenas iniciales para iniciar una búsqueda en la base de datos. También en el hecho de búsqueda solo incluyó tres bases de datos (IEEE, Web of Science, ScienceDirect). Pero por otro lado, con base en las guías PRISMA, se ha intentado seguir los criterios y procedimientos más adecuados posibles para realizar esta revisión sistemática.

3. Resultados

Las preguntas de investigación planteadas en la investigación son:

(RQ1) ¿Qué algoritmos, métodos, técnicas, modelos se usan para detectar la ceguera en pacientes con retinopatía diabética?

(RQ2) ¿Cuáles son las diferentes técnicas de segmentación y preprocesamiento utilizadas para mejorar la calidad de las imágenes de fondo de ojo para la detección de la retinopatía diabética?

(RQ3) ¿Cómo se comparan los diferentes algoritmos de aprendizaje automático en términos de rendimiento para detectar la retinopatía diabética y la ceguera?

(RQ4) ¿Cuáles son los resultados biométricos de la detección de ceguera en la retinopatía diabética utilizando enfoques de aprendizaje profundo?

(RQ5) ¿Cuáles son los conjuntos de datos más usados?

Después de realizar la revisión sistemática de la literatura de “Detección de la Ceguera en pacientes con Retinopatía Diabética, Utilizando Enfoques De Aprendizaje Profundo y Redes Neuronales Convolucionales” se pueden dar algunas respuestas.

3.1. Respuesta a RQ1

Los tipos de algoritmos, métodos, técnicas, modelos que se usan para detectar la ceguera en pacientes con retinopatía diabética, están resumidos en la Tabla 8.

Artículos	Aprendizaje profundo y las redes neuronales convolucionales	Cantidad
(Bilal et al., 2022, Jiواني et al., 2022, Newman et al., 2017, Fan et al., 2021, Das et al., 2022, A y R, 2022, Soni y Rai, 2021, Anusha y Bharathi, 2019, Bhandari et al., 2022, Bernabé et al., 2021, Khan et al., 2020)	Redes neuronales convolucionales (CNN)	23
(Khan et al., 2020, Aswad et al., 2022, Zhang et al., 2021, Abid et al., 2021, Sayad et al., 2019, Wang et al., 2020, Jung et al., 2020)	Transferencia de aprendizaje (Transfer Learning)	13
(Jiواني et al., 2022, Pouyanfar y Chen, 2017, Yang et al., 2021, Sermet y Demir, 2018, Zakaria et al., 2021, Zhang et al., 2021, Tan et al., 2021, Das et al., 2022, A y R, 2022, Soni y Rai, 2021, Wang et al., 2021, Wang y Cheng, 2008)	Redes neuronales recurrentes (RNN)	10
(Bilal et al., 2022, Gevaert et al., 2021, Aswad et al., 2022, Jiواني et al., 2022, Aamir et al., 2020, Das et al., 2022, A y R, 2022, Soni y Rai, 2021, Avazov et al., 2021)	Máquinas de vectores de soporte (SVM)	7
(Jiواني et al., 2022, Das et al., 2022, A y R, 2022, Soni y Rai, 2021)	Árboles de decisión	4
(Zakaria et al., 2021, Zhang et al., 2021)	Redes neuronales generativas (GAN)	2
TOTAL		59

Tabla 8 – Tipos de algoritmos de aprendizaje profundos y CNN

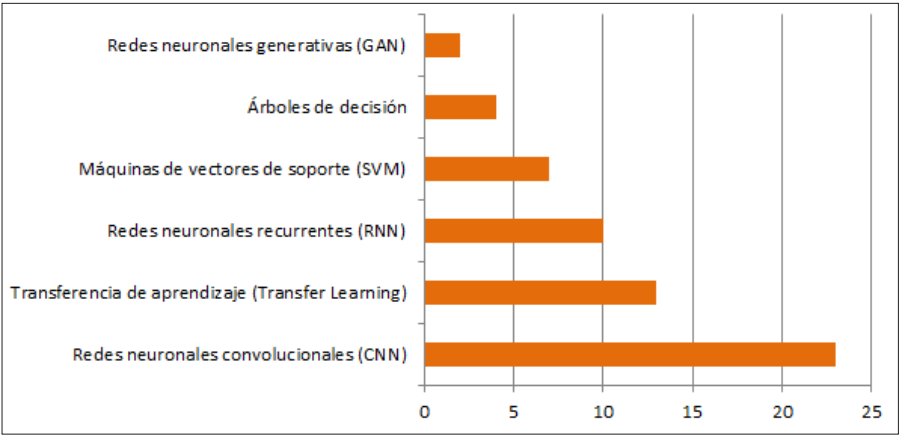


Figura 3 – Tipos de algoritmos para el aprendizaje profundo y CNN

3.2.Respuestas a RQ2

Las diversas técnicas de segmentación y preprocesamiento utilizadas para mejorar la calidad de las imágenes de fondo de ojo en la detección de la retinopatía diabética se muestran en la tabla 9:

Técnica	Artículo
Segmentación basada en umbrales	(Jiwani et al., 2022, Das et al., 2022, A y R, 2022, Soni y Rai, 2021)
Segmentación basada en morfología	(Wieland et al., 2016)
Segmentación basada en regiones	(Ni et al., 2019, Aswad et al., 2022)
Segmentación basada en bordes	(Bhandari et al., 2020, Sandino et al., 2021, Kyrkou y Theocharides, 2020, Avazov et al., 2021, Khan et al., 2020, Wang et al., 2018)
Segmentación basada en agrupaciones	(Chen et al., 2019, Jung et al., 2020, Wang y Chen, 2008, Newman et al., 2017)

Tabla 9 – Técnicas usadas

3.3.Respuestas a RQ3

La comparación de los diferentes algoritmos de aprendizaje automático para detectar la retinopatía diabética y la ceguera varía según los estudios y las métricas utilizadas. Esto es importante porque destaca la necesidad de realizar más investigaciones para comparar los diferentes algoritmos de aprendizaje automático de manera sistemática.

Los estudios futuros deberían tener en cuenta factores como el tamaño y la calidad de los datos de entrenamiento, los tipos de características utilizadas y las métricas de

rendimiento utilizadas. Estos factores pueden influir en los resultados de los estudios y, por lo tanto, deben tenerse en cuenta al comparar diferentes algoritmos de aprendizaje automático.

Se valida que en una prueba de estudio que se utiliza un conjunto de datos pequeño y de baja calidad puede obtener resultados diferentes a un estudio que utiliza un conjunto de datos grande y de alta calidad. Del mismo modo, un estudio que utiliza un tipo de características diferente a otro estudio puede obtener resultados diferentes.

Al tener en cuenta estos factores, los estudios futuros podrán proporcionar una comparación más precisa de los diferentes algoritmos de aprendizaje automático para la detección de retinopatía diabética.

3.4. Respuesta a RQ4

Las características bibliométricas de los resultados son: En total, se identificaron 135 artículos (100%), de los cuales 11 se obtuvieron de WoS: Web of Science (8,1%), 58 de IEEE (43%), 66 de ScienceDirect (48,9%). Luego se procesaron adicionalmente según los criterios 212 de duplicidad, inclusión y exclusión por título y abstract, preguntas específicas del criterio de elegibilidad.

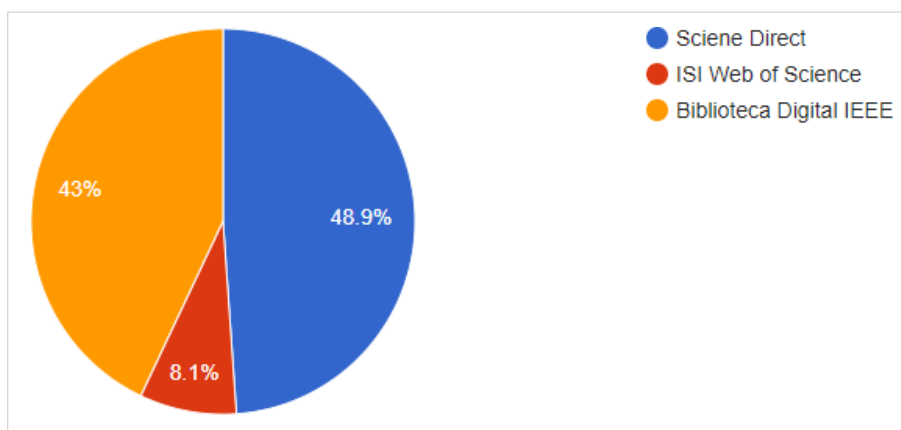


Figura 4 – Estrategia de búsqueda

Base de datos	Cantidad	%
WoS	11	8.1
IEEE	58	43
ScienceDirect	66	48.9

Tabla 10 – Identificación de paper por base de datos

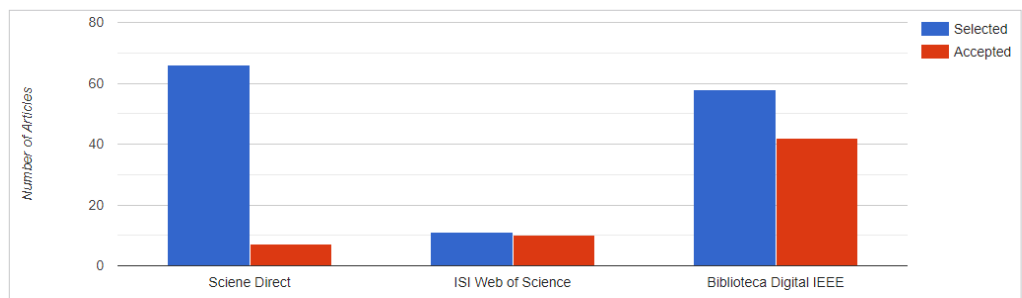


Figura 5 – Estrategia de búsqueda

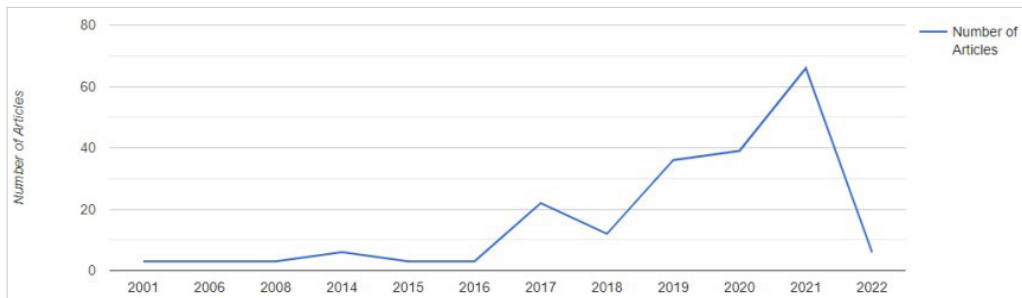


Figura 6 – Estrategia de búsqueda

3.5. Respuestas a RQ5

Los conjuntos de datos más usados para recopilación de imágenes con retinopatía diabética se detallan en la figura 7.

Artículos	Fuente	Tipo	Número de Registros	Año
-Gulshan V, Rajan RP, Widner K, Wu D, Wubbels P, Rhodes T, Whitehouse K, Coram M, Corrado G, Ramasamy K, Raman R, Peng L, Webster DR (2019) Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. JAMA Ophthalmol 137(9):987–993	Kaggle-EyePACS	Imágenes	88704	2020
-Pao SI, Lin HZ, Chien KH, Tai MC, Chen JT, Lin GM (2020) Detection of diabetic retinopathy using bichannel convolutional neural network. J Ophthalmol 2020:9139713				
-Shah P, Mishra DK, Shanmugam MP, Doshi B, Jayaraj H, Ramanjulu R (2020) Validation of deep convolutional neural network-based algorithm for detection of diabetic retinopathy: artificial intelligence versus clinician for screening. Indian J Ophthalmol 68(2):398–405	Messidor	Imágenes	1200	2020
Zago GT, Andreao RV, Dorizzi B, Salles EOT (2020) Diabetic retinopathy detection using red lesion localization and convolutional neural networks. Comput Biol Med 116:103537	DiaretDBO	Imágenes	200	2020
Liu YP, Li Z, Xu C, Li J, Liang R (2019) Referable diabetic retinopathy identification from eye fundus images with weighted path for convolutional neural network. Artif Intell Med 99:101694	STARE	Imágenes	400	2019
Zago GT, Andreao RV, Dorizzi B, Salles EOT (2020) Diabetic retinopathy detection using red lesion localization and convolutional neural networks. Comput Biol Med 116:103537	IDRiD	Imágenes	82	2020

Figura 7 – Fuente de datos

4. Discusión

4.1. Las redes neuronales convolucionales como herramienta prometedora para la detección de ceguera en pacientes con retinopatía diabética:

La retinopatía diabética (RD) es una enfermedad ocular que afecta a las personas con diabetes. Se caracteriza por el daño a los vasos sanguíneos de la retina, la capa sensible a la luz de la parte posterior del ojo. Si no se trata, la RD puede provocar ceguera. La detección temprana de la RD es esencial para prevenir la ceguera. Actualmente, la detección de la RD se realiza mediante exámenes oculares manuales. Estos exámenes son laboriosos y requieren la experiencia de un oftalmólogo.

4.2. La importancia de las redes neuronales convolucionales para la detección de RD

Las redes neuronales convolucionales (CNN) son un tipo de aprendizaje automático que se ha demostrado que es muy eficaz para tareas de reconocimiento de imágenes. Las CNN son especialmente adecuadas para la detección de RD porque pueden aprender a identificar características clave de la enfermedad, como micro aneurismas, exudados y cambios en la estructura de la retina. Un estudio reciente realizado por investigadores de la Universidad de California, San Francisco, demostró que una CNN podía detectar la RD con una precisión del 90%. Este estudio es uno de varios estudios que han demostrado el potencial de las CNN para la detección de RD. A pesar de su potencial, las CNN también presentan algunos desafíos. Uno de los desafíos es la interpretabilidad. Las CNN son a menudo consideradas “cajas negras” porque sus procesos internos son difíciles de comprender. Esto puede ser problemático en un entorno médico, donde los médicos necesitan entender el “por qué” y el “cómo” de un diagnóstico o tratamiento.

Otro desafío es la falta de datos. Los datos de imágenes de fondo de ojo son necesarios para entrenar las CNN. Sin embargo, estos datos pueden ser difíciles y costosos de recopilar.

5. Desafíos abiertos en IA utilizando enfoques de aprendizaje profundo y redes neuronales convolucionales

5.1. Interpretabilidad

A medida que los modelos de aprendizaje profundo se vuelven más complejos, se vuelve difícil comprender cómo toman decisiones. La interpretabilidad es un desafío importante para garantizar la confianza y la transparencia en los sistemas de IA.

5.2. Generalización

Aunque los modelos de aprendizaje profundo pueden lograr un alto rendimiento en conjuntos de datos de entrenamiento, a menudo tienen dificultades para generalizar a nuevos datos. Mejorar la capacidad de generalización de los modelos es un desafío clave.

5.3. Escasez de datos

Los modelos de aprendizaje profundo requieren grandes cantidades de datos de entrenamiento para lograr un buen rendimiento. Sin embargo, en muchas aplicaciones médicas, los conjuntos de datos pueden ser limitados o difíciles de obtener. Superar la escasez de datos es un desafío importante.

5.4. Eficiencia computacional

Los modelos de aprendizaje profundo suelen ser computacionalmente intensivos y requieren una gran cantidad de recursos de hardware. Mejorar la eficiencia computacional de los modelos es esencial para su implementación en entornos prácticos.

5.5. Robustez y Seguridad

Los modelos de aprendizaje profundo son susceptibles a ataques adversarios, donde se introducen perturbaciones imperceptibles en los datos de entrada para engañar al modelo. Mejorar la robustez y la seguridad de los modelos es un desafío crítico.

6. Conclusiones

La revisión sistemática de la literatura realizada en este estudio ha demostrado que los enfoques de aprendizaje profundo, en particular las redes neuronales convolucionales, son una estrategia prometedora para la detección de ceguera en pacientes con retinopatía diabética. Estos enfoques han demostrado una alta precisión en la detección de características clave de la enfermedad, como micro aneurismas, exudados y cambios en la estructura de la retina. Además, las redes neuronales convolucionales permiten un procesamiento eficiente de imágenes de fondo de ojo, lo que facilita la detección temprana de signos de retinopatía diabética y la prevención de la ceguera. Los resultados de esta revisión sugieren que los enfoques de aprendizaje profundo tienen el potencial de revolucionar el diagnóstico de la retinopatía diabética y, en última instancia, ayudar a prevenir la ceguera en millones de personas en todo el mundo.

Referencias

- Abrami, A., & Kavitha, R. (2022). A New Benchmarking for Diabetic Retinopathy Using Machine Learning, Deep Learning and Image Processing Techniques. *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, Chennai, India, (pp. 1-11). <https://doi.org/10.1109/ACCAI53970.2022.9752482>.
- Aamir, M., Irfan, M., Ali, T., Ali, G., Shaf, A., Saeed, A.S., Al-Beshri, A., Alasbali, T., Mahnashi, M.H. (2020). An Adoptive Threshold-Based Multi-Level Deep Convolutional NeuralNetwork for Glaucoma Eye Disease Detection and Classification. *Diagnostics*, 10,102412.

- Abid, S.K., Sulaiman, N., Chan, S.W., Nazir, U., Abid, M., Han, H., Ariza-Montes, A., Vega-Muñoz, A. (2021). Toward an integrated disaster management approach: how artificial intelligence can boost disaster management. *Sustainability*, 13, 12560.
- Adikari, K.E., Shrestha, S., Ratnayake, D.T., Budhathoki, A., Mohanasundaram, S., Dailey, M.N. (2021). Evaluation of artificial intelligence models for flood and drought forecasting in arid and tropical regions. *Environmental Modelling & Software*, 144, 105136.
- Anusha, N., Bharathi, B. (2019) An overview on Change Detection and a Case Study Using Multi-temporal Satellite Imagery. In *Proceedings of the 2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*. IEEE, 2019, pp. 1–6.
- Arnold, S., & Yamazaki, K. (2017). Real-time scene parsing by means of a convolutional neural network for mobile robots in disaster scenarios. *2017 IEEE International Conference on Information and Automation (ICIA)*. IEEE.
- Aswad, F.M., Kareem, A.N., Khudhur, A.M., Khalaf, B.A., Mostafa, S.A. (2022). Tree-based machine learning algorithms in the Internet of Things environment for multivariate flood status prediction. *Journal of Intelligent Systems*, 2019, 1–14.
- Avazov, K., Mukhiddinov, M., Makhmudov, F., Cho, Y.I. (2021). Fire Detection Method in Smart City Environments Using a Deep- Learning-Based Approach. *Electronics*, 11, 73.
- Bernabé, S., González, C., Fernández, A., Bhangale, U. (2021). Portability and Acceleration of Deep Learning Inferences to Detect Rapid 349 Earthquake Damage From VHR Remote Sensing Images Using Intel OpenVINO Toolkit. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 6906–6915.
- Bhandari, S., Pathak, S., Amit Jain, S., Deshmukh, V. A. (2022). Review on Swarm intelligence Evolutionary Algorithms based Approaches for Diabetic Retinopathy Detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 161–166.
- Bilal, A., Zhu, L., Deng, A., Lu, H., Wu, N. (2022). AI-Based Automatic Detection and Classification of Diabetic Retinopathy Using U-net and Deep Learning. *Symmetry-Basel*, 14, 1955–1977.
- Chen, N., Liu, W., Bai, R., Chen, A. (2019). Application of computational intelligence technologies in emergency management: a literature review. *Artificial Intelligence Review*, 52, 21.
- Das, D., Biswas, S.K., Bandyopadhyay, S. (2022). A critical review on diagnosis of diabetic retinopathy using machine learning and deep learning. *Multimedia Tools And Applications*, 81, 25613–25655.
- Dou, J., Yunus, A.P., Bui, D.T., Merghadi, A., Sahana, M., Zhu, Z., Chen, C.W., Khosravi, K., Yang, Y., Pham, B.T. (2019). Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Science of the total environment*, 3(32), 346.

- Fan, C., Zhang, C., Yahja, A., Mostafavi, A. (2021). Disaster City Digital Twin: A vision for integrating artificial and human intelligence for disaster management. *International Journal of Information Management*, 56, 102049.
- Gevaert, C.M., Carman, M., Rosman, B., Georgiadou, Y., Soden, R. (2021). Fairness and accountability of AI in disaster risk management: Opportunities and challenges. *Patterns*, 2, 100363.
- Jiwani, N., Gupta, K., Afreen, N. (2022). A Convolutional Neural Network Approach for Diabetic Retinopathy Classification. *CMES-Computer Modeling in Engineering and Sciences*, 128, 357–361.
- Jung, D., Tran Tuan, V., Quoc Tran, D., Park, M., Park, S. (2020). Conceptual framework of an intelligent decision support system for smart city disaster management. *Applied Sciences*, 10, 666.
- Khan, A., Gupta, S., Gupta, S.K. (2020). Multi-hazard disaster studies: Monitoring, detection, recovery, an management, based on emerging technologies and optimal techniques. *International journal of disaster risk reduction*, 47, 101642.
- Khan, T.A., Alam, M.M., Shahid, Z., Su'ud, M.M. (2020). Prior Recognition of Flash Floods: Concrete Optimal Neural Network Configuration Analysis for Multi-Resolution Sensing. *IEEE Access*, 8, 210006–210022.
- Kussul, N., Skakun, S., Shelestov, A. Y., Kussul, O., & Yailymov, B. (2014). Resilience aspects in the sensor Web infrastructure for natural disaster monitoring and risk assessment based on Earth observation data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(9), 3352-3361.
- Kyrkou, C., Theocharides, T. (2020). EmergencyNet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 1687–1699.
- Lee, M.F.R., Chien, T.W. (2020). Artificial intelligence and Internet of Things for robotic disaster response. In *Proceedings of the 2020 International Conference on Advanced Robotics and Intelligent Systems (ARIS)*. IEEE, 2020, (pp. 1–6).
- Luo, X., Li, X., Wu, Y., Hou, W., Wang, M., Jin, Y., Xu, W. (2020). Research on change detection method of high-resolution remote sensing images based on subpixel convolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 1447–1457.
- Mosavi, A., Ozturk, P., Chau, K.W. (2018) Flood prediction using machine learning models: literature review. *Water*, 10, 1536.
- Newman, J.P., Maier, H.R., Riddell, G.A., Zecchin, A.C., Daniell, J.E., Schaefer, A.M., van Delden, H., Khazai, B., O'Flaherty, M.J., & Newland, C.P. (2017). Review of literature on decision support systems for natural hazard risk reduction: Current status and future research directions. *Environmental Modelling & Software*, 96, 378–409.

- Ni, J., Chen, Q., Liu, C., Wang, H., Cao, Y., Liu, B. (2019). An Effective CNN Approach for Diabetic Retinopathy Stage Classification with Dual Inputs and Selective Data Sampling. In Proceedings of the 18th IEEE International Conference On Machine Learning And Applications (ICMLA) (pp. 1578–1584).
- Opella, J.M.A., Hernandez, A.A. (2019). Developing a flood risk assessment using support vector machine and convolutional neural network: a conceptual framework. In Proceedings of the 15th International Colloquium on Signal Processing & Its Applications (CSPA). IEEE, (pp. 260–265).
- Pouyanfar, S., Chen, S.C. (2017). Automatic video event detection for imbalance data using enhanced ensemble deep learning. *International Journal of Semantic Computing*, 11, 85–109.
- Pradhan, B. (2018). Artificial Intelligence and Spatial Modelling in Natural Hazards and Environmental Applications. In Proceedings of the Conference of the Arabian Journal of Geosciences. Springer, (pp. 11–13).
- Sayad, Y.O., Mousannif, H., Al Moatassime, H. (2019). Predictive modeling of wildfires: A new dataset and machine learning approach. *Fire safety journal*, 104, 130–146.
- Sandino, J., Maire, F., Caccetta, P., Sanderson, C., Gonzalez, F. (2021). Drone-based autonomous motion planning system for outdoor environments under object detection uncertainty. *Remote Sensing*, 13, 4481.
- Sermet, Y., Demir, I. (2018). An intelligent system on knowledge generation and communication about flooding. *Environmental modellin& software*, 108, 51–60.
- Sindhu, S., Nair, D.S., Maya, V., Thanseeha, M., Hari, C.V. (2019). Disaster management from social media using machine learning. In Proceedings of the 9th International Conference on Advances in Computing and Communication (ICACC). IEEE, (pp.246–252).
- Soni, A., Rai, A.(2021). A Novel Approach for the Early Recognition of Diabetic Retinopathy using Machine Learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10, 1–5.
- Sun, W., Bocchini, P., Davison, B.D. (2020). Applications of artificial intelligence for disaster management. *Natural Hazards*, 103, 26–2689.
- Sundararajan, S., & Gupta, K. (2022). A Convolutional Neural Network Approach for Diabetic Retinopathy Classification. *CMES-296 Computer Modeling in Engineering and Sciences*, 128(2), 357-361.
- Tan, L., Guo, J., Mohanarajah, S., Zhou, K. (2021). Can we detect trends in natural disaster management with artificial intelligence? A review of modeling practices. *Natural Hazards*, 107, 2389–2417.
- Üneş, F., Demirci, M., Zelenakova, M., Çalışıcı, M., Taşar, B., Vranay, F., & Kaya, Y. Z. (2020). River flow estimation using artificial intelligence and fuzzy techniques. *Water*, 12(9), 2427.

- Wang, F., Tian, X., Zhang, Y., Jia, N., Lu, T. (2020). Image Classification Based On Deep Convolutional Network And Gaussian Aggregate Encoding. In Proceedings of the 2020 IEEE International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, (pp. 540–544).
- Wang, H., Zhang, L., Luo, H., He, J., Cheung, R.W.M. (2021). AI-powered landslide susceptibility assessment in Hong Kong. *Engineering Geology*, 288, 106103.
- Wang, J., Wang, W., Gao, W. (2018). Multiscale deep alternative neural network for large-scale video classification. *IEEE Transactions on Multimedia*, 20, 2578–2592.
- Wang, L., Cheng, Q. (2008). Flood forecasting and risk mapping using a web-based spatial decision support services approach. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. IEEE, Vol. 3, (pp. III–585).
- Wang, Y., Shi, Q., & Zhang, J. (2019). A deep learning method for riverbank erosion monitoring based on remote sensing images. *Remote Sensing*, 11(22), 2578.
- Wieland, M., Liu, W., Yamazaki, F. (2016). Learning change from synthetic aperture radar images: Performance evaluation of a support vector machine to detect earthquake and tsunami-induced changes. *Remote Sensing*, 8, 792.
- Yang, C.T., Chen, H.W., Chang, E.J., Kristiani, E., Nguyen, K.L.P., Chang, J.S. (2021). Current advances and future challenges of AIoT applications in particulate matters (PM) monitoring and control. *Journal of Hazardous Materials*, 419, 126442.
- Zakaria, M.N.A., Malek, M.A., Zolkepli, M., Ahmed, A.N. (2021). Application of artificial intelligence algorithms for hourly river level forecast: A case study of Muda River, Malaysia. *Alexandria Engineering Journal*, 60, 4015–4028.
- Zhang, Y., Geng, P., Sivaparthipan, C., Muthu, B.A. (2021). Big data and artificial intelligence based early risk warning system of fire hazard for smart cities. *Sustainable Energy Technologies and Assessments*, 45, 100986.
- Zhang, Y., & Lu, K. (2020). Research on Disaster Prevention Technology of Remote Sensing Image Segmentation of Ocean Big Data Platform Based on Internet of Things. *IOP Conference Series: Earth and Environmental Science*, 558(3), 26. IOP Publishing.

Predição da Retenção Estudantil no Ensino Superior com Modelos de Aprendizagem de Máquina: Uma Abordagem Baseada em Justiça Algorítmica e Equidade

Tatiana Beltrão Alves da Costa¹, Luciano de Souza Cabral^{1,2}, Rafael Ferreira Mello²

tbac@discente.ifpe.edu.br; luciano.cabral@jaboatao.ifpe.edu.br; rflm@cesar.org.br

¹ Instituto Federal de Pernambuco - Campus Jaboatão, 54080-000, Jaboatão dos Guararapes-PE, Brasil

² Centro de Estudos Avançados do Recife (CESAR) / CESAR School, Recife - PE, 50030-220, Brasil

DOI: 10.17013/risti.62.23-37

Resumo: A utilização de modelos de Aprendizagem de Máquina para apoiar sistemas de alerta precoce na educação superior tem crescido significativamente. No entanto, poucos estudos avaliam simultaneamente o desempenho preditivo e a equidade algorítmica desses modelos, particularmente sob diferentes estratégias de balanceamento de classes e otimização de hiperparâmetros. Este estudo investiga a aplicação de técnicas de Inteligência Artificial, com ênfase em Aprendizagem de Máquina, na predição da retenção estudantil utilizando o conjunto de dados Higher Education: Predictors of Student Retention, originalmente disponibilizado no UCI Machine Learning Repository (Realinho et al., 2021) e posteriormente difundido por meio da plataforma Kaggle. A metodologia integra técnicas de balanceamento (SMOTE e ADASYN), otimização de hiperparâmetros (GridSearch e Optuna) e a avaliação de seis algoritmos supervisionados. Para além das métricas tradicionais de desempenho, foram empregadas métricas de equidade algorítmica, nomeadamente Demographic Parity Difference e Equalized Odds Difference, permitindo analisar potenciais disparidades entre grupos. Os resultados demonstram que a combinação entre ADASYN e Optuna proporcionou o melhor desempenho preditivo, com o XGBoost apresentando os maiores F1-Scores nas três classes avaliadas. Verificou-se, contudo, que modelos com maior desempenho nem sempre apresentaram maior equidade algorítmica, evidenciando um compromisso entre essas duas dimensões. Os resultados contribuem para o desenvolvimento de sistemas de apoio à decisão mais éticos, transparentes e socialmente responsáveis para a educação superior.

Palavras-chave: aprendizagem de máquina; justiça algorítmica; equidade algorítmica; retenção estudantil; inteligência artificial responsável.

Predicting Student Retention in Higher Education using Machine Learning Models: An Algorithmic Fairness-Based Approach

Abstract: The use of Machine Learning to support early warning systems in higher education has increased significantly in recent years. However, few

studies simultaneously evaluate predictive performance and algorithmic fairness, particularly under different class balancing and hyperparameter optimization strategies. This study investigates the application of Artificial Intelligence (AI), with an emphasis on Machine Learning, to predict student retention using the Higher Education: Predictors of Student Retention dataset, originally published in the UCI Machine Learning Repository (Realinho et al., 2021) and subsequently disseminated through the Kaggle platform. The proposed methodology combines class balancing techniques (SMOTE and ADASYN), hyperparameter optimization (GridSearch and Optuna), and the evaluation of six supervised learning algorithms. In addition to conventional performance metrics, algorithmic fairness was assessed using Demographic Parity Difference and Equalized Odds Difference to identify potential disparities between protected groups. The results show that the combination of ADASYN and Optuna achieved the best predictive performance, with XGBoost obtaining the highest F1-scores across the three target classes. Nevertheless, models with superior predictive performance did not necessarily exhibit higher algorithmic fairness, highlighting the trade-off between these two dimensions. These findings contribute to the development of more ethical, transparent, and socially responsible decision-support systems for higher education.

Keywords: machine learning; algorithmic fairness; student retention; higher education; learning analytics.

1. Introdução

A crescente adoção de sistemas de Inteligência Artificial (IA) em processos de apoio à decisão tem transformado diversos setores, incluindo a educação superior, onde modelos de Aprendizagem de Máquina vêm sendo utilizados para identificar estudantes em risco de evasão, apoiar políticas institucionais e subsidiar sistemas de alerta precoce.

A crescente disponibilidade de dados e o avanço das técnicas computacionais têm ampliado significativamente o uso de modelos de Aprendizagem de Máquina em diversos setores, como saúde, finanças, segurança e educação. No contexto educacional, essas técnicas têm sido utilizadas para analisar fatores relacionados ao desempenho acadêmico, à permanência e ao abandono de estudantes, permitindo que instituições de ensino identifiquem padrões e desenvolvam estratégias para melhorar a retenção estudantil.

Apesar de seu grande potencial, a aplicação de sistemas baseados em inteligência artificial também levanta desafios importantes relacionados à equidade, à transparência e à responsabilização nas decisões automatizadas. Modelos preditivos treinados com dados históricos podem reproduzir ou até ampliar desigualdades existentes, caso possíveis vieses presentes nos dados não sejam adequadamente identificados e tratados durante o desenvolvimento dos algoritmos.

Nesse contexto, a justiça algorítmica constitui um conjunto de princípios voltados à redução de decisões discriminatórias produzidas por sistemas automatizados, sendo a equidade (fairness) uma de suas principais dimensões de avaliação. A análise da equidade em sistemas de Aprendizagem de Máquina torna-se, portanto, um aspecto fundamental para garantir que esses modelos produzam resultados mais justos e responsáveis.

No âmbito do ensino superior, modelos preditivos podem auxiliar instituições educacionais na identificação precoce de estudantes em risco de evasão, contribuindo para o desenvolvimento de políticas de apoio académico e estratégias institucionais voltadas à permanência estudantil. Entretanto, a utilização desses modelos exige atenção especial à possibilidade de geração de previsões enviesadas que possam afetar diferentes grupos de estudantes de maneira desigual.

Diante desse cenário, este estudo investiga a aplicação de modelos de Aprendizagem de Máquina na predição da retenção de estudantes no ensino superior, utilizando o conjunto de dados “*Higher Education: Predictors of Student Retention*”, originalmente disponibilizado no UCI Machine Learning Repository (Realinho et al., 2021) e posteriormente difundido pela plataforma Kaggle. A pesquisa busca analisar o desempenho desses modelos e avaliar aspectos relacionados à imparcialidade e à justiça algorítmica, contribuindo para o desenvolvimento de aplicações de inteligência artificial mais éticas e responsáveis no contexto educacional.

A avaliação de justiça algorítmica torna-se particularmente relevante em contextos educacionais, uma vez que decisões baseadas em modelos preditivos podem influenciar políticas institucionais e afetar diretamente oportunidades académicas de diferentes grupos de estudantes. Nesse sentido, a análise da equidade (*fairness*) em sistemas de inteligência artificial integra o debate contemporâneo sobre ética e responsabilidade no uso de tecnologias digitais na educação.

Embora a literatura reporte avanços significativos na aplicação de modelos de aprendizagem de máquina à predição da retenção estudantil, poucos estudos analisam simultaneamente o desempenho preditivo e a equidade algorítmica, bem como os impactos de diferentes estratégias de balanceamento de classes e otimização de hiperparâmetros sobre essas duas dimensões.

Assim, esta investigação procura responder à seguinte questão de investigação (RQ1):

RQ1: Como diferentes técnicas de balanceamento de classes e estratégias de otimização de hiperparâmetros influenciam simultaneamente o desempenho preditivo e a equidade algorítmica em modelos de predição da retenção estudantil?

Este estudo contribui para a literatura em três aspetos principais. Primeiro, compara diferentes algoritmos supervisionados combinados com técnicas de balanceamento de classes e otimização de hiperparâmetros para a predição da retenção estudantil. Segundo, integra métricas tradicionais de desempenho com métricas de equidade algorítmica, permitindo analisar simultaneamente precisão preditiva e potenciais vieses. Terceiro, fornece evidências empíricas que podem apoiar o desenvolvimento de sistemas de alerta precoce mais éticos, transparentes e socialmente responsáveis na educação superior.

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta a revisão da literatura. A Seção 3 descreve a metodologia adotada. A Seção 4 apresenta e discute os resultados experimentais. Finalmente, a Seção 5 apresenta as conclusões, limitações e trabalhos futuros.

2. Revisão de literatura

2.1. Inteligência Artificial e Justiça Algorítmica

A Inteligência Artificial (IA) tem apresentado avanços significativos nas últimas décadas, impulsionados principalmente pelo desenvolvimento de técnicas de Aprendizagem de Máquina (Machine Learning). Essas técnicas permitem a construção de modelos capazes de identificar padrões em grandes volumes de dados e realizar previsões ou classificações com alto nível de precisão, sendo amplamente utilizadas em diferentes áreas do conhecimento, como saúde, finanças, segurança e educação (Zhou et al., 2021).

A Aprendizagem de Máquina pode ser definida como um conjunto de métodos computacionais que permitem aos sistemas aprender a partir de dados, melhorando seu desempenho em determinadas tarefas sem a necessidade de programação explícita para cada situação. Esses métodos são geralmente classificados em aprendizagem supervisionada, não supervisionada e por reforço, dependendo da forma como os dados são utilizados no processo de treino dos modelos (Zhou et al., 2021).

Apesar dos avanços tecnológicos e das inúmeras aplicações práticas da IA, a crescente utilização de sistemas automatizados também tem levantado preocupações relacionadas à equidade e à transparência nas decisões tomadas por algoritmos. Nesse contexto, a discussão sobre justiça algorítmica tem ganhado destaque, especialmente em situações nas quais decisões baseadas em modelos computacionais podem impactar diretamente indivíduos ou grupos sociais (Raji et al., 2020).

A presença de vieses em sistemas de aprendizagem de máquina pode ocorrer por diferentes razões, incluindo conjuntos de dados de treino não representativos, presença de desigualdades históricas nos dados ou escolhas inadequadas no processo de modelagem (Baker & Hawn, 2022). Como resultado, algoritmos podem produzir previsões que favoreçam determinados grupos em detrimento de outros, reproduzindo ou até ampliando desigualdades existentes (Dwork et al., 2012).

Nos últimos anos, a discussão sobre IA responsável evoluiu significativamente, impulsionada por diretrizes internacionais que enfatizam transparência, responsabilização, equidade e supervisão humana no desenvolvimento de sistemas inteligentes (UNESCO, 2022; NIST, 2023). Esses princípios reforçam que modelos de aprendizagem de máquina devem ser avaliados não apenas quanto ao desempenho preditivo, mas também quanto à transparência, à equidade e aos seus potenciais impactos sociais.

Embora essas diretrizes forneçam princípios gerais para o desenvolvimento de sistemas de IA confiáveis, sua operacionalização depende da adoção de métricas quantitativas capazes de avaliar possíveis disparidades entre grupos protegidos. Nesse contexto, métricas de equidade tornaram-se componentes essenciais da avaliação de modelos preditivos utilizados em domínios sensíveis, como saúde, finanças e educação (UNESCO, 2022; NIST, 2023).

Estudos empíricos demonstram que sistemas baseados em inteligência artificial podem apresentar diferenças de desempenho entre grupos demográficos distintos. Um exemplo amplamente discutido na literatura refere-se a sistemas de reconhecimento facial que

apresentam taxas de erro significativamente maiores para determinados grupos raciais e de género, evidenciando limitações relacionadas à representatividade dos dados utilizados no treino dos modelos (Buolamwini & Gebru, 2018).

Diante desses desafios, diferentes abordagens têm sido propostas para identificar e mitigar vieses em sistemas de Aprendizagem de Máquina. Essas abordagens incluem técnicas de pré-processamento dos dados, modificações nos algoritmos de aprendizado e métodos de pós-processamento aplicados às previsões geradas pelos modelos (Kamishima et al., 2012; Bolukbasi et al., 2016).

Além disso, pesquisadores têm desenvolvido métricas específicas para avaliar a justiça algorítmica em sistemas automatizados. Essas métricas permitem medir possíveis disparidades nas previsões realizadas pelos modelos entre diferentes grupos, contribuindo para a identificação de situações em que decisões algorítmicas possam resultar em tratamento desigual (Barocas et al., 2023; Friedler et al., 2019).

Entre as diversas métricas propostas na literatura, *Demographic Parity* e *Equalized Odds* destacam-se por avaliarem dimensões complementares da equidade. Enquanto a primeira mede diferenças na distribuição das decisões entre grupos protegidos, a segunda considera simultaneamente taxas de verdadeiros positivos e falsos positivos, sendo amplamente empregadas na avaliação de modelos de classificação supervisionada (Barocas et al., 2023; Friedler et al., 2019).

Nesse sentido, a incorporação de critérios de justiça algorítmica no desenvolvimento de modelos de aprendizagem de máquina torna-se um aspecto essencial para garantir que essas tecnologias sejam utilizadas de forma ética e socialmente responsável. A análise sistemática de vieses e desigualdades nos resultados dos modelos constitui, portanto, um passo fundamental para o avanço de aplicações de inteligência artificial mais transparentes e confiáveis.

2.2. Aprendizagem de Máquina Aplicada à Retenção Estudantil

A retenção de estudantes no ensino superior constitui um dos principais desafios enfrentados por instituições educacionais em diferentes países. A evasão acadêmica pode gerar impactos significativos tanto para os estudantes quanto para as instituições, afetando o desempenho académico, os indicadores institucionais e a eficiência dos sistemas educacionais. Nesse contexto, a identificação precoce de fatores associados ao abandono ou à permanência dos estudantes tem se tornado um tema relevante de pesquisa.

Com o avanço das tecnologias de análise de dados, os modelos de Aprendizagem de Máquina têm sido cada vez mais utilizados para investigar padrões associados ao desempenho académico e ao comportamento dos estudantes. Esses métodos permitem analisar grandes volumes de dados educacionais e identificar relações complexas entre variáveis socioeconômicas, académicas e institucionais, possibilitando a construção de modelos preditivos capazes de estimar a probabilidade de evasão ou permanência de estudantes (Zhou et al., 2021).

Esses modelos normalmente utilizam algoritmos aplicados à retenção estudantil normalmente utilizam algoritmos de classificação supervisionada, nos quais o modelo é

treinado com dados históricos contendo informações sobre o perfil dos estudantes e seus resultados acadêmicos. A partir desse treino, o algoritmo aprende padrões presentes nos dados e passa a ser capaz de prever categorias relacionadas ao status acadêmico do estudante, como permanência, evasão ou conclusão do curso (Zhou et al., 2021).

Diversos algoritmos de Aprendizagem de Máquina têm sido utilizados nesse tipo de aplicação, incluindo Árvores de Decisão, Random Forest, Support Vector Machines, Gradient Boosting e métodos baseados em redes neurais. A escolha do algoritmo mais adequado depende de fatores como o tamanho do conjunto de dados, a complexidade das variáveis analisadas e os objetivos específicos do estudo. Além disso, técnicas de otimização de hiperparâmetros e métodos de balanceamento de dados podem ser utilizados para melhorar o desempenho preditivo dos modelos (Goodfellow, Bengio & Courville, 2016).

Estudos recentes reforçam o potencial da Aprendizagem de Máquina para a predição da evasão estudantil. Nesse contexto, Aco Tito, Hanco Condori e Pérez Vera (2023) realizaram uma análise comparativa de diferentes algoritmos de classificação, incluindo Regressão Logística, Support Vector Machine, Árvores de Decisão, Random Forest e Redes Neurais, concluindo que a Regressão Logística apresentou o melhor desempenho para o conjunto de dados analisado. Embora esses resultados reforcem a viabilidade da utilização de modelos preditivos no contexto educacional, a avaliação concentrou-se exclusivamente em métricas tradicionais de desempenho, não contemplando aspectos relacionados à justiça algorítmica e à equidade.

Estudos recentes indicam que modelos baseados em aprendizagem de máquina têm alcançado elevado desempenho na predição da retenção estudantil; entretanto, a maioria concentra-se na maximização das métricas tradicionais de desempenho, dedicando menor atenção à avaliação sistemática da equidade algorítmica entre diferentes grupos de estudantes (Vaarma & Li, 2024). Essa lacuna evidencia a necessidade de abordagens que conciliem desempenho preditivo e justiça algorítmica.

Nesse contexto, a utilização de modelos preditivos no contexto educacional também levanta preocupações relacionadas à equidade das previsões geradas pelos algoritmos. Caso os dados utilizados no treino apresentem desequilíbrios ou representem inadequadamente determinados grupos de estudantes, os modelos podem gerar previsões enviesadas que reproduzam desigualdades existentes no sistema educacional. Dessa forma, torna-se necessário avaliar não apenas o desempenho preditivo dos modelos, mas também possíveis disparidades nos resultados produzidos para diferentes grupos de estudantes (Friedler et al., 2019).

Nesse sentido, a incorporação de métricas de justiça algorítmica no processo de avaliação dos modelos pode contribuir para identificar e reduzir possíveis vieses presentes nas previsões. Métricas como Demographic Parity e Equalized Odds, por exemplo, permitem analisar diferenças nas taxas de classificação entre grupos distintos, auxiliando na construção de sistemas de decisão mais equitativos (Barocas et al., 2023).

Dessa forma, a aplicação de algoritmos de Aprendizagem de Máquina na análise da retenção estudantil apresenta elevado potencial para apoiar instituições de ensino na formulação de políticas e estratégias de intervenção. Contudo, a literatura ainda dedica maior atenção à maximização do desempenho preditivo do que à avaliação sistemática

da justiça algorítmica e da equidade entre diferentes grupos de estudantes. Essa lacuna evidencia a necessidade de abordagens que integrem simultaneamente desempenho preditivo e justiça algorítmica, motivando a proposta metodológica apresentada neste trabalho. Assim, a integração entre desempenho preditivo e justiça algorítmica representa uma tendência emergente na aplicação de Aprendizagem de Máquina à educação, exigindo metodologias capazes de equilibrar precisão, transparência e equidade.

3. Metodologia

Este estudo adota uma abordagem baseada em Aprendizagem de Máquina para avaliar a predição da retenção de estudantes no ensino superior, com ênfase na mitigação de vieses e na promoção de justiça algorítmica. A metodologia integra etapas de pré-processamento, balanceamento de dados, construção e otimização de modelos, além da avaliação do desempenho preditivo e da equidade algorítmica. A seguir, detalham-se os procedimentos adotados.

O fluxo experimental adotado compreendeu as seguintes etapas: (i) pré-processamento e divisão estratificada em treino (70%) e teste (30%); (ii) normalização das variáveis numéricas utilizando *StandardScaler*; (iii) aplicação do SMOTE/ADASYN no conjunto de treino; (iv) otimização dos hiperparâmetros; (v) treinamento dos modelos; e (vi) avaliação por meio de métricas de desempenho e de justiça algorítmica. Todos os experimentos foram conduzidos utilizando *random_state = 42*, de modo a assegurar a reprodutibilidade dos experimentos.

3.1. Base de Dados

O conjunto de dados utilizado foi o *Higher Education: Predictors of Student Retention*, originalmente disponibilizado no UCI Machine Learning Repository (Realinho et al., 2021) e posteriormente difundido na plataforma Kaggle¹. Este conjunto de dados contém informações sobre fatores acadêmicos, socioeconômicos e demográficos de estudantes de ensino superior, incluindo dados sobre matrícula, conclusão, evasão e desempenho acadêmico. A escolha deste conjunto de dados justifica-se por sua amplitude, qualidade e relevância para estudos de predição de retenção estudantil, além de permitir análises sobre desigualdades e possíveis vieses em predições.

3.2. Pré-processamento de Dados

Devido à natureza desbalanceada do dataset, em que determinadas classes (por exemplo, estudantes que evadem ou concluem o curso) possuem menor representação, aplicaram-se técnicas de *oversampling* para balanceamento das classes, mitigando o impacto do desbalanceamento no treino dos modelos:

- SMOTE (*Synthetic Minority Over-sampling Technique*): gera exemplos sintéticos de classes minoritárias interpolando entre instâncias vizinhas, aumentando a representatividade das classes menos frequentes (Chawla et al., 2002).

¹ <https://www.kaggle.com/datasets/lucamorn/higher-education-predictors-of-student-retention>

- **ADASYN (Adaptive Synthetic Sampling):** cria exemplos sintéticos de forma adaptativa, focando em instâncias difíceis de classificar, o que favorece a aprendizagem de padrões complexos em classes minoritárias (He et al., 2008).

Essas técnicas foram aplicadas para reduzir os efeitos do desequilíbrio entre classes durante o treino dos modelos, favorecendo uma aprendizagem mais equilibrada. Ressalta-se, contudo, que métodos de oversampling não garantem, por si só, maior equidade entre grupos sensíveis, motivo pelo qual a avaliação de justiça algorítmica foi realizada separadamente por meio de métricas específicas.

3.3. Modelos de Aprendizagem de Máquina

Foram avaliados seis algoritmos supervisionados amplamente utilizados em problemas de classificação: *Decision Tree*, *Random Forest*, *Support Vector Machine*, *Gradient Boosting*, *XGBoost* e *CatBoost*. Esses algoritmos foram selecionados por apresentarem diferentes níveis de complexidade, interpretabilidade e capacidade de generalização.

O *Decision Tree* é um modelo simples e interpretável, permitindo visualizar o processo de decisão. O *Random Forest* consiste num conjunto de árvores de decisão, robusto ao *overfitting* e adequado para variáveis categóricas e contínuas. O *Support Vector Machine* (SVM) é particularmente indicado para classificação em espaços multidimensionais. O *Gradient Boosting* constrói modelos sequenciais que minimizam iterativamente a função de perda. O *XGBoost* constitui uma versão otimizada do *Gradient Boosting*, frequentemente utilizada em problemas de classificação devido ao seu elevado desempenho. Por fim, o *CatBoost* destaca-se pelo tratamento eficiente de variáveis categóricas e pela redução do risco de *overfitting*.

Os modelos foram inicialmente avaliados sem otimização e posteriormente ajustados por estratégias de otimização de hiperparâmetros, incluindo *GridSearch* e *Optuna* (Akiba et al., 2019). Os resultados apresentados correspondem às configurações que obtiveram melhor desempenho para cada modelo.

3.4. Divisão de Dados e Normalização

Os dados foram divididos em 70% para treino e 30% para teste, utilizando a função *train_test_split* do *Scikit-learn*. Em seguida, aplicou-se normalização das variáveis numéricas com *StandardScaler*, garantindo que todas as features tivessem escala uniforme, evitando que variáveis com magnitudes maiores dominassem o aprendizado dos modelos (Goodfellow, Bengio, & Courville, 2016).

3.5. Otimização de Hiperparâmetros

A otimização dos modelos foi conduzida por meio de duas estratégias distintas.

- **GridSearch:** busca exaustiva em uma grade de combinações predefinidas de hiperparâmetros, garantindo cobertura completa do espaço de parâmetros selecionado.
- **Optuna:** abordagem de otimização bayesiana que ajusta automaticamente os hiperparâmetros visando maximizar o F1-Score. Cada estudo contou com 20 iterações, permitindo explorar de forma eficiente o espaço de parâmetros (Akiba et al., 2019).

A escolha de ambas as estratégias permite comparar abordagens tradicionais e modernas, avaliando *trade-offs* entre desempenho, custo computacional e eficiência. Foram otimizados hiperparâmetros específicos de cada algoritmo, incluindo profundidade máxima (*max_depth*), número mínimo de amostras por folha (*min_samples_leaf*), critério de divisão (*criterion*), taxa de aprendizagem (*learning_rate*), parâmetros de regularização, número de iterações e parâmetros específicos de cada modelo, conforme aplicável. Os hiperparâmetros foram avaliados por validação cruzada estratificada de 10 dobras utilizando F1-Score ponderado como função objetivo. Na avaliação final, foram reportados os F1-Scores por classe, além das demais métricas de desempenho e de equidade.

3.6. Métricas de Avaliação

Para avaliar o desempenho dos modelos, utilizou-se o F1-Score por classe evasão (*Dropout*), inscrito (*Enrolled*) e diplomado (*Graduate*). Esta métrica pondera precisão e sensibilidade (*recall*), sendo adequada para datasets desbalanceados, pois penaliza tanto falsos positivos quanto falsos negativos. Além disso, foram analisadas matrizes de confusão e métricas complementares de acurácia e sensibilidade (*recall*) para validação do comportamento do modelo em cada classe.

3.7. Avaliação de Fairness

Para garantir a imparcialidade das previsões, os modelos foram avaliados em relação à variável sensível de género. As métricas de justiça algorítmica utilizadas foram:

- *Demographic Parity Difference* (Diferença de Paridade Demográfica): avalia a diferença na probabilidade de previsão positiva entre grupos, indicando se todos os grupos têm chances similares de receber o mesmo resultado.
- *Equalized Odds Difference* (Diferença de Oportunidades Equalizadas): mede diferenças na taxa de verdadeiros positivos e falsos positivos entre grupos, avaliando se o modelo mantém equidade no desempenho entre subpopulações.

As métricas foram calculadas utilizando a biblioteca Fairlearn (Bird et al., 2020). A análise da equidade algorítmica visa garantir que os modelos não reproduzam ou ampliem desigualdades, alinhando o estudo às diretrizes de ética e responsabilidade em inteligência artificial. A escolha das métricas *Demographic Parity Difference* e *Equalized Odds Difference* fundamenta-se em sua ampla utilização na literatura para avaliar, respectivamente, a igualdade de distribuição das decisões e a igualdade das taxas de erro entre grupos protegidos.

A escolha da variável género deve-se à sua disponibilidade estruturada no conjunto de dados e à sua ampla utilização como atributo protegido em estudos de justiça algorítmica. Contudo, essa escolha não impede a utilização da mesma metodologia com múltiplas variáveis sensíveis, como idade, estatuto socioeconómico ou nacionalidade, desde que tais atributos estejam disponíveis.

4. Resultados e Discussão

Nesta seção, apresentam-se os resultados obtidos a partir da aplicação dos modelos de aprendizagem de máquina para predição da retenção de estudantes no ensino superior,

utilizando as técnicas de balanceamento SMOTE e ADASYN, além de otimização de hiperparâmetros via GridSearch e Optuna. Os modelos foram avaliados com base nos F1-Scores para três classes: Evasão (*Dropout*), Inscrito (*Enrolled*) e Diplomado (*Graduate*). Além disso, foram também analisadas métricas de equidade algorítmica (*Demographic Parity Difference* e *Equalized Odds Difference*) considerando a variável sensível de gênero.

4.1. Desempenho dos Modelos (F1-Score)

A Tabela 1 resume os F1-Scores obtidos por cada modelo, técnica de balanceamento e método de otimização.

Modelo	Balanceamento	Otimização	Dropout	Enrolled	Graduate
Decision Tree	SMOTE	GridSearch	0,7482	0,7310	0,6521
Decision Tree	ADASYN	Optuna	0,7732	0,7587	0,6747
Random Forest	SMOTE	Optuna	0,8321	0,8475	0,7890
Random Forest	ADASYN	Optuna	0,8453	0,8566	0,7963
SVM	SMOTE	Optuna	0,8110	0,7902	0,7483
SVM	ADASYN	Optuna	0,8172	0,7983	0,7568
Gradient Boosting	ADASYN	Optuna	0,8504	0,8461	0,8098
XGBoost	ADASYN	Optuna	0,8588	0,8600	0,8158
CatBoost	ADASYN	Optuna	0,8452	0,8517	0,8003

Tabela 1 – F1-Scores dos Modelos por Classe e Técnica de Balanceamento

Os resultados apresentados na Tabela 1 demonstram que a combinação entre o balanceamento ADASYN e a otimização por Optuna produziu, de forma consistente, os melhores desempenhos preditivos entre os modelos avaliados. O XGBoost destacou-se por alcançar os maiores F1-Scores nas três classes analisadas, evidenciando elevada capacidade de generalização. Gradient Boosting e CatBoost apresentaram desempenho semelhante, enquanto Decision Tree obteve os menores resultados, sobretudo na classe Diplomado (*Graduate*), indicando maior dificuldade na modelação dessa categoria.

4.2. Avaliação de Fairness

Para avaliar a equidade dos modelos, utilizou-se a métrica Demographic Parity Difference e Equalized Odds Difference, considerando o gênero como variável sensível. Valores próximos de zero indicam maior equidade entre os grupos.

Modelo	Balanceamento	Otimização	Dropout	Enrolled	Graduate
Decision Tree	ADASYN	Optuna	0,0953	0,1421	0,1604
Random Forest	ADASYN	Optuna	0,1021	0,1845	0,2032
SVM	ADASYN	Optuna	0,0905	0,1234	0,1456
Gradient Boosting	ADASYN	Optuna	0,0846	0,1567	0,1921

Modelo	Balanceamento	Otimização	Dropout	Enrolled	Graduate
<i>XGBoost</i>	ADASYN	Optuna	0,0885	0,1489	0,1790
<i>CatBoost</i>	ADASYN	Optuna	0,0890	0,2162	0,1983

Tabela 2 – Demographic Parity Difference por Modelo

Modelo	Balanceamento	Otimização	Dropout	Enrolled	Graduate
<i>Decision Tree</i>	ADASYN	Optuna	0,0621	0,1014	0,1372
<i>Random Forest</i>	ADASYN	Optuna	0,0897	0,1243	0,1457
<i>SVM</i>	ADASYN	Optuna	0,0464	0,0723	0,0890
<i>Gradient Boosting</i>	ADASYN	Optuna	0,0875	0,1083	0,1427
<i>XGBoost</i>	ADASYN	Optuna	0,0532	0,0941	0,1102
<i>CatBoost</i>	ADASYN	Optuna	0,0591	0,1267	0,1586

Tabela 3 – Equalized Odds Difference por Modelo

Observa-se que o SVM com ADASYN apresentou os melhores resultados em termos de equidade algorítmica, com menores diferenças de paridade demográfica e oportunidades equalizadas, especialmente na classe Dropout. Em contrapartida, Random Forest e Gradient Boosting apresentaram maiores disparidades, principalmente para a classe Diplomado, indicando potenciais vieses nos modelos. Embora as diferenças observadas sejam relativamente moderadas, verifica-se que os modelos apresentam comportamentos distintos quanto à distribuição das decisões entre os grupos analisados.

4.3. Discussão

A utilização de modelos preditivos para identificar estudantes em risco de evasão pode apoiar estratégias institucionais de permanência. No entanto, a adoção dessas tecnologias exige cuidados éticos, pois classificações automatizadas podem influenciar decisões académicas e administrativas que impactam diretamente trajetórias educacionais individuais.

A análise dos resultados obtidos evidencia diferenças importantes tanto no desempenho quanto na equidade dos modelos de Aprendizagem de Máquina aplicados à predição da retenção estudantil. De maneira geral, os algoritmos mais complexos, como XGBoost e CatBoost, apresentaram os maiores F1-Scores, especialmente quando combinados com o balanceamento ADASYN e otimização via Optuna. Esses modelos demonstraram elevada capacidade de capturar padrões complexos nos dados, garantindo maior acurácia na previsão de estudantes que abandonam (*Dropout*), permanecem matriculados (*Enrolled*) ou concluem o curso (*Graduate*).

Os resultados obtidos permitem responder à questão de investigação proposta. Observou-se que tanto a técnica de balanceamento quanto o método de otimização influenciam simultaneamente o desempenho preditivo e a equidade algorítmica. Enquanto o ADASYN associado ao Optuna proporcionou os melhores F1-Scores, a melhoria do

desempenho não implicou necessariamente maior equidade entre grupos protegidos, evidenciando a existência de um compromisso (trade-off) entre essas duas dimensões.

Em contrapartida, modelos mais simples, como Decision Tree e SVM, embora apresentem F1-Scores ligeiramente inferiores, mostraram comportamento mais equilibrado em termos de equidade algorítmica. No caso do SVM otimizado com ADASYN, as métricas de Demographic Parity Difference e Equalized Odds Difference foram consistentemente menores, indicando previsões menos enviesadas entre grupos de gênero. Os resultados sugerem que modelos mais complexos podem apresentar maior propensão a disparidades entre grupos protegidos, embora essa relação dependa das características dos dados e do algoritmo utilizado.

A escolha das técnicas de balanceamento mostrou-se determinante para o desempenho e a equidade dos modelos. O ADASYN, ao gerar amostras sintéticas adaptativas para classes minoritárias, aprimorou a capacidade dos modelos de identificar padrões associados às classes minoritárias. Já o SMOTE, embora eficaz, apresentou desempenho ligeiramente inferior, possivelmente devido à criação de instâncias sintéticas menos representativas da complexidade do conjunto de dados.

A otimização de hiperparâmetros via Optuna revelou-se superior ao GridSearch, especialmente em modelos com grande espaço de parâmetros, como XGBoost e Gradient Boosting. A abordagem bayesiana permitiu explorar combinações de hiperparâmetros mais eficazes, maximizando o F1-Score e garantindo ajustes finos para melhorar a performance preditiva.

As análises de equidade algorítmica revelam que mesmo modelos com alto desempenho preditivo podem apresentar disparidades relevantes. Por exemplo, o XGBoost otimizado com ADASYN, apesar de alcançar os melhores F1-Scores em todas as classes, apresentou diferenças de paridade demográfica mais elevadas na classe Inscrito, sugerindo que a previsão de estudantes matriculados pode favorecer ou prejudicar certos grupos de gênero. Esse resultado reforça a necessidade de avaliar métricas de equidade paralelamente às métricas tradicionais de desempenho, sobretudo em contextos educacionais, onde decisões automatizadas impactam diretamente trajetórias individuais.

Do ponto de vista prático, os resultados fornecem diretrizes valiosas para instituições de ensino superior. Modelos como XGBoost e CatBoost, combinados com técnicas de sobreamostragem (*oversampling*) e otimização avançada, permitem identificar estudantes em risco de evasão com elevada precisão. Contudo, políticas de retenção devem considerar não apenas o desempenho preditivo, mas também a equidade das previsões. Modelos como SVM-ADASYN, apesar de apresentarem desempenho ligeiramente inferior, podem funcionar como ferramentas complementares, auxiliando no monitorização ética e imparcial das decisões institucionais.

Em conjunto, os resultados demonstram que a seleção de modelos para sistemas de alerta precoce não deve basear-se exclusivamente em métricas tradicionais de desempenho. A incorporação de métricas de equidade algorítmica permite identificar soluções que conciliam elevada capacidade preditiva com menor risco de introduzir disparidades entre grupos de estudantes, contribuindo para o desenvolvimento de sistemas de apoio à decisão mais responsáveis.

Os resultados obtidos podem auxiliar instituições de ensino superior no desenvolvimento de sistemas de alerta precoce para identificação de estudantes em risco de evasão. Contudo, a implementação dessas ferramentas deve ser acompanhada por mecanismos contínuos de monitorização da equidade algorítmica, garantindo que decisões institucionais apoiadas por modelos preditivos sejam mais transparentes, responsáveis e socialmente justas, sem reforçar desigualdades preexistentes.

Por fim, algumas limitações do estudo devem ser destacadas. A análise de equidade algorítmica considerou apenas a variável sensível de género, sem explorar outros fatores potenciais de desigualdade, como etnia, rendimento socioeconómico ou idade. Além disso, o estudo utilizou um único conjunto de dados, o que pode limitar a generalização dos resultados para outras instituições ou contextos educacionais. Estudos futuros poderão expandir a análise para múltiplas variáveis sensíveis, validar os modelos em diferentes bases de dados e incorporar técnicas de explicabilidade, como SHAP e LIME, contribuindo para aumentar a robustez, a transparência e a aplicabilidade das recomendações propostas.

5. Conclusões, Limitações e Trabalhos Futuros

Este estudo investigou como diferentes estratégias de balanceamento de classes e otimização de hiperparâmetros influenciam simultaneamente o desempenho preditivo e a equidade algorítmica em modelos de predição da retenção estudantil.

A aplicação de técnicas avançadas de sobreamostragem (*oversampling*), como o ADASYN, combinadas à otimização bayesiana via Optuna, demonstrou potencial significativo para melhorar o desempenho preditivo de modelos de aprendizagem de máquina na predição da retenção estudantil. Entre os algoritmos avaliados, o XGBoost otimizado apresentou os melhores F1-Scores para as três categorias analisadas, Evasão (*Dropout*), Inscritos (*Enrolled*) e Diplomado (*Graduate*), destacando-se como a alternativa mais robusta para prever padrões complexos de permanência académica.

Além do desempenho preditivo, a análise de equidade algorítmica revelou-se crucial. Modelos como SVM, quando combinados com ADASYN, apresentaram menores valores de *Demographic Parity Difference* e *Equalized Odds Difference*, especialmente para a classe Evasão (*Dropout*), indicando previsões mais equitativas entre diferentes grupos de género. Em contraste, Random Forest e Gradient Boosting mostraram disparidades mais elevadas, sobretudo para a classe Diplomado, evidenciando a necessidade de considerar métricas de equidade em paralelo às métricas tradicionais de desempenho.

Esses resultados reforçam a importância de integrar desempenho preditivo e equidade algorítmica na avaliação de modelos aplicados à educação. A utilização conjunta de técnicas de sobreamostragem (*oversampling*) e otimização avançada pode aumentar o desempenho dos modelos, desde que acompanhada por avaliações sistemáticas de equidade, contribuindo para o desenvolvimento de sistemas de apoio à decisão mais responsáveis.

Quanto às limitações, a análise de equidade algorítmica concentrou-se exclusivamente na variável sensível de género, não explorando outros fatores potenciais de desigualdade, como etnia, renda ou idade. Além disso, o estudo utilizou apenas um conjunto de dados,

o que pode restringir a generalização dos resultados para diferentes instituições ou contextos educacionais.

Para trabalhos futuros, recomenda-se expandir a investigação considerando múltiplas variáveis sensíveis e validar os modelos em bases de dados de outras instituições e países. Além disso, recomenda-se avaliar a metodologia em diferentes contextos institucionais, de forma a analisar sua capacidade de generalização para distintos sistemas de ensino superior. Também é pertinente explorar técnicas adicionais de mitigação de viés e métodos de interpretabilidade de modelos, visando oferecer recomendações mais detalhadas para políticas educacionais inclusivas.

Embora modelos baseados em boosting, como o XGBoost, apresentem elevado desempenho preditivo, sua interpretabilidade pode ser limitada em comparação a modelos mais simples. Nesse sentido, técnicas de explicabilidade, como SHAP ou análise de importância de variáveis, podem ser utilizadas em estudos futuros para tornar as decisões algorítmicas mais transparentes.

Referências

- Aco Tito, A. E., Hanco Condori, B. O., & Pérez Vera, Y. (2023). Análisis comparativo de técnicas de machine learning para la predicción de casos de deserción universitaria. *RISTI – Revista Ibérica de Sistemas e Tecnologías de Informação*, (51), 84–98. <https://doi.org/10.17013/risti.51.84-98>
- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (pp. 2623–2631). <https://doi.org/10.1145/3292500.3330701>
- Andrade-Girón, D., Sandivar-Rosas, J., Marín-Rodríguez, W., Susanibar-Ramirez, E., Toro-Dextre, E., Ausejo-Sanchez, J., & Angeles-Morales, J. (2023). Predicting student dropout based on machine learning and deep learning: A systematic review. *EAI Endorsed Transactions on Scalable Information Systems*, 10(5), 3586. <https://doi.org/10.4108/eetsis.3586>
- Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32(4), 1052–1092. <https://doi.org/10.1007/s40593-021-00285-9>
- Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press. <https://fairmlbook.org>
- Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft Research*. <https://fairlearn.org>
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. <https://dl.acm.org/doi/10.5555/3157382.3157584>

- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, pp. 77–91). <https://proceedings.mlr.press/v81/buolamwini18a.html>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). <https://doi.org/10.1145/2090236.2090255>
- Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 329–338). <https://doi.org/10.1145/3287560.3287589>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. <https://www.deeplearningbook.org/>
- He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the IEEE International Joint Conference on Neural Networks* (pp. 1322–1328). <https://doi.org/10.1109/IJCNN.2008.4633969>
- Kamishima, T., Akaho, S., & Asoh, H. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 35–50). https://doi.org/10.1007/978-3-642-33486-3_3
- NIST. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). <https://doi.org/10.1145/3351095.3372873>
- Realinho, V., Vieira Martins, M., Machado, J., & Baptista, L. (2021). Predict Students' Dropout and Academic Success (Dataset). UCI Machine Learning Repository. <https://doi.org/10.24432/C5MC89>
- UNESCO. (2022). *Recommendation on the ethics of artificial intelligence*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>
- Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning. *Technology in Society*, 76, 102474. <https://doi.org/10.1016/j.techsoc.2024.102474>
- Zhou, Z.-H. (2021). *Machine learning*. Springer. <https://doi.org/10.1007/978-981-15-1967-3>

Modelos de Machine Learning para la Predicción de la Deserción en una Institución de Educación Superior en Colombia

Ludwing Bernart Carreño Patiño¹

lbcarreño@fucsulud.edu.co

¹ Facultad de Ciencias Administrativas en Salud, Fundación Universitaria de Ciencias de la Salud – FUCS, Bogotá D.C., Colombia.

DOI: 10.17013/risti.62.38–46

Resumen: La deserción estudiantil en la educación superior es un problema multifactorial con impacto directo en la competitividad nacional y en los ingresos de las instituciones educativas. Este trabajo desarrolló modelos predictivos de deserción mediante herramientas de Machine Learning aplicadas a registros académicos y socioeconómicos de una Facultad de Ciencias Administrativas en Salud en Bogotá D.C., Colombia, siguiendo la metodología CRISP-DM. Se construyeron tres bases de datos ($n = 70-522$ registros) y se aplicaron técnicas de balanceo de clases (SMOTE y RUS). Los modelos Random Forest y Extra Trees obtuvieron el mejor desempeño ($AUC \geq 0.99$, $F1 \geq 0.97$). Las variables de mayor relevancia predictiva fueron la pertenencia al programa distrital Jóvenes a la U, el rendimiento en la Cátedra FUCS y el promedio acumulado.

Palabras-clave: abandono escolar; aprendizaje automático; CRISP-DM; educación superior; minería de datos educativos.

Machine Learning Models for Dropout Prediction in a Higher Education Institution in Colombia

Abstract: Student dropout in higher education is a multifactorial problem with a direct impact on national competitiveness and institutional revenues. This study developed dropout prediction models using Machine Learning tools applied to academic and socioeconomic records of a Health Administrative Sciences faculty in Bogotá D.C., Colombia, following the CRISP-DM methodology. Three databases were built ($n = 70-522$ records) and class imbalance was handled using SMOTE and RUS techniques. Random Forest and Extra Trees models showed the best performance ($AUC \geq 0.99$, $F1 \geq 0.97$). The most relevant predictive variables were membership in the Jóvenes a la U district program, first-semester performance in the FUCS introductory course, and cumulative GPA.

Keywords: CRISP-DM; data mining; educational data mining; higher education; machine learning; student dropout.

1. Introducción

La deserción puede definirse como el abandono de la formación académica por parte del estudiante, producto de variables socioeconómicas, individuales, institucionales y académicas (Ministerio de Educación Nacional, 2009). Este fenómeno puede comprenderse también desde perspectivas ecológicas —micro, meso y macrosistemas— que enmarcan las interacciones del estudiante con su entorno próximo, su institución y el contexto económico y político (Castro-Martínez & Machuca-Téllez, 2023).

Los algoritmos de Machine Learning (ML) emergen como herramientas prometedoras para predecir la deserción y encuadrar estrategias de prevención, dando lugar al campo del Educational Data Mining (EDM) (Kabathova & Drlik, 2021; Niyogisubizo et al., 2022). Investigaciones previas en el contexto latinoamericano han explorado esta vía con resultados alentadores. Castiblanco Vargas et al. (2021), Fandiño Benavides & Camargo (2022) y Rodríguez Vásconez & Benalcázar Palacios (2023) lograron clasificar estudiantes en riesgo con hasta el 87% de exactitud mediante regresión logística y redes neuronales. Pérez-Gutiérrez (2020) demostró que algoritmos simples alcanzan niveles de precisión adecuados. Guerra et al. (2020) reportaron un 92% de exactitud con redes neuronales sobre datos universitarios. Aco Tito et al. (2023) y Fandiño Benavides & Camargo (2022) señalaron que, según el contexto y la composición de los datos, la regresión logística o el Random Forest pueden resultar superiores.

El presente trabajo desarrolló modelos predictivos de deserción aplicados a registros de una Facultad de Ciencias Administrativas en Salud en Bogotá D.C., con la contribución específica de emplear CRISP-DM de forma sistemática, combinar técnicas de balanceo de clases y validar los resultados mediante AutoML y validación cruzada en un contexto institucional del sector salud escasamente documentado en la literatura.

2. Metodología

La metodología empleada fue CRISP-DM (Cross-Industry Standard Process for Data Mining), que comprende seis etapas: comprensión del negocio, comprensión de los datos, preparación, modelado, evaluación e implementación (Raschka & Mirjalili, 2019). La fase de implementación no se desarrolla completamente en este trabajo.

2.1. Fuentes de datos y bases de datos construidas

Se construyeron tres bases de datos a partir de registros académicos y socioeconómicos de la Facultad: (a) *df_22* con 522 registros tras eliminación de valores nulos; (b) *df_24* con 215 registros que incorporaron lugar de nacimiento y situación laboral; y (c) *df_26* con 70 registros enriquecidos con puntajes Saber 11 consultados en la plataforma del ICFES. Se incluyeron estudiantes activos e inactivos con inicio de vigencia desde 2019; se excluyeron estudiantes de otras facultades y aquellos con cierre de materias previo a 2019. El procesamiento y análisis de datos se realizó en el entorno de Google Colaboratory, utilizando el lenguaje de programación Python (versión 3.10). Los datos iniciales se estructuraron en una hoja de cálculo en Microsoft Excel. Las bases de datos definitivas fueron alojadas y gestionadas en la nube. Tabla 1 presenta las principales variables del estudio.

Variable	Descripción	Tipo de dato
PROMEDIO ACUMULADO	Promedio histórico de todos los semestres cursados	Númerica continua
PROMEDIO ÚLTIMO SEM.	Promedio del último semestre registrado	Númerica continua
NUM_SEMESTRES	Número de semestres cursados	Númerica discreta
NUM_ASIGNATURAS	Número de asignaturas cursadas	Númerica discreta
BAJO RENDIMIENTO	Situación académica de bajo rendimiento activa	Binaria (0/1)
JÓVENES A LA U	Pertenencia al programa distrital Jóvenes a la U	Binaria (0/1)
BECA / DESCUENTO	Beneficio económico activo en matrícula	Binaria (0/1)
ESTRATO	Estrato socioeconómico del estudiante	Ordinal (1-6)
CONTINUIDAD	Avance dentro del ciclo propedéutico	Binaria (0/1)
NOTAS 1er SEM.	Calificaciones de asignaturas del primer semestre	Númerica continua
PUNTAJE SABER 11	Puntaje en las pruebas de Estado (ICFES)	Númerica continua
DESERCIÓN (objetivo)	Desvinculación académica ≥ 2 periodos sin graduarse	Binaria (0/1)

Tabla 1 – Variables principales de la investigación

2.2. Balanceo de clases

Dado el habitual desequilibrio de clases en fenómenos sociales, se aplicaron dos técnicas de balanceo: SMOTE (Synthetic Minority Over-sampling Technique) y RUS (Random Undersampling). La Tabla 2 muestra los registros resultantes. RUS solo se aplicó a df_22 por limitaciones del algoritmo con clases minoritarias muy pequeñas.

Base de datos	Técnica	Registros tras balanceo
Grande (df_22)	SMOTE	566
Grande (df_22)	RUS	146
Mediana (df_24)	SMOTE	340
Pequeña (df_26)	SMOTE	96

Tabla 2 – Técnicas de balanceo de clases aplicadas

2.3. AutoML y selección de modelos

Se implementó AutoML mediante la librería pycaret (preprocess = True, partición 80/20 entrenamiento/prueba) para evaluar múltiples algoritmos simultáneamente. Los modelos seleccionados se sometieron a validación cruzada k-fold, reportando AUC, F1-score, Precision y Recall como métricas principales. Para los modelos de Regresión Logística se aplicó análisis SHAP para interpretabilidad de las predicciones.

3. Resultados

3.1. Análisis exploratorio: variables categóricas

La Figura 1 muestra el comportamiento de la deserción frente a las variables categóricas más relevantes. La muestra presenta predominio femenino (81.4%), aunque el fenómeno de deserción es equivalente entre sexos. El estrato socioeconómico 2 concentra la mayor proporción de desertores, consistente con el rol protector del soporte financiero directo (Villegas & Núñez Lira, 2024, citando a Behr et al., 2020). El programa Jóvenes a la U — al que pertenece el 61.9% de los estudiantes— actúa como factor protector: la deserción es predominante en quienes no participan en él. El bajo rendimiento académico muestra la asociación más pronunciada con la deserción, con proporciones cercanas al 90%.

3.2. Análisis exploratorio: variables numéricas

El análisis cuantitativo reveló que la mayor proporción de desertores se concentra en promedios acumulados entre 1 y 3, y que un promedio de fallas por semestre superior a 20 es un predictor relevante. Las calificaciones en asignaturas del primer semestre —en particular la Cátedra FUCS— muestran que estudiantes con calificaciones inferiores a 2.0 presentan mayor propensión a desertar. Los puntajes Saber 11 entre 200 y 250 puntos se asociaron a mayores tasas de deserción, consistente con lo reportado en (Gartner et al., 2016; Rose-Parra et al., 2023).

3.3. Selección con AutoML

La Tabla 3 presenta los mejores modelos identificados por AutoML sobre los datos balanceados. Los modelos de árboles de decisión ensemble —Random Forest y Extra Trees— obtuvieron los mejores resultados sobre df_22 con SMOTE (Accuracy = 0.9543; AUC > 0.99). Las bases más pequeñas mostraron métricas aparentemente superiores, pero son susceptibles de sobreajuste dado su reducido número de registros.

Base	Modelo	Accuracy	AUC	F1	Kappa
df_22_SMOTE	Random Forest	0.9543	0.9923	0.9548	0.9085
df_22_SMOTE	Extra Trees	0.9543	0.9931	0.9543	0.9085
df_22_RUS	Regresión Logística	0.9018	0.9513	0.9089	0.8045
df_22_RUS	Extra Trees	0.8836	0.9543	0.8892	0.7679
df_24_SMOTE	Random Forest	0.9958	1.000	0.9957	0.9917
df_26_SMOTE	Random Forest	0.9857	1.000	0.9857	0.9720

AUC: área bajo la curva ROC. F1: media armónica de Precision y Recall. Métricas calculadas sobre el 20% de datos de prueba.

Tabla 3 – Mejores modelos identificados con AutoML en datos balanceados

3.4. Validación final y métricas por clase

La Tabla 4 y la Figura 2 presentan los resultados de la validación final sobre df_22. El Random Forest obtuvo F1 = 0.97 para la clase no deserción y F1 = 0.86 para desertores,

con AUC = 0.95 y Accuracy promedio en validación cruzada = 0.93. El Extra Trees mostró mayor estabilidad (Accuracy = 0.96; AUC = 0.96), superando en un punto porcentual al Random Forest. Con técnica RUS, la Regresión Logística alcanzó Accuracy = 0.89 y AUC = 0.96, siendo adecuada cuando se prioriza interpretabilidad.

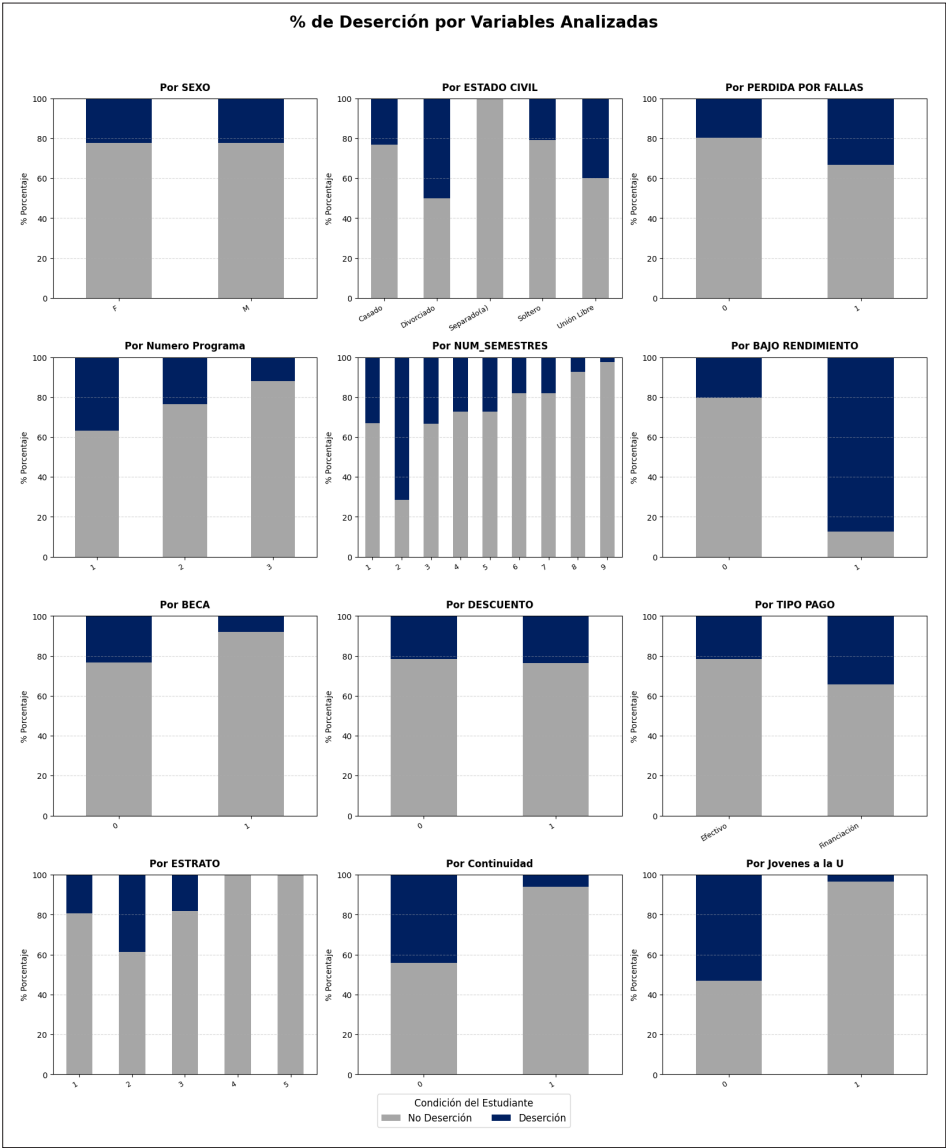


Figura 1 – Distribución de la deserción por variables categóricas (valor 1, azul).
Se observa el efecto protector del programa Jóvenes a la U y la alta asociación entre bajo rendimiento y deserción.

Base / Modelo	Clase	Precision	Recall	F1-Score	Acc. prom.
df_22_SMOTE / Random Forest	No deserción	0.96	0.97	0.97	0.93
	Deserción	0.88	0.83	0.86	
df_22_SMOTE / Extra Trees	No deserción	0.95	0.99	0.97	0.96
	Deserción	0.93	0.78	0.85	
df_22_RUS / Reg. Logística	No deserción	0.97	0.90	0.93	0.89
	Deserción	0.70	0.89	0.78	

Clase 0 = no deserción; Clase 1 = deserción. Acc. prom. = Accuracy promedio en validación cruzada k-fold.

Tabla 4 – Resultados de los modelos finales aplicados a df_22

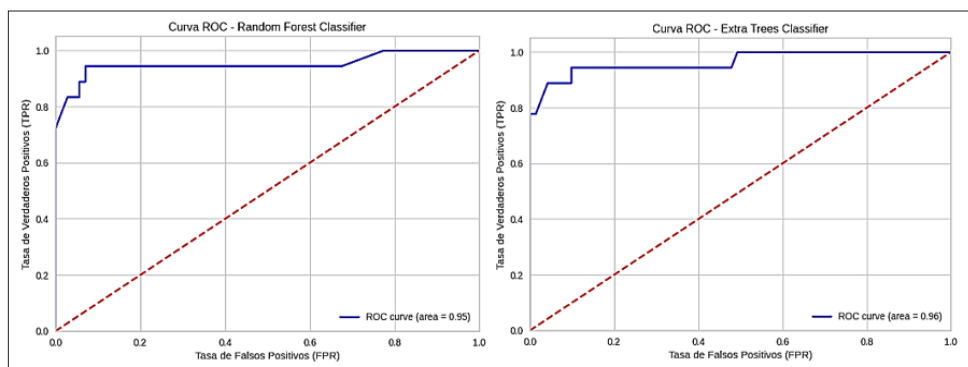


Figura 2 – Curvas ROC para Random Forest (AUC = 0.95) y Extra Trees (AUC = 0.96) sobre df_22 con SMOTE. Ambos modelos muestran elevada capacidad discriminativa bajo distintos umbrales de decisión.

3.5. Importancia de características

Los tres modelos coinciden en las variables de mayor peso predictivo: (1) pertenencia al programa Jóvenes a la U, (2) rendimiento en la Cátedra FUCS del primer semestre, y (3) promedio acumulado anterior. El análisis SHAP sobre la Regresión Logística confirma que valores altos en Jóvenes a la U y en la Cátedra FUCS se asocian con menor probabilidad de desertar ($p < 0.05$). Variables como el sexo y el estado civil, frecuentemente mencionadas en la literatura, perdieron relevancia en los modelos.

4. Discusión

Los hallazgos son consistentes con lo reportado por Fandiño Benavides & Camargo (2022), quienes identificaron al Random Forest como el mejor modelo con exactitudes superiores al 95%. La coincidencia entre Random Forest y Extra Trees en la importancia

de variables refuerza la robustez de los predictores identificados. La estabilidad ligeramente superior del Extra Trees es coherente con su mayor aleatorización en la división de nodos, que reduce la correlación entre árboles.

El efecto protector de Jóvenes a la U merece atención especial. Este programa distrital no solo reduce la carga económica del estudiante, sino que puede generar mayor compromiso institucional y red de apoyo entre pares. Su peso predictivo —superior al del estrato socioeconómico— sugiere que las intervenciones de política pública orientadas al acceso tienen efectos que van más allá de la financiación.

La brecha entre el desempeño de `df_22` y las bases más pequeñas evidencia el rol crítico del volumen de datos en la generalización de modelos de ML para la deserción (Kabathova & Drlik, 2021; Niyogisubizo et al., 2022). Las métricas perfectas en `df_24` y `df_26` son una señal de sobreajuste que debe interpretarse con cautela. La transferencia de este enfoque a sectores como el financiero, las telecomunicaciones y la salud abre líneas de investigación de alto impacto (Ahmad et al., 2019; Habehh & Gohel, 2021).

Las principales limitaciones son: (a) ausencia de variables psicosociales señaladas como determinantes en la literatura (Castro-Martínez & Machuca-Téllez, 2023); (b) tamaño reducido de `df_24` y `df_26`; y (c) alcance de una sola facultad e institución. Futuros estudios deberían ampliar la muestra, incorporar variables de bienestar estudiantil e instrumentos validados, y replicar el modelo en otras instituciones del sector salud en Colombia y Latinoamérica.

5. Conclusiones

Los modelos Random Forest, Extra Trees y Regresión Logística permitieron identificar variables clave para predecir la deserción en una Facultad de Ciencias Administrativas en Salud. El rendimiento académico, la pertenencia al programa Jóvenes a la U y el desempeño en la Cátedra FUCS emergen como predictores robustos y consistentes en los tres modelos. El Extra Trees Classifier, infrecuente en la literatura revisada sobre deserción, superó ligeramente al Random Forest en estabilidad y AUC. El uso combinado de SMOTE y RUS resultó determinante para mejorar la confiabilidad de los resultados ante el desbalanceo de clases. La predicción temprana del riesgo de deserción, sustentada en modelos validados, puede traducirse en estrategias de intervención temprana que mejoren la retención estudiantil y la competitividad institucional.

Agradecimientos

Se agradece al profesor Mike Parra por su colaboración en la descarga y organización de las bases de datos. Este trabajo forma parte del proyecto institucional SIDI 14943 de la Fundación Universitaria de Ciencias de la Salud – FUCS, con financiación no desembolsable en el curso académico 2024/2025.

Referencias

- Aco Tito, A. E., Hanco Condori, B. O., & Pérez Vera, Y. (2023). Análisis comparativo de técnicas de Machine Learning para la predicción de casos de deserción universitaria. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, (51), 84-98. <https://doi.org/10.17013/risti.51.84-98>
- Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6(1), 1-24. <https://doi.org/10.1186/S40537-019-0191-6>
- Behr, A., Giese, M., Tegum, H. and Theune, K. (2020). Dropping out of university: a literature review. *Review of Education*, 8(2), 1-39. <https://doi.org/10.1002/rev3.3202>
- Castiblanco Vargas, W., Fonseca Gómez, L. R., & Pineda-Ríos, W. (2021). Detección de alertas tempranas para la prevención de la deserción estudiantil en una IES mediante Machine Learning. *Conocimiento Global*, 6(S2), 408-426.
- Castro-Martínez, J., & Machuca-Téllez, G. (2023). La deserción universitaria en América Latina: una perspectiva ecológica. *Estudios Pedagógicos*, 49(2), 87-108. <https://doi.org/10.4067/S0718-07052023000200087>
- Fandiño Benavides, A., & Camargo, J. E. (2022). Modelo de predicción para la deserción en la Fundación Universitaria Konrad Lorenz. Konrad Lorenz Editores. <https://repositorio.konradlorenz.edu.co/handle/001/4965>
- Gartner, L., Arango, M., & Villa, E. (2016). Caracterización de la deserción estudiantil en la Universidad de Caldas 2009-2013 a partir del SPADIES. *Revista Latinoamericana de Estudios Educativos*, 12(1), 132-158.
- Guerra, L., Rivero, D., Ortiz, A., Diaz, E., & Quishpe, S. (2020). Modelo de predicción de la deserción universitaria mediante analítica de datos: estrategia para la sustentabilidad. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, (35), 38-47.
- Habehh, H., & Gohel, S. (2021). Machine learning in healthcare. *Current Genomics*, 22(4), 291-300. <https://doi.org/10.2174/1389202922666210705124359>
- Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. *Applied Sciences*, 11(7), 3130. <https://doi.org/10.3390/APP11073130>
- Ministerio de Educación Nacional. (2009). Deserción estudiantil en la educación superior colombiana: metodología de seguimiento, diagnóstico y elementos para su prevención. Ministerio de Educación Nacional de Colombia. https://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-254702_libro_desercion.pdf

- Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: a novel stacked generalization. *Computers and Education: Artificial Intelligence*, 3, 100066. <https://doi.org/10.1016/J.CAEAI.2022.100066>
- Pérez-Gutiérrez, B. R. (2020). Comparación de técnicas de minería de datos para identificar indicios de deserción estudiantil a partir del desempeño académico. *Revista UIS Ingenierías*, 19(1), 193-204. <https://doi.org/10.18273/revuin.v19n1-2020018>
- Raschka, S., & Mirjalili, V. (2019). *Aprendizaje automático con Python*. MARCOMBO.
- Rodríguez Vásquez, C. A., & Benalcázar Palacios, M. E. (2023). Aplicación de algoritmos de Machine Learning para predecir la deserción estudiantil en universidades públicas del Ecuador. Universidad Técnica de Ambato. <http://repositorio.uta.edu.ec/jspui/handle/123456789/38615>
- Rose-Parra, C., Barajas, M., & Restrepo, G. (2023). Factores que influyen en la deserción estudiantil en la educación terciaria de Colombia. *Cienciamatria*, 9(17), 45-56. <https://doi.org/10.35381/cm.v9i17.1122>
- Villegas, B. R., & Núñez Lira, L. A. (2024). Factores asociados a la deserción estudiantil en el ámbito universitario. Una revisión sistemática 2018-2023. *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, 14(28), e671. <https://doi.org/10.23913/ride.v14i28.1923>

Diseño e implementación de un prototipo para la estimación del tiempo de espera en la cafetería del edificio #3 de la Universidad Tecnológica de Panamá basado en visión artificial

Lucia Cárdenas¹, Andrea Pipino¹, Paulo Picota²

lucia.cardenas@utp.ac.pa; andrea.pipino@utp.ac.pa; paulo.picota@utp.ac.pa

¹ Universidad Tecnológica de Panamá, Facultad de Sistemas Computacionales, Lic. en Ingeniería de Sistemas y Computación, Panamá

² Universidad Tecnológica de Panamá, Facultad de Sistemas Computacionales, Departamento de Programación de Computadoras, Panamá

DOI: 10.17013/risti.62.47–62

Resumen: Este estudio detalla el diseño e implementación de un sistema automatizado para la cuantificación del tiempo de espera en la cafetería del edificio #3 de la Universidad Tecnológica de Panamá, basado en técnicas de visión artificial. La arquitectura emplea cámaras IP vinculadas mediante el protocolo RTSP, integrando el modelo YOLOv8 para la detección y seguimiento de personas en una zona de interés delimitada por un polígono. Mediante el uso de algoritmos de conteo y una implementación multihilo para el procesamiento concurrente, el sistema calcula la permanencia en fila y el promedio general del tiempo de espera, notificando los resultados en tiempo real a través de una aplicación móvil institucional. Las pruebas de campo validaron la consistencia del sistema en la detección y precisión métrica. Esta solución busca optimizar la experiencia del usuario y la gestión operativa, demostrando la viabilidad de la visión por computadora en la mejora de servicios públicos.

Palabras-clave: Visión artificial, tiempo de espera, detección de personas, sistema multihilo.

Design and implementation of a prototype for estimating waiting time in the cafeteria of building #3 of the Technological University of Panama based on artificial vision

Abstract: This study details the design and implementation of an automated system for quantifying waiting time in the cafeteria of Building #3 at the Technological University of Panama, based on computer vision techniques. The architecture employs IP cameras connected via the RTSP protocol, integrating the YOLOv8 model for the detection and tracking of people within a region of interest defined by a polygon. Through the use of counting algorithms and a multithreaded implementation for concurrent processing, the system calculates queue dwell

time and the overall average waiting time, delivering results in real time through an institutional mobile application. Field tests validated the system's consistency in detection and measurement accuracy. This solution aims to optimize user experience and operational management, demonstrating the viability of computer vision in improving public services.

Keywords: Computer vision, wait time, people detection, multi-threaded system.

1. Introducción

En la Universidad Tecnológica de Panamá, las cafeterías no solo constituyen un espacio de alimentación y socialización para los estudiantes, sino también un entorno donde el tiempo se convierte en un recurso invaluable, especialmente durante la hora de almuerzo. Debido a la alta concurrencia de estudiantes en este periodo, las filas se vuelven inevitables y el tiempo que se permanece en ellas puede influir de manera positiva o negativa tanto en estudiantes como en personal administrativo, particularmente en aspectos psicológicos como el aumento del estrés, la ansiedad y la sensación de desaprovechamiento del tiempo (Caruelle et al., 2023).

Esta situación puede generar una percepción negativa de la experiencia y del servicio brindado. La ansiedad asociada a las filas no se debe únicamente a la incertidumbre sobre su duración, sino también a la percepción de estas como un obstáculo para alcanzar el objetivo final (Chebat & Filiatrault, 1993), que en este caso corresponde al servicio de alimentación proporcionado por la cafetería universitaria. De este modo, el tiempo en fila se ha convertido en un aspecto crítico y fundamental para mejorar la calidad del servicio y la experiencia del usuario.

Este proyecto tiene como objetivo desarrollar e implementar un sistema automático de conteo de personas y medición del tiempo en fila mediante el uso de tecnologías de visión artificial. La calidad del servicio que reciben las personas suele medirse en función de la fluidez del proceso y del tiempo transcurrido desde que ingresan a la fila hasta que reciben el servicio deseado. El uso de inteligencia artificial no solo facilita el manejo de los datos en tiempo real, sino que también contribuye a una mejor gestión de los recursos y de la atención, brindando al estudiante una experiencia más eficiente y una mayor calidad en el servicio.

El uso de sistemas de visión artificial plantea consideraciones importantes relacionadas con la privacidad y la protección de datos personales de los usuarios. En el diseño de la herramienta propuesta en este trabajo, estos aspectos fueron considerados desde las etapas iniciales del desarrollo. En particular, el sistema fue concebido para operar sin almacenar ni transmitir imágenes o video que permitan la identificación de personas, restringiendo el procesamiento a la extracción de métricas agregadas de tiempo de espera y cantidad de personas.

El presente artículo se estructura de la siguiente manera: en la primera sección, se presentan trabajos relacionados y estudios similares; en la segunda, se describe la metodología del proyecto; en la tercera, se detalla el desarrollo del sistema; en la cuarta, su implementación en la cafetería; en la quinta, la implementación del proyecto

junto con sus pruebas de campo; finalmente, en las dos últimas secciones, se exponen los resultados obtenidos y las conclusiones.

2. Trabajos Relacionados

2.1. Predicción

Diversos estudios han abordado la estimación de tiempos de espera desde un enfoque predictivo, entrenando modelos con datos históricos para anticipar el comportamiento de una fila sin necesidad de monitoreo visual en tiempo real. (Kyritsis & Deriaz, 2019) utilizan técnicas de aprendizaje automático para modelar el comportamiento de filas bancarias a partir de conjuntos de datos históricos, proponiendo además un sistema de creación de filas virtuales que permite a terceros aplicar su modelo optimizado a nuevos escenarios. Su lógica de cálculo del tiempo de permanencia en fila guarda similitud conceptual con la empleada en este trabajo; sin embargo, su propuesta depende de infraestructura adicional para instanciar filas virtuales, lo que introduce una capa de configuración manual ausente en un sistema de monitoreo automático.

En la misma línea, (Rahaman et al., 2018) emplean regresión por vecinos más cercanos ponderada (KNN) para predecir tiempos de espera en filas de taxis aeroportuarios, analizando patrones heterogéneos de viaje mediante variables como demanda y disponibilidad. Los propios autores reconocen que este método de regresión presenta una velocidad de inferencia relativamente lenta y aún susceptible de mejora. (Zhang et al., 2013) abordan un problema similar en el contexto de oficinas de tránsito vehicular (DMV), comparando múltiples enfoques predictivos, clasificación, SVM, regresión lineal, redes bayesianas dinámicas, modelos ocultos de Markov y árboles de decisión, sin identificar un método claramente superior para todos los escenarios evaluados.

Si bien estos trabajos demuestran la viabilidad de anticipar tiempos de espera a partir de datos históricos, ninguno realiza monitoreo en tiempo real de la fila física, condición indispensable para el objetivo de este proyecto: informar al estudiante sobre el estado actual de la cafetería antes de dirigirse a ella.

2.2. Simulación de Eventos Discretos

Un segundo grupo de trabajos aborda la problemática de las filas mediante simulación de eventos discretos en lugar de monitoreo directo. (Usluer, 2025) emplea el software Simio para modelar la relación dinámica entre los horarios de clase de los estudiantes y su afluencia a un comedor universitario, con el fin de optimizar la gestión de turnos de atención. De manera similar, (da Rocha Nascimento et al., 2021) utiliza el software Arena para simular la gestión de una única fila en una agencia bancaria en Brasil, evaluando distintas configuraciones de atención antes de su implementación física.

Estos enfoques resultan valiosos para la planeación estratégica y el rediseño de procesos de servicio, pero dependen de modelos construidos a partir de supuestos y datos de entrada estimados, en lugar de observaciones directas del comportamiento real de la fila. A diferencia de estos trabajos, el sistema propuesto en este estudio no simula el comportamiento de la fila, sino que la mide directamente mediante visión artificial, lo

que permite reflejar condiciones reales y cambiantes, como aglomeraciones inesperadas, que un modelo de simulación no captura en tiempo real.

2.3. Visión Artificial

El grupo de trabajos más cercano al enfoque de este estudio corresponde a sistemas basados en visión artificial para el conteo y seguimiento de personas. (Gozler et al., 2022) proponen un método que combina el detector YOLOv4 con algoritmos de tracking (evaluando KCF, CSRT y MOSSE) para reducir la carga computacional de la detección continua, alternando detección cada 30 cuadros con seguimiento en los cuadros intermedios. Su enfoque comparte con el presente trabajo el uso de detección basada en YOLO y una fórmula de distancia punto-línea equivalente para delimitar la fila; sin embargo, definen la zona de interés mediante un vector simple de dos puntos en lugar de un polígono, lo que limita su aplicabilidad a colas con geometría irregular como la de una cafetería universitaria. Además, no especifican un escenario de despliegue concreto ni un canal de entrega de la información al usuario final, y no contemplan el uso de múltiples cámaras ni una arquitectura concurrente para escalar el sistema a varios puntos de servicio simultáneamente.

(Wang et al., 2021) también emplean YOLO combinado con una red neuronal profunda para estimar tiempos de espera en cajas registradoras, pero con un enfoque orientado a la administración y eficiencia operativa del establecimiento, más que a informar directamente al usuario final. Este enfoque resulta complementario y podría orientar líneas de trabajo futuro de este proyecto, aunque se aleja del objetivo actual de proveer información al estudiante antes de su llegada.

Por su parte, (Piragua & Gomez, 2016) en su tesis utiliza métodos tradicionales de visión artificial, descriptores HOG combinados con clasificadores SVM, para la detección de personas, prescindiendo de modelos de aprendizaje profundo como YOLO. Si bien este enfoque evita la dependencia de arquitecturas de red más complejas, no logra una mejora sustancial en eficiencia computacional frente a modelos más pesados, lo que limita su viabilidad para aplicaciones de conteo en tiempo real con múltiples flujos de video simultáneos.

3. Metodología

El sistema (véase la Figura 1) tiene como objetivo principal medir el tiempo de permanencia de los estudiantes en la fila de la cafetería del edificio 3, utilizando técnicas de visión artificial de detección y seguimiento de objetos, las cuales han mostrado buen desempeño en problemas similares de análisis de filas (Gozler et al., 2022; Wang et al., 2021). Para efectos de este proyecto, se considera que el tiempo de permanencia en el área de interés definida es representativo del tiempo que un estudiante tardaría en comprar la comida, dado que dicha área fue delimitada específicamente para excluir zonas de permanencia no relacionadas con el servicio (ej. mesas o zonas de espera adyacentes), de modo que todo estudiante presente en ella se considera parte de la fila. El cálculo se realiza generando un promedio del tiempo de permanencia de los estudiantes que completaron su recorrido dentro de la zona de interés, es decir, aquellos

cuyo tiempo total, calculado desde su ingreso hasta su salida, ya fue registrado, lo cual permite aproximar el tiempo que experimentaría un nuevo estudiante al integrarse a la fila.

Dado que el sistema propuesto emplea cámaras para la monitorización de personas dentro de la cafetería universitaria, se incorporaron mecanismos orientados a proteger la privacidad de los pasajeros y minimizar la recolección de datos personales. El procesamiento de las imágenes capturadas por la cámara se realiza de forma local, sin transmisión ni almacenamiento persistente de los fotogramas originales. Del proceso de inferencia únicamente se extraen datos agregados, como el conteo de usuarios y el nivel de ocupación estimado a través de archivo de tipo txt, los cuales son transmitidos a través de la red hacia el sistema central. En ningún momento se almacenan ni transmiten imágenes, fotogramas individuales o representaciones que permitan la identificación de las personas detectadas, lo que reduce significativamente el riesgo de exposición de datos personales conforme a lo establecido en la Ley 81 de 2019 sobre Protección de Datos Personales de la República de Panamá (Asamblea Nacional de Panamá, 2019). Asimismo, dado que el objetivo del sistema es exclusivamente la estimación cuantitativa de ocupación y no el reconocimiento o seguimiento de individuos, el diseño no contempla funcionalidades de identificación biométrica ni de reidentificación de usuarios entre distintas unidades o trayectos.

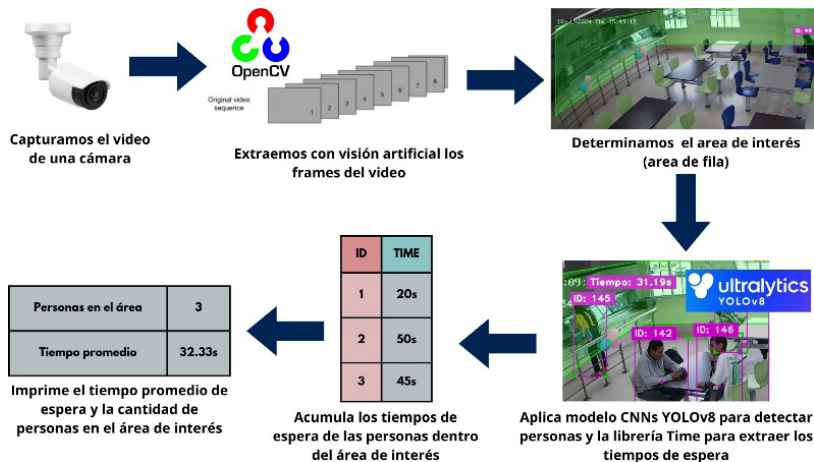


Figura 1 – Esquema del funcionamiento del sistema

El uso de estos datos tiene como objetivo facilitar al estudiante una estimación del tiempo que le tomaría comprar su comida en la cafetería, especialmente en horarios de alta demanda como el almuerzo y la cena. Asimismo, al mostrar la cantidad de personas presentes, aunque se trate de un valor aproximado debido a la posición y al alcance visual de la cámara, se proporciona información suficiente para que el estudiante pueda tomar una mejor decisión al momento de elegir a cuál cafetería dirigirse.

4. Desarrollo de la Herramienta

4.1. Conexión con la cámara

La conexión de la cámara se realiza utilizando el protocolo RTSP (Real Time Streaming Protocol) para la transmisión de los datos en tiempo real desde la cámara en conjunto con la biblioteca de OpenCV. Se utiliza una función de captura de video en la cual se inicializa la conexión con la cámara a través de la dirección IP de esta además de las credenciales de autenticación predefinidas por el protocolo. Debido a las características del tipo de transmisión de datos utilizados y las condiciones de la conexión de red, se implementa en el código una función que garantiza la estabilidad de la conexión al reconectar la cámara al esta perder contacto con el programa.

4.2. Detección de personas y área de interés

Con el objetivo de analizar únicamente la región donde se forma la fila, se delimita un área de interés en la imagen mediante un polígono personalizado, construido a partir de coordenadas correspondientes a los píxeles donde se ubican sus vértices. Para delinear visualmente esta área en la imagen de entrada, se utiliza la librería OpenCV, la cual cuenta con funciones para el manejo de polígonos y permite dibujarlos sobre la imagen a partir de una lista de puntos (ver Figura 2). Esta área se define principalmente para limitar el alcance de la detección en los pasos siguientes, ya que se busca analizar únicamente a los estudiantes en la fila, excluyendo por completo a quienes se encuentran sentados o socializando fuera de ella.

Una vez configurados los límites, el sistema emplea el modelo YOLOv8 para la detección de personas dentro del área definida, con un umbral de confianza de

0.3 limitándolo únicamente a la clase “persona”. Para determinar si una persona detectada se encuentra dentro de dicha área, se verifica si el centroide de su cuadro delimitador (bounding box) se ubica dentro del polígono. Entre los componentes de YOLOv8, la columna vertebral (backbone) desempeña un papel clave en la extracción de características multiescala a partir de la imagen de entrada, generando mapas de características en múltiples resoluciones mediante el apilamiento sucesivo de capas convolucionales y bloques especializados. El modelo finalmente produce cuadros delimitadores y las etiquetas de clase correspondientes para los objetos detectados, tras recibir y analizar los mapas de características provenientes del cuello (neck).

YOLOv8 emplea además un enfoque sin anclas (anchor-free), el cual mejora la precisión de la localización de los cuadros delimitadores y acelera las predicciones. Si bien YOLOv10 es el modelo más rápido y posee la menor cantidad de parámetros gracias a su mecanismo sin NMS (supresión de no-máximos), presenta la menor precisión entre los modelos comparados. YOLOv8 y YOLO11, en cambio, logran un mejor equilibrio entre precisión y velocidad de inferencia, con valores reportados en la literatura de 0.76 de precisión y 0.72 de mAP (Resti et al., 2026).

Para complementar la detección, dado que también es necesario seguir a las personas a lo largo de todo el periodo de grabación, las detecciones generadas por YOLOv8 en cada cuadro se entregan al algoritmo de seguimiento SORT (Simple Online

and Realtime Tracking), ampliamente recomendado para soluciones de bajo costo computacional (Bewley et al., 2016). Este algoritmo destaca por ser liviano, muy rápido (260 fps con alta precisión según sus autores) y compatible con cualquier detector de objetos. Sin embargo, una de sus principales limitaciones, que se retoma más adelante en el desarrollo de los algoritmos, es que no maneja correctamente los objetos que desaparecen temporalmente de la escena.

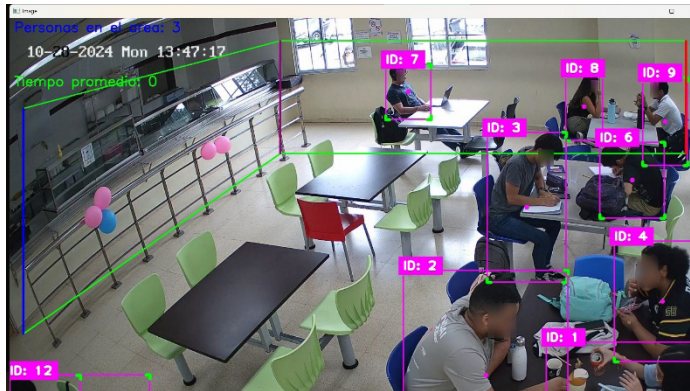


Figura 2 – Limites virtuales definidos en el algoritmo 1, en donde la línea roja es la inicial de la fila y la línea azul la del final del recorrido

4.3. Conteos del tiempo y cálculo de promedio

Algoritmo 1: Este primer algoritmo implementará un sistema de seguimiento mediante límites virtuales o zonas de entrada y salida. Su objetivo es garantizar que una persona solo sea contada si ha pasado por todos los límites establecidos dentro del área de interés. Inicialmente se definen los límites de la zona, se visualizan como líneas colocadas en los extremos y el medio por las cuales las personas pasan sucesivamente. A medida que cada persona pasa la primera línea se inicia el conteo del tiempo, almacenando el momento exacto en el que se cruzó la línea, si la persona pasa los tres límites de forma ordenada se mantiene el seguimiento hasta que abandone la zona de interés y cruce el límite final donde se efectuará el cálculo del tiempo total.

Este proceso se realiza utilizando principalmente funciones y una lógica condicional. Para determinar si la persona ha cruzado cada línea, la función calcula la distancia entre el centro de la persona y cada línea a medida que avanza. Si esta distancia es menor o igual al radio del círculo que representa a la persona en la imagen, se considera que ha cruzado la línea correspondiente. El cálculo de la distancia realizado en el programa se puede expresar de la siguiente forma:

$$d = \frac{|(x_2 - x_1)(y_1 - c_y) - (x_1 - c_x)(y_2 - y_1)|}{\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}} \quad (1)$$

Para el cálculo del tiempo promedio individual que una persona permanece dentro del área, se utiliza una función que realiza la diferencia de tiempo entre el paso por la primera línea y el paso por la línea media, y posteriormente entre la línea media y la línea final, sumando ambos intervalos.

Finalmente, para determinar el tiempo promedio general, se toman en cuenta los resultados promedio de cada persona que haya completado la trayectoria y se realiza una sumatoria dividida por la cantidad de personas que han sido registradas, lo cual puede ser expresado de la siguiente forma:

$$T_{\text{prom}} = \frac{1}{n} \sum_{i=1}^n T_{\text{total}}^{(i)} \quad (2)$$

Algoritmo 2: El segundo algoritmo simplifica el proceso de cálculo del tiempo utilizando únicamente la zona de interés definida. Este enfoque se centra en dar seguimiento a las personas mientras se encuentran dentro del área y calcula el tiempo de permanencia durante su estancia en la zona determinada, eliminando así la necesidad de establecer límites y de utilizar funciones para calcular la distancia del centro de la persona con respecto a las líneas como en la Ecuación (1).

Cada vez que una persona ingresa a la zona, se registra su momento de entrada. Mientras esta permanece dentro del área, el algoritmo actualiza su tiempo de estancia en función la diferencia entre el tiempo de entrada y el tiempo actual al salir de la zona, el tiempo total de permanencia de la persona se almacena en un registro, lo que permite calcular el tiempo promedio en la zona para todas las personas detectadas.

Para cubrir la mayoría de los casos que se presentan durante la ejecución del algoritmo, además de almacenar el tiempo cuando la persona sale del área, también se registra el tiempo cuando algún ID desaparece del frame, situación que ocurre con frecuencia cuando la conexión con la cámara fluctúa. Esto permite considerar un mayor número de mediciones y obtener un valor promedio más cercano al real.



Figura 3 – Área de interés definida en el algoritmo 2

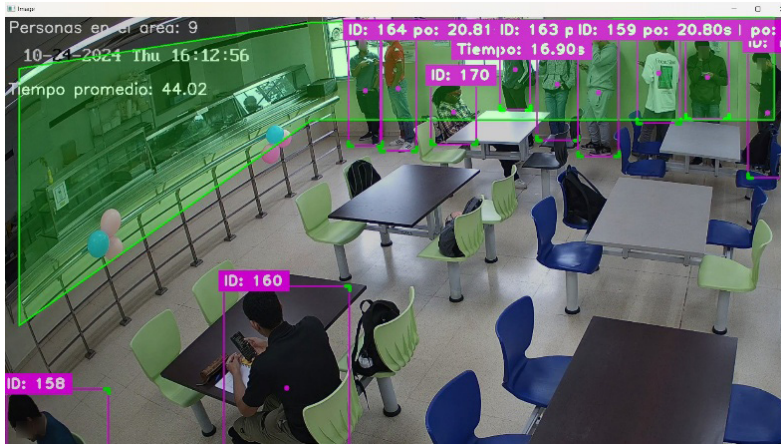


Figura 4 – Cálculo de tiempo en algoritmo 2

5. Implementación y Pruebas

En la implementación del sistema dentro de la cafetería de la universidad primero se cambió el tipo de cámara IP por una donde la conexión de red y de alimentación sea la misma, esto nos permite la conexión directa con la red universitaria (en este caso la red de cada cafetería), facilitando su manejo y por tanto la extracción de su dirección IP.

El sistema de comunicación para el proyecto ubicado en el edificio está dividido en tres sub-redes las cuales son: las redes de servidores, la cámara de video vigilancia y la seguridad Institucional. Como podemos observar en el diagrama (ver la Figura 5) se encuentra el equipo principal del edificio, etiquetado como “Core Edificio 3”, adicional se pueden observar tres equipos de comunicación. Es importante destacar que estas redes se encuentran con seguridad de listas de acceso y configuraciones que permiten protegerlas de acceso no autorizado.

Tomando en cuenta que se desea implementar en todas las cafeterías disponibles en el Campus de la universidad es necesario cambiar el algoritmo a un escenario en donde se capture datos de diferentes cafeterías del campus de manera simultánea, para esto se aplicó el uso de hilos, creando una instancia por cámara, permitiendo que el procesamiento de video de cada una se ejecute en paralelo, de esta manera se crea un programa escalable a la cantidad de cafeterías y/o cámaras de la universidad sin que afecte directamente al algoritmo principal.

Cada hilo se inicializa con el nombre de la cámara, el URL del stream RTSP de la cámara, los puntos que forman el polígono del área de interés, el modelo YOLOv8, cargado para hacer la detección y los archivos donde se guarda el tiempo promedio de permanencia y la cantidad de personas detectadas dentro del área. También se inicializa el “tracker” para el seguimiento de personas y un diccionario para guardar cuándo una persona fue detectada por primera vez. Con el uso de los hilos cada cámara se procesa en tiempo real

sin bloquear a las demás, y mejoramos a su vez el uso de la CPU y manejo óptimo de los recursos de la universidad.

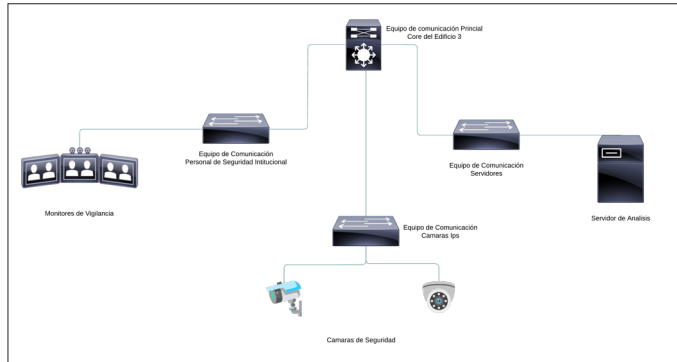


Figura 5 – Diagrama de red del sistema de comunicación del Edificio #3

Finalmente, para mostrar la información obtenida por el programa en la aplicación de la universidad, se debe contar inicialmente con un servidor al cual se pueda enviar los datos. En el caso de este proyecto, se utiliza Apache configurado en un entorno local (xampp). Se crea una carpeta dentro del “document root” (htdocs) de dicho entorno ya que esta permitirá que los archivos sean accesibles de forma remota cuando el servidor esté funcionando. Por último, a través del código se realiza la escritura de dos archivos de texto que contengan el tiempo promedio y la cantidad de personas en la fila respectivamente dentro de la carpeta creada.

5.1. Pruebas de campo

Las pruebas pueden ser divididas en etapas según el avance del desarrollo e investigación del proyecto. Es importante destacar que se describen en relación con la implementación y las pruebas del algoritmo 2, ya que este fue el utilizado como producto final. Todo el procesamiento principal se realizó en una computadora: AMD Ryzen 7 5700U, AMD Radeon Graphics y 16Gb de RAM.

Etapas de prueba 1: Inicialmente, para verificar los conceptos ya conocidos y probar la adición del área de interés al algoritmo, se realizaron pruebas con la cámara web integrada de la computadora y el protocolo HTTP para comprobar la funcionalidad. Las métricas consideradas fueron la cantidad real de personas dentro del área versus la cantidad estimada, y la cantidad real de personas fuera del área versus la cantidad estimada.

Etapas de prueba 2: Esta fase se enfocó en probar el funcionamiento del protocolo RTSP en el conteo de personas dentro del área estimada. Para ello, se utilizó un celular con una aplicación de cámara web (IP Webcam) que permitía verificar la IP del dispositivo y el puerto al cual estaba conectado, para así poder ingresarlo en el algoritmo. Además, para esa etapa se implementó el algoritmo de medición de tiempo, por lo que también se tomó en cuenta la precisión al medir el tiempo dentro del área.

Las métricas a considerar son:

- Tiempo estimado por el programa vs. tiempo real de la transmisión
- Cantidad de personas reales dentro del área vs. cantidad de personas estimadas dentro del área
- Cantidad de personas reales fuera del área vs. cantidad de personas estimadas fuera del área

Etapas de prueba 3: En este punto, se realiza una maqueta a pequeña escala de cómo sería el sistema real cuando esté instalado, utilizando una cámara de vigilancia marca Ezviz, modelo C3TN, con conectividad WiFi y resolución de 1080p para verificar la transmisión en tiempo real y el acceso a los datos generados por el programa. Sin embargo, el protocolo RTSP utilizado no transmitía todos los frames capturados por la cámara, lo que generaba fluctuaciones en los paquetes de datos y provocaba fallos en la ejecución del programa. Para resolverlo, se implementó un módulo de código que detecta la pérdida de paquetes, evita la desconexión de la cámara y permite su reconexión automática. Las métricas importantes fueron la frecuencia de reconexión de la cámara y la conexión correcta del servidor con la aplicación.

Etapas de prueba 4: Finalmente, se ejecutó el sistema en la cámara conectada en la cafetería (ver la Figura 4), donde el propósito final fue validar el funcionamiento remoto del programa y comprobar que todas las métricas mencionadas en las pruebas anteriores fueran registradas correctamente. En esta prueba se realizaron ajustes mínimos al código para que este considerara a todas las personas y las condiciones de tiempo específicas de salida y entrada, además de permitir la reconexión, de modo que todos los identificadores fueran considerados y no hubiera fallas en la detección. Se utilizó la misma cámara de vigilancia marca Ezviz, modelo C3TN, con conectividad WiFi y resolución de 1080p, esta se ubica en un pilar situado en la sección central de la cafetería del edificio 3, apuntando directamente hacia la fila y abarcando desde la caja (donde los estudiantes pagan la comida) hasta la esquina de la pared en la cual esta se forma (ver Figura 3).

6. Resultados

Con el fin de evaluar la efectividad del sistema propuesto para la medición del tiempo de espera en la cafetería universitaria, se llevaron a cabo diversas pruebas de campo controladas en tres etapas progresivas. Estas pruebas permitieron analizar la precisión del conteo de personas, la exactitud en la estimación del tiempo de espera y la estabilidad de la conexión con la cámara.

Para cuantificar el desempeño del sistema, se emplearon las siguientes métricas: error absoluto, error relativo y precisión.

El error absoluto (EA) representa la diferencia, sin considerar el signo, entre el valor estimado por el sistema y el valor real observado manualmente (ground truth), y se calcula como:

$$EA = |V_{real} - V_{estimado}| \quad (3)$$

donde V_{real} corresponde al valor de referencia obtenido mediante conteo manual, y $V_{estimado}$ al valor generado por el sistema propuesto.

Dado que el error absoluto no permite comparar mediciones realizadas en escalas distintas (por ejemplo, distintos niveles de afluencia), se calculó también el error relativo (ER), expresado como porcentaje del valor real:

$$ER = \frac{|V_{real} - V_{estimado}|}{V_{real}} \times 100 \% \tag{4}$$

Esta métrica facilita la interpretación del desempeño del sistema de forma independiente de la magnitud absoluta de cada prueba.

Finalmente, se calculó la precisión (P) del sistema como el complemento del error relativo promedio, expresada de la siguiente manera:

$$P = \left(1 - \frac{1}{n} \sum_{i=1}^n \frac{|V_{real,i} - V_{estimado,i}|}{V_{real,i}} \right) \times 100 \% \tag{5}$$

donde n representa el número total de observaciones realizadas durante las pruebas de campo. Esta métrica permite resumir en un solo valor el desempeño global del sistema a lo largo de las distintas etapas de evaluación.

6.1. Precisión en la detección de personas

Se empleó un cronómetro manual como referencia para validar los resultados del sistema. En total, se realizaron cuatro pruebas, en las cuales se evaluaron tanto el tiempo calculado mostrado por el sistema como su actualización conforme aumentaba la cantidad de personas en la fila. Los resultados evidenciaron una precisión promedio del 83.6% según la Ecuación (5), con un margen de error más elevado en escenarios de alta densidad, como se observa en la Figura 6 en las pruebas 3 y 4, donde se incrementa la diferencia de segundos con respecto al tiempo real de espera.

Este comportamiento se atribuye principalmente a la superposición de individuos en la imagen capturada y al uso de un valor promedio para el cálculo del tiempo. En consecuencia, tanto la cantidad de personas presentes como el momento de incorporación influyen directamente en el resultado final mostrado por el sistema.

Prueba	Tiempo del sistema (s)	Error absoluto (s)	Error relativo (%)
1	37.46	1.54	3.95
2	600.00	25.80	4.12
3	480.00	120.00	20.00
4	687.00	187.00	37.40

Tabla 1 – Resultados de cálculos de error absoluto y relativo en las pruebas de estimación del tiempo de espera por persona, según las Ecuaciones (3) y (4).

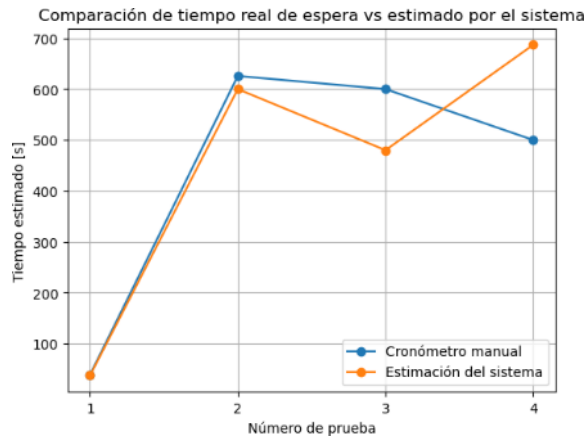


Figura 6 – Gráfico de líneas comparando el tiempo de espera real (tomado por cronómetro) vs el tiempo de espera estimado por el sistema. Se debe tomar en cuenta que este tiempo final mostrado es por persona

6.2. Precisión en la medición del tiempo de espera

Se comparó la cantidad real de personas presentes en la zona de interés con la cantidad estimada por el sistema. Las pruebas revelaron una precisión promedio del 89.72% según la Ecuación (5). Como se observa en la Figura 7, el error aumenta progresivamente con la cantidad de personas en la fila, similar a lo ocurrido con la estimación de tiempo. No obstante, estas diferencias de detección son de apenas 2 o 3 personas, por lo que no representa un problema significativo. Este margen de error se atribuye principalmente a la forma de agrupación entre individuos. Adicionalmente, la iluminación influyó en el desempeño del conteo, ya que la formación de filas junto a ventanas expuestas a la luz solar generó deslumbramiento, afectando la calidad de las detecciones.

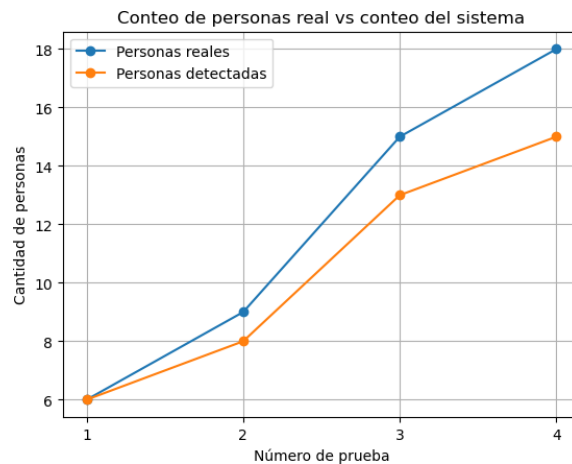


Figura 7 – Gráfico de líneas comparando el número real de personas en la fila vs el número detectado por el sistema

6.3. Estabilidad del sistema y reconexión automática

Durante las pruebas iniciales con las cámaras IP, se registraron desconexiones esporádicas causadas por la inestabilidad de la red y el envío de los datos a través del protocolo RTSP. El módulo de reconexión implementado permitió restablecer la conexión en un promedio de 2.3 segundos, evitando pérdida de datos y mejoras en los valores de estimación de tiempo. Además, el uso de la tecnología de programación multihilo proporcionó un programa escalable que registraba y estimaba el tiempo de varias cámaras al mismo tiempo para de esta manera facilitar la implementación en más cafeterías de la universidad.

6.4. Limitaciones de las pruebas

Si bien los resultados obtenidos durante las pruebas de campo permitieron validar la viabilidad técnica del sistema propuesto, es importante señalar ciertas limitaciones asociadas al proceso de evaluación. El número de observaciones recolectadas fue relativamente reducido, condicionado principalmente por las restricciones de financiamiento y por la disponibilidad de facilidades institucionales para la realización de pruebas controladas dentro de la cafetería universitaria. Estas condiciones limitaron tanto la cantidad de sesiones de prueba realizadas como la diversidad de escenarios de afluencia evaluados.

Adicionalmente, a la fecha de redacción de este trabajo, ya no es posible replicar el entorno específico en el que se llevaron a cabo las pruebas originales, debido a cambios en las condiciones operativas y de acceso a las instalaciones universitarias utilizadas. Esto restringe la posibilidad de ampliar el conjunto de datos bajo las mismas condiciones experimentales o de realizar pruebas adicionales de validación cruzada en el mismo escenario.

7. Conclusiones y Trabajos a futuro

El presente trabajo demostró la viabilidad de implementar un sistema automatizado de estimación de tiempo de espera en cafeterías universitarias mediante visión por computadora y algoritmos de detección de personas. Los resultados obtenidos revelan que el sistema alcanzó valores de precisión que validan su aplicabilidad en entornos reales de alta concurrencia.

El análisis de las pruebas de campo evidenció que el desempeño del sistema se ve afectado principalmente en escenarios de alta densidad, donde la superposición de individuos y las condiciones de iluminación generan mayores márgenes de error. No obstante, las diferencias registradas no comprometen significativamente la utilidad del sistema para proporcionar estimaciones orientativas a los usuarios en escenarios reales.

En consecuencia, si bien los resultados presentados son representativos del desempeño del sistema bajo las condiciones evaluadas, se reconoce que un conjunto de datos más amplio y diverso, idealmente recolectado en múltiples sesiones, horarios y niveles de afluencia, permitiría obtener conclusiones más robustas sobre la generalización del sistema. Se identifica este aspecto como una oportunidad de mejora para trabajos futuros,

en los que se recomienda gestionar con anticipación el acceso sostenido a instalaciones de prueba y, de ser posible, contar con financiamiento dedicado que permita ampliar tanto la escala como la duración de la fase experimental.

La arquitectura multihilo implementada demostró ser robusta y escalable, permitiendo la gestión simultánea de múltiples cámaras IP y reconexiones que mejoran la calidad de los datos. Esta característica resulta fundamental para la expansión del sistema a otras cafeterías dentro del campus universitario, facilitando una red integrada de monitoreo que optimice la experiencia del usuario.

En términos de costo-beneficio, el prototipo utilizó una cámara de vigilancia genérica junto con equipo de cómputo estándar. Si bien no se realizó un análisis financiero formal, la reutilización de infraestructura de red ya disponible en el campus sugiere que una implementación a mayor escala sería económicamente viable.

En síntesis, la aplicabilidad de soluciones basadas en visión por computadora para la gestión de servicios se consolida como una tendencia actual orientada a mejorar tanto la eficiencia operativa como la satisfacción del usuario. La integración de este tipo de tecnologías permite transformar datos visuales en información útil para la toma de decisiones, optimizando la asignación de recursos y reduciendo la percepción negativa asociada a los tiempos de espera. Además, su implementación en entornos universitarios abre la posibilidad de desarrollar sistemas inteligentes escalables que contribuyan a la modernización de los servicios, fortaleciendo la experiencia del estudiante y promoviendo el uso eficiente del tiempo.

Referencias

- Asamblea Nacional de Panamá. (2019). *Ley No. 81 de 26 de marzo de 2019, sobre Protección de Datos Personales*. https://s3-legispan.asamblea.gob.pa/legispan/NORMAS/2010/2019/LEY/Administrador%20Legispan_28743-A_2019_3_29_ASAMBLEA%20NACIONAL_81.pdf
- Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). Simple online and Realtime tracking. <https://github.com/abewley/sort>
- Caruelle, D., Lervik-Olsen, L., & Gustafsson, A. (2023). The clock is ticking—Or is it? Customer satisfaction response to waiting shorter vs. longer than expected during a service encounter. *Journal of Retailing*, 99, 247-264. <https://doi.org/10.1016/j.jretai.2023.03.003>
- Chebat, J. C., & Filiatrault, P. (1993). The impact of waiting in line on consumers. *International Journal of Bank Marketing*, 11, 35-40. <https://doi.org/10.1108/02652329310025938>
- da Rocha Nascimento, M. A., dos Santos, L. M., da Silva, A. M., Bueno, R. C., Machado, S. T., & Tanaka, W. Y. (2021). Discrete event simulation applied to single queue management: a case study at a bank agency. *Independent Journal of Management & Production*, 12, s831-s842. <https://doi.org/10.14807/ijmp.v12i9.1632>

- Gozler, D., Isik, B., & Topal, C. (2022). A Real-time Queue Tracking Method for Waiting Time Estimation. *2022 30th Signal Processing and Communications Applications Conference, SIU 2022*. <https://doi.org/10.1109/SIU55565.2022.9864749>
- Kyritsis, A. I., & Deriaz, M. (2019). A machine learning approach to waiting time prediction in queueing scenarios. In *Proceedings - 2019 2nd International Conference on Artificial Intelligence for Industries, AI4I 2019*, 17-21. <https://doi.org/10.1109/AI4I46381.2019.00013>
- Piragua, A. M. G., & Gomez, W. A. M. (2016). *Sistema De Conteo Automático De Flujo De Personas Por Medio De Visión Artificial* (inf. téc.). Universidad Distrital Francisco José de Caldas.
- Rahaman, M. S., Ren, Y., Hamilton, M., & Salim, F. D. (2018). Wait time prediction for airport taxis using weighted nearest neighbor regression. *IEEE Access*, 6, 74660-74672. <https://doi.org/10.1109/ACCESS.2018.2882580>
- Resti, J., Hidayatullah, P., Syakrani, N., Sholahuddin, M. R., Gelar, T., & Tubagus, R. (2026). Comprehensive Architectural Comparative Review. *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, 10, 341-354. <https://doi.org/10.29207/resti.v10i2.6598>
- Usluer, F. O. (2025). Solving Queue Problems in a Campus Dining Service with Discrete Event Simulation. *Ege Akademik Bakis*, 25, 351-360. <https://doi.org/10.21121/EAB.20250207>
- Wang, B., Li, Y., Cai, X., Rajvanshi, A., & Pramod, D. (2021). Queue Time Estimation in Checkout Counters Using Computer Vision and Deep Neural Network. *Journal of Physics: Conference Series*, 1964, 062108. <https://doi.org/10.1088/1742-6596/1964/6/062108>
- Zhang, Y., Nguyen, L. T., & Zhang, J. (2013). Wait time prediction: How to avoid waiting in lines? *UbiComp 2013 Adjunct - Adjunct Publication of the 2013 ACM Conference on Ubiquitous Computing*, 481-490. <https://doi.org/10.1145/2494091.2495980>

Riesgos de ciberseguridad y estrategias de mitigación en el Gobierno Digital: una revisión sistemática

Néstor Giankeiler Burga Vásquez¹

nburgav@ucvvirtual.edu.pe

¹ Universidad César Vallejo, Lima, Perú.

DOI: 10.17013/risti.62.63–85

Resumen: El avance del Gobierno Digital ha incrementado la exposición de las instituciones públicas a riesgos de ciberseguridad que comprometen la integridad de la información y la continuidad de los servicios estatales. El objetivo del estudio fue identificar los principales riesgos de ciberseguridad en el Gobierno Digital y las estrategias implementadas para su mitigación. Se realizó una revisión sistemática de la literatura siguiendo el protocolo PRISMA 2020 en las bases de datos Scopus, SciELO, Redalyc y Google Académico, seleccionándose 20 estudios publicados entre 2020 y 2025. Los resultados evidencian que los riesgos más frecuentes corresponden a ataques de phishing, vulnerabilidades tecnológicas, deficiencias de infraestructura y limitada capacitación del personal. Asimismo, las estrategias de mitigación más efectivas incluyen autenticación multifactor, monitoreo continuo, cifrado de datos y programas de formación especializada. Se concluye que la ciberseguridad constituye un componente estratégico para fortalecer la confianza ciudadana y garantizar la sostenibilidad del Gobierno Digital.

Palabras clave: Ciberseguridad; Gobierno Digital; Gestión pública; Mitigación de riesgos; Transformación digital.

Cybersecurity Risks and Mitigation Strategies in Digital Government: A Systematic Review

Abstract: The advancement of Digital Government has increased the exposure of public institutions to cybersecurity risks that compromise information integrity and the continuity of public services. The objective of this study was to identify the main cybersecurity risks in Digital Government and the strategies implemented for their mitigation. A systematic literature review was conducted following the PRISMA 2020 protocol using the Scopus, SciELO, Redalyc, and Google Scholar databases, selecting 20 studies published between 2020 and 2025. The results show that the most frequent risks include phishing attacks, technological vulnerabilities, infrastructure deficiencies, and limited staff training. Likewise, the most effective mitigation strategies include multi-factor authentication, continuous monitoring, data encryption, and specialized training programs. It is concluded that cybersecurity constitutes a strategic component for strengthening citizen trust and ensuring the sustainability of Digital Government.

Keywords: Cybersecurity; Digital Government; Public management; Risk mitigation; Digital transformation.

1. Introducción

El acelerado desarrollo de las Tecnologías de la Información y las Comunicaciones (TIC) ha transformado de manera significativa la administración pública en distintos países, promoviendo la digitalización de servicios y la consolidación del Gobierno Digital como un nuevo modelo de gestión pública. El Gobierno Digital busca optimizar los procesos administrativos, fortalecer la transparencia institucional y mejorar la interacción entre el Estado y la ciudadanía mediante el uso estratégico de herramientas tecnológicas (Organización para la Cooperación y el Desarrollo Económico [OCDE], 2022). En este contexto, la transformación digital del sector público ha permitido ampliar la cobertura de los servicios estatales y mejorar la eficiencia operativa de las instituciones gubernamentales.

No obstante, la expansión del Gobierno Digital también ha incrementado la exposición de las instituciones públicas a riesgos de ciberseguridad, los cuales comprometen la confidencialidad, integridad y disponibilidad de la información pública. La Agencia de la Unión Europea para la Ciberseguridad (European Union Agency for Cybersecurity [ENISA], 2021) advierte que el incremento del uso de plataformas digitales gubernamentales ha generado nuevas vulnerabilidades asociadas con ataques de phishing, ransomware, robo de credenciales y accesos no autorizados a bases de datos institucionales. Estas amenazas afectan no solo la continuidad operativa de los servicios públicos, sino también la confianza ciudadana en los procesos de modernización digital del Estado.

En este contexto, la identificación de estos riesgos y la implementación de estrategias de mitigación se han convertido en prioridades para garantizar la sostenibilidad del Gobierno Digital. La OCDE (2022) sostiene que la digitalización del sector público debe ir acompañada de políticas de seguridad que permitan reducir los riesgos tecnológicos y proteger la información sensible administrada por las instituciones públicas. De manera similar, ENISA (2020) señala que muchos gobiernos han avanzado en la implementación de marcos regulatorios para la seguridad digital; sin embargo, persisten limitaciones relacionadas con la capacitación del personal, la interoperabilidad tecnológica y la ausencia de protocolos homogéneos entre instituciones.

En América Latina, la situación presenta desafíos adicionales debido a restricciones presupuestarias, infraestructura tecnológica insuficiente y escasa especialización del recurso humano en materia de seguridad informática. Cornejo (2025) indica que numerosas administraciones públicas de la región continúan operando con sistemas obsoletos y políticas fragmentadas de protección digital, lo que incrementa la probabilidad de incidentes cibernéticos. Asimismo, el Instituto Nacional de Estadística y Geografía (INEGI, 2021) reportó que menos de la mitad de las entidades públicas en México cuentan con políticas formales de ciberseguridad, evidenciando brechas estructurales en la región para garantizar entornos digitales seguros.

En el contexto peruano, si bien se han desarrollado iniciativas orientadas a modernizar la administración pública mediante la Política Nacional de Transformación Digital (Presidencia del Consejo de Ministros [PCM], 2020), la implementación de estrategias de ciberseguridad aún presenta importantes desigualdades entre instituciones y niveles de gobierno. Esta situación genera vulnerabilidades operativas que afectan la continuidad de los servicios digitales y debilitan la confianza de los ciudadanos en las plataformas estatales. Según León-Gutiérrez et al. (2022), muchas instituciones públicas peruanas todavía presentan debilidades relacionadas con la actualización de software, protección de datos y gestión de incidentes de seguridad.

Desde una perspectiva conceptual, la ciberseguridad puede definirse como el conjunto de procesos destinados a proteger los sistemas de información frente a amenazas internas y externas que puedan comprometer la confidencialidad, integridad o disponibilidad de los datos (International Organization for Standardization [ISO/IEC], 2018). Shamala, Ahmad, & Yusof (2017) sostienen que la gestión de riesgos de ciberseguridad debe considerar no solo factores tecnológicos, sino también aspectos humanos y organizacionales. En la misma línea, Von Solms & Van Niekerk (2013) afirman que la seguridad digital en el sector público requiere una visión integral que combine gobernanza institucional, prevención tecnológica y formación continua del personal.

Con el propósito de fortalecer la protección digital de los gobiernos, diversos organismos internacionales han propuesto marcos de referencia orientados a la gestión de riesgos. Entre los más relevantes se encuentran el NIST Cybersecurity Framework desarrollado por el National Institute of Standards and Technology (NIST, 2018) y la norma ISO/IEC 27005:2018, que establecen lineamientos para la identificación, evaluación y tratamiento de amenazas digitales en organizaciones públicas. Ambos enfoques promueven la incorporación de mecanismos de seguridad desde la fase de diseño de los servicios digitales, permitiendo una respuesta más eficiente frente a incidentes y una mayor resiliencia institucional.

A pesar del crecimiento del Gobierno Digital, la literatura científica muestra resultados dispersos sobre cuáles son los principales riesgos de ciberseguridad que enfrentan las instituciones públicas y qué estrategias han demostrado mayor efectividad para su mitigación. Esta dispersión limita la formulación de lineamientos integrales para fortalecer la seguridad digital en el sector público. En este contexto, Castillo, Bruzza & Tupia (2021) desarrollaron una revisión sistemática sobre las amenazas no intencionales de ciberseguridad generadas por personal interno en instituciones públicas, destacando que los factores humanos constituyen una fuente relevante de vulnerabilidades en los servicios de gobierno electrónico. No obstante, su estudio se centró exclusivamente en las amenazas internas y no analizó de manera integrada los riesgos tecnológicos, de infraestructura y las estrategias de mitigación en el contexto del Gobierno Digital, lo que evidencia la necesidad de revisiones más amplias como la presentada en esta investigación.

Frente a esta problemática, la presente investigación plantea la siguiente pregunta de investigación: ¿Cuáles son los principales riesgos de ciberseguridad en el Gobierno Digital y qué estrategias se implementan para su mitigación? A partir de esta interrogante, se

establece como objetivo general identificar los principales riesgos de ciberseguridad en el Gobierno Digital y las estrategias implementadas para su mitigación. Como objetivos específicos, se plantean clasificar los tipos de riesgos de ciberseguridad reportados en la literatura científica, analizar las estrategias aplicadas para mitigar los riesgos de ciberseguridad en entornos de Gobierno Digital y conocer la percepción sobre la relación entre ciberseguridad y confianza ciudadana.

En ese sentido, la presente revisión sistemática busca sintetizar la evidencia científica disponible con el propósito de identificar los principales riesgos de ciberseguridad presentes en el entorno del Gobierno Digital y analizar las estrategias implementadas para su mitigación en distintos contextos institucionales, contribuyendo al fortalecimiento de la seguridad digital y la sostenibilidad de los servicios públicos digitales.

2. Método

2.1. Diseño del estudio

La presente investigación se desarrolló mediante una revisión sistemática de la literatura con enfoque cualitativo-documental, orientada a identificar y sintetizar la evidencia científica disponible sobre los principales riesgos de ciberseguridad en el Gobierno Digital y las estrategias implementadas para su mitigación. Este tipo de estudio permite recopilar, evaluar e integrar de manera estructurada los resultados de investigaciones previas, proporcionando una visión amplia y fundamentada del estado del conocimiento sobre un fenómeno específico (Olaya Moncada, 2024).

El diseño metodológico siguió las directrices del protocolo PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), empleadas para garantizar la transparencia, reproducibilidad y trazabilidad del proceso de revisión sistemática. En consecuencia, la selección de los estudios comprendió las etapas de identificación, cribado, evaluación de elegibilidad e inclusión, documentando de manera sistemática las decisiones adoptadas durante cada fase del proceso (Yepes-Núñez et al., 2021).

2.2. Estrategia de búsqueda

La búsqueda bibliográfica se realizó de manera sistemática entre enero y febrero de 2026 en las bases de datos Scopus, SciELO, Redalyc y Google Académico, seleccionadas por su amplia cobertura de investigaciones relacionadas con Gobierno Digital, administración pública, transformación digital y ciberseguridad. La inclusión de Google Académico tuvo como finalidad complementar la recuperación de estudios potencialmente relevantes no indexados en las demás bases de datos, reduciendo el riesgo de omisión de evidencia científica relevante.

La estrategia de búsqueda fue diseñada utilizando descriptores en español e inglés relacionados con el objeto de estudio. Para ampliar la recuperación documental se emplearon operadores booleanos (AND y OR), permitiendo combinar términos equivalentes y conceptos relacionados. Asimismo, las estrategias de búsqueda fueron adaptadas a las características y opciones de búsqueda disponibles en cada base de

datos, con el fin de optimizar la recuperación de literatura científica sin comprometer la pertinencia temática de los estudios seleccionados.

Se aplicaron filtros para recuperar únicamente artículos científicos publicados entre 2020 y 2025, escritos en español o inglés, con acceso al texto completo y relacionados con los riesgos de ciberseguridad y las estrategias de mitigación en el contexto del Gobierno Digital.

Posteriormente, los registros recuperados fueron exportados a una matriz de análisis elaborada en Microsoft Excel, donde se consolidó la información bibliográfica, se eliminaron los registros duplicados y se organizó la evidencia para las etapas de cribado, evaluación de elegibilidad e inclusión de estudios, siguiendo las directrices establecidas por el protocolo PRISMA 2020.

Los principales descriptores empleados durante la búsqueda se presentan en la Tabla 1, mientras que las cadenas de búsqueda utilizadas en cada base de datos se muestran en la Tabla 2.

Categoría	Palabras clave en español	Palabras clave en inglés
Riesgos de ciberseguridad	Riesgos de Ciberseguridad	Cybersecurity, Cyber Security, Security Risks
Gobierno Digital	Gobierno Digital	Digital Government, E-Government, Electronic Government
Estrategias de mitigación	Estrategias de Seguridad, Mitigación de Riesgos	Security Strategies, Risk Mitigation, Cybersecurity Management

Tabla 1 –Descriptores y palabras clave empleadas en la búsqueda bibliográfica

Base de datos	Estrategia de búsqueda	Registros identificados
Scopus	("Cybersecurity" OR "Security Risks") AND ("Digital Government" OR "E-Government")	50
SciELO	("Cybersecurity" OR "Riesgos de Ciberseguridad") AND ("Digital Government" OR "Gobierno Digital")	35
Redalyc	("Gobierno Digital") AND ("Ciberseguridad") AND ("Estrategias de Seguridad")	25
Google Académico	("Cybersecurity" AND "Digital Government") OR ("Riesgos de Ciberseguridad" AND "Gobierno Digital")	30
Total		140

Tabla 2 –Cadenas de búsqueda utilizadas en las bases de datos

2.3.Criterios de selección

La selección de los estudios se realizó mediante criterios de inclusión y exclusión previamente definidos, con el propósito de garantizar la pertinencia, actualidad y calidad metodológica de la evidencia científica analizada. Estos criterios fueron establecidos antes del inicio del proceso de cribado, en concordancia con las recomendaciones del

protocolo PRISMA 2020, con el fin de asegurar la transparencia y reproducibilidad de la revisión sistemática.

Se incluyeron estudios empíricos publicados entre 2020 y 2025, en idioma español o inglés, que abordaran de manera explícita los riesgos de ciberseguridad, las estrategias de mitigación o ambos aspectos en el contexto del Gobierno Digital o del gobierno electrónico en instituciones públicas. Asimismo, los estudios debían estar disponibles en texto completo para permitir la evaluación integral de su contenido.

Se excluyeron publicaciones anteriores a 2020, documentos en otros idiomas, revisiones narrativas, artículos de opinión, editoriales, capítulos de libro, actas de congresos y estudios desarrollados exclusivamente en organizaciones del sector privado o sin evidencia empírica relacionada con el objeto de estudio. También fueron descartados los documentos sin acceso al texto completo.

Los criterios utilizados durante el proceso de selección documental se presentan en la Tabla 3.

Criterios de inclusión	Criterios de exclusión
Publicaciones entre 2020 y 2025	Publicaciones anteriores a 2020
Artículos científicos con evidencia empírica	Revisiones narrativas, artículos de opinión, editoriales, capítulos de libro y actas de congresos
Estudios relacionados con riesgos de ciberseguridad y/o estrategias de mitigación en Gobierno Digital	Estudios centrados exclusivamente en organizaciones del sector privado o sin relación con el Gobierno Digital
Publicaciones en español o inglés	Publicaciones en otros idiomas
Texto completo disponible	Documentos sin acceso al texto completo

Tabla 3 – Criterios de inclusión y exclusión

2.4.Procedimiento de selección

El proceso de selección de los estudios siguió las cuatro etapas propuestas por la declaración PRISMA 2020: identificación, cribado, elegibilidad e inclusión.

Durante la fase de identificación se recuperaron 140 registros provenientes de las bases de datos consultadas: Scopus (50), SciELO (35), Redalyc (25) y Google Académico (30). Todos los registros fueron consolidados en una matriz de análisis en Microsoft Excel, donde se verificó la existencia de duplicados y se organizó la información bibliográfica para su posterior evaluación.

En la etapa de eliminación de duplicados se identificaron y excluyeron 40 registros repetidos, obteniéndose un total de 100 estudios únicos para la fase de cribado.

Posteriormente, durante el cribado, se revisaron los títulos y resúmenes de los artículos con el propósito de verificar su relación con el objetivo de la investigación. En esta etapa se excluyeron 45 estudios por no abordar de forma específica los riesgos de ciberseguridad o las estrategias de mitigación en el contexto del Gobierno Digital.

Los 55 artículos restantes fueron evaluados mediante la lectura del texto completo durante la fase de elegibilidad. En esta etapa se excluyeron 35 estudios debido a que no cumplían con los criterios metodológicos previamente establecidos, principalmente por presentar un enfoque diferente al objeto de estudio, carecer de evidencia empírica o desarrollarse en contextos distintos al sector público.

Finalmente, la revisión sistemática incluyó 20 artículos científicos, los cuales conformaron la muestra definitiva para el análisis cualitativo de la evidencia. El proceso completo de identificación, selección, elegibilidad e inclusión de los estudios se presenta mediante el diagrama de flujo PRISMA 2020.

2.5. Análisis de la información

Los artículos seleccionados fueron organizados en una matriz de análisis documental que permitió registrar información relacionada con autor, año de publicación, país de procedencia, tipo de institución pública analizada, riesgos de ciberseguridad identificados, estrategias de mitigación aplicadas y principales hallazgos.

Posteriormente, se realizó un análisis temático comparativo de la evidencia seleccionada para identificar patrones recurrentes y clasificar los hallazgos en dos dimensiones principales: riesgos de ciberseguridad y estrategias de mitigación en el Gobierno Digital.

Este procedimiento permitió sintetizar de manera estructurada la evidencia disponible sobre la gestión de riesgos de ciberseguridad en gobiernos digitales, facilitando la interpretación de las tendencias predominantes reportadas en la literatura científica reciente.

2.6. Evaluación de la calidad metodológica

La calidad metodológica de los estudios incluidos fue evaluada mediante las listas de verificación crítica del Joanna Briggs Institute (JBI), seleccionando el instrumento correspondiente según el diseño metodológico identificado en cada investigación. Esta evaluación permitió valorar la rigurosidad de los estudios seleccionados y garantizar la confiabilidad de la evidencia sintetizada en la presente revisión sistemática.

Los criterios considerados incluyeron la claridad de los objetivos de investigación, la adecuación del diseño metodológico, la descripción de la población o muestra, los procedimientos utilizados para la recolección de información, la consistencia del análisis de datos y la correspondencia entre los resultados obtenidos y las conclusiones planteadas por los autores.

Cada estudio fue evaluado mediante una escala dicotómica de cumplimiento (Sí = 1; No = 0; No aplica = excluido del cálculo). Posteriormente, se calculó el porcentaje de cumplimiento metodológico para cada artículo mediante la relación entre los criterios cumplidos y el número total de criterios aplicables del instrumento utilizado.

Para interpretar los resultados de la evaluación se establecieron los siguientes niveles de calidad metodológica: calidad alta ($\geq 80\%$), calidad moderada ($60\% - 79\%$) y calidad baja ($< 60\%$). Estos criterios permitieron determinar la consistencia metodológica de los estudios incluidos y sustentar la confiabilidad de los hallazgos obtenidos.

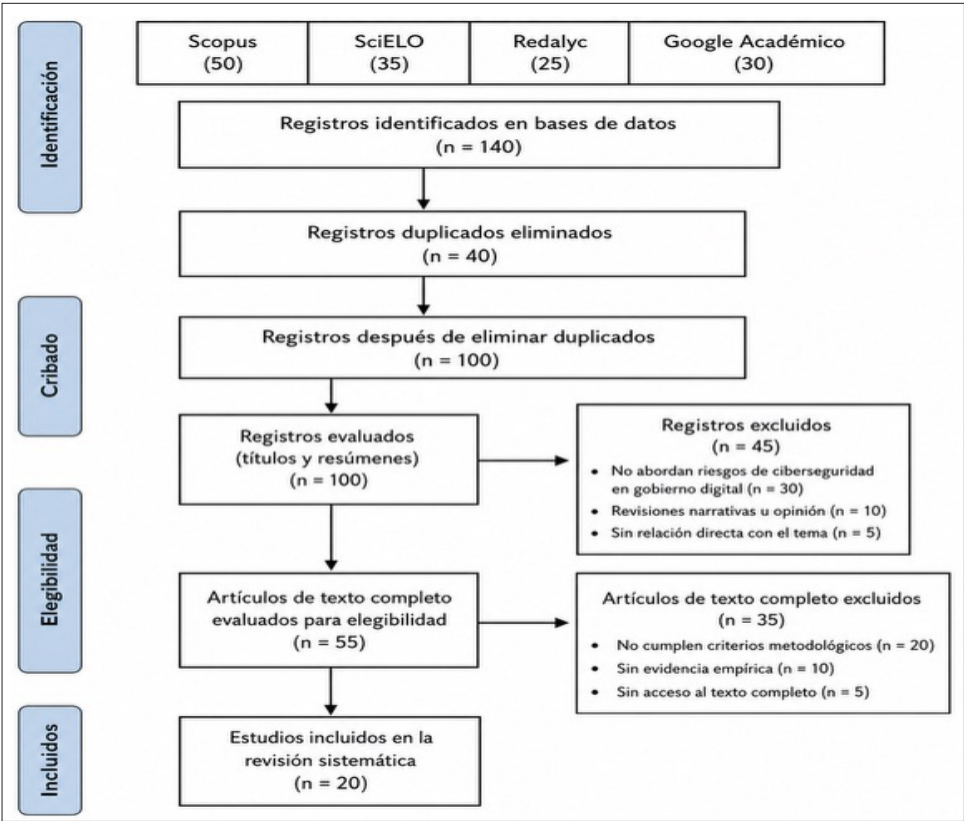


Figura 1 – Diagrama de flujo PRISMA del proceso de selección de estudios.

Como resultado de la evaluación, los 20 estudios seleccionados alcanzaron niveles metodológicos aceptables, predominando investigaciones clasificadas con calidad alta, por lo que fueron incorporadas para el análisis cualitativo final de la revisión sistemática. La síntesis de la evaluación metodológica se presenta en la Tabla 4.

Nº	Estudio	Tipo de evidencia evaluada	Instrumento JBI aplicado	Resultado de evaluación	Nivel de calidad
1	Caldarulo, Welch y Feeney (2022)	Encuesta a gobiernos locales (321 entidades)	JBICritical Appraisal Checklist for Analytical Cross Sectional Studies	Cumple la mayoría de criterios metodológicos, con adecuada descripción de muestra y análisis de resultados	Alta

Nº	Estudio	Tipo de evidencia evaluada	Instrumento JBI aplicado	Resultado de evaluación	Nivel de calidad
2	Pethig, Kroenung y Nöltner (2021)	Encuesta a usuarios de servicios de gobierno electrónico (742 participantes)	JBIC Critical Appraisal Checklist for Analytical Cross Sectional Studies	Presenta objetivos claros, muestra definida y procedimientos de análisis adecuados	Alta
3	Ahn y Chen (2022)	Encuesta a funcionarios públicos (214 participantes)	JBIC Critical Appraisal Checklist for Analytical Cross Sectional Studies	Cumple criterios relacionados con selección de participantes y medición de variables	Alta
4	Chang y Huang (2023)	Análisis de organizaciones participantes en redes de información (96 organizaciones)	JBIC Critical Appraisal Checklist for Analytical Cross Sectional Studies	Presenta metodología clara y adecuada relación entre objetivos y resultados	Alta
5	Oliveira y Santos (2023)	Estudio sobre gobiernos municipales brasileños (74 municipalidades)	JBIC Critical Appraisal Checklist for Analytical Cross Sectional Studies	Evidencia adecuada para analizar estrategias de seguridad institucional	Alta
6	Hassan y Duarte (2024)	Análisis documental de literatura científica (185 artículos)	JBIC Critical Appraisal Checklist for Text and Opinion / revisión documental	Presenta criterios de análisis definidos y síntesis temática consistente	Moderada
7	Rossi y Bianchi (2022)	Análisis de plataformas de gobierno electrónico (58 municipalidades)	JBIC Critical Appraisal Checklist for Analytical Cross Sectional Studies	Describe adecuadamente el contexto institucional y los resultados obtenidos	Alta
8	Kim y Park (2022)	Evaluación de agencias gubernamentales (120 entidades)	JBIC Critical Appraisal Checklist for Analytical Cross Sectional Studies	Presenta metodología estructurada y evidencia suficiente para sustentar resultados	Alta
9	Chen y Li (2024)	Análisis de plataformas digitales gubernamentales (100 plataformas)	JBIC Critical Appraisal Checklist for Analytical Cross Sectional Studies	Identifica claramente variables relacionadas con vulnerabilidades tecnológicas	Alta
10	Sadik-Zada, Aliyev y Mammadov (2024)	Análisis comparativo internacional (40 países)	JBIC Critical Appraisal Checklist for Analytical Cross Sectional Studies	Presenta diseño comparativo y coherencia entre resultados y conclusiones	Alta
11	Martínez y Gómez (2023)	Encuesta a funcionarios municipales (132 participantes)	JBIC Critical Appraisal Checklist for Analytical Cross Sectional Studies	Cumple criterios de claridad metodológica y análisis de resultados	Alta

Nº	Estudio	Tipo de evidencia evaluada	Instrumento JBI aplicado	Resultado de evaluación	Nivel de calidad
12	Frandell y Feeney (2022)	Encuesta a gobiernos locales (500 entidades)	JBICritical Appraisal Checklist for Analytical Cross Sectional Studies	Presenta adecuado enfoque sociotécnico y consistencia metodológica	Alta
13	Norris y Mateczun (2023)	Encuesta a responsables de TI de 14 gobiernos locales	JBICritical Appraisal Checklist for Analytical Cross Sectional Studies	Metodología claramente descrita y resultados alineados con objetivos	Alta
14	Altamirano-de-la-Borda (2021)	Análisis de seguridad de información en 52 instituciones públicas	JBICritical Appraisal Checklist for Analytical Cross Sectional Studies	Presenta evidencia relevante, aunque con limitaciones por alcance institucional	Moderada
15	León-Gutiérrez et al. (2022)	Análisis documental de 64 documentos institucionales	JBICritical Appraisal Checklist for Text and Opinion / revisión documental	Presenta análisis estructurado y evidencia pertinente para el contexto peruano	Moderada
16	Cornejo (2025)	Evaluación mediante modelo de madurez en organizaciones iberoamericanas	JBICritical Appraisal Checklist for Analytical Cross Sectional Studies	Presenta metodología definida y criterios de evaluación establecidos	Alta
17	Henman, Graham y Lata (2021)	Análisis de 38 portales gubernamentales	JBICritical Appraisal Checklist for Analytical Cross Sectional Studies	Presenta criterios claros para evaluar gobernanza de datos	Alta
18	Ribeiro, Fonte y Silva (2025)	Análisis de 67 plataformas gubernamentales	JBICritical Appraisal Checklist for Analytical Cross Sectional Studies	Presenta evidencia suficiente sobre vulnerabilidades digitales	Alta
19	Yang y Yang (2025)	Análisis de asociaciones público-tecnológicas (82 asociaciones)	JBICritical Appraisal Checklist for Analytical Cross Sectional Studies	Metodología consistente con el objetivo del estudio	Alta
20	Busetti y Scanni (2025)	Análisis de 110 reportes institucionales	JBICritical Appraisal Checklist for Analytical Cross Sectional Studies	Presenta criterios definidos para evaluar mecanismos de reporte	Alta

Tabla 4 – Evaluación de calidad metodológica mediante listas de verificación JBI

La matriz completa de evaluación metodológica mediante las listas JBI se encuentra disponible como material suplementario.

3. Resultados

La revisión sistemática permitió analizar 20 estudios científicos relacionados con los riesgos de ciberseguridad y las estrategias de mitigación en el Gobierno Digital. La información extraída se organizó considerando el país de procedencia, año de publicación, objetivo del estudio, tipo de muestra y principales hallazgos. Esta sistematización permitió identificar los riesgos de ciberseguridad más frecuentes en los entornos gubernamentales digitales y las estrategias implementadas para su mitigación.

3.1. Caracterización general de los estudios analizados.

Del total de investigaciones incluidas, dos estudios correspondieron al contexto peruano, representando el 10 % de la producción científica analizada. El 90 % restante se distribuyó entre Estados Unidos, Alemania, Taiwán, Brasil, Italia, Corea del Sur, China, Azerbaiyán, España, Australia, Portugal y Ecuador. La mayor parte de la evidencia científica provino de países con procesos consolidados de transformación digital en el sector público, destacando especialmente Estados Unidos por la presencia de estudios enfocados en la capacidad institucional, gestión de riesgos y respuesta ante incidentes de ciberseguridad en gobiernos locales.

Como se observa en la Tabla 5, los estudios revisados muestran que la digitalización de los servicios públicos incrementa la eficiencia administrativa; sin embargo, también expone a las instituciones a nuevas amenazas relacionadas con la protección de datos, la continuidad operativa y la confianza ciudadana.

Nº	Autor	Año	País	Objetivo	Muestra	Principal hallazgo	Objetivos abordados
1	Caldarulo, Welch & Feeney	2022	Estados Unidos	Analizar la capacidad de ciberseguridad en gobiernos locales.	321 gobiernos locales.	La capacitación y la planificación institucional fortalecen la resiliencia digital.	OE1, OE2
2	Pethig, Kroenung & Nöltner	2021	Alemania	Examinar los factores de evasión del uso de servicios de gobierno digital.	742 usuarios de e-government.	La desconfianza en la seguridad reduce la adopción de plataformas digitales.	OG, OE1
3	Ahn & Chen	2022	Estados Unidos	Analizar el impacto de la inteligencia artificial y la transformación digital en la administración pública.	214 funcionarios públicos.	La formación tecnológica mejora el uso seguro de herramientas digitales.	OE2
4	Chang & Huang	2023	Taiwán	Examinar redes de intercambio de información sobre ciberseguridad.	96 organizaciones multisectoriales.	El intercambio de información fortalece la resiliencia institucional.	OE2

Nº	Autor	Año	País	Objetivo	Muestra	Principal hallazgo	Objetivos abordados
5	Oliveira & Santos	2023	Brasil	Evaluar estrategias de seguridad en gobiernos municipales.	74 municipalidades.	El monitoreo continuo reduce riesgos tecnológicos.	OE2
6	Hassan & Duarte	2024	Brasil	Analizar la evolución de la ciberseguridad en la administración pública.	185 artículos científicos indexados.	La capacitación fortalece la gestión de seguridad institucional.	OE1, OE2
7	Rossi & Bianchi	2022	Italia	Analizar la seguridad en plataformas de gobierno electrónico.	58 municipalidades italianas.	El cifrado de información mejora la protección de datos.	OE2
8	Kim & Park	2022	Corea del Sur	Evaluar estrategias de mitigación de riesgos digitales.	120 agencias gubernamentales.	El monitoreo constante reduce incidentes cibernéticos.	OE2
9	Chen & Li	2024	China	Analizar riesgos cibernéticos en plataformas gubernamentales.	100 plataformas digitales.	Persisten vulnerabilidades críticas en los servicios públicos digitales.	OE1
10	Sadik-Zada, Aliyev & Mammadov	2024	Azerbaiyán	Analizar la relación entre digitalización y fraude administrativo.	Datos de 40 países.	La ciberseguridad fortalece la legitimidad institucional.	OG, OE2
11	Martínez & Gómez	2023	España	Analizar riesgos y estrategias digitales en municipios.	132 funcionarios municipales.	La capacitación mejora la seguridad institucional.	OE1, OE2
12	Frاندell & Feeney	2022	Estados Unidos	Analizar las amenazas de ciberseguridad en gobiernos locales desde una perspectiva sociotécnica	500 gobiernos locales	La integración de factores sociales y técnicos fortalece la capacidad institucional frente a incidentes cibernéticos	OE1, OE2
13	Norris & Mateczun	2023	Estados Unidos	Analizar los ciberataques y las estrategias de respuesta implementadas en gobiernos locales.	14 gobiernos locales de Estados Unidos (3 condados y 11 ciudades), representados por directores de TI y seguridad informática.	Los gobiernos locales evidencian limitaciones en preparación organizacional, capacidades institucionales y mecanismos de respuesta frente a incidentes de ciberseguridad.	OE2

Nº	Autor	Año	País	Objetivo	Muestra	Principal hallazgo	Objetivos abordados
14	Altamirano-de-la-Borda	2021	Perú	Analizar la seguridad de la información en entidades públicas.	52 instituciones públicas.	Persisten debilidades en los estándares de protección de datos.	OE1
15	León-Gutiérrez et al.	2022	Perú	Revisar los avances de la ciberseguridad en el gobierno digital peruano.	64 documentos institucionales.	Existen brechas en infraestructura y capacitación tecnológica.	OE1, OE2
16	Cornejo	2025	Ecuador	Analizar la madurez en ciberseguridad y la resiliencia digital en organizaciones iberoamericanas.	PYMES iberoamericanas evaluadas mediante modelo de madurez.	La madurez en ciberseguridad fortalece la resiliencia digital y la gestión preventiva de riesgos.	OE2
17	Henman, Graham & Lata	2021	Australia	Analizar desafíos de gobernanza de datos en gobierno digital.	38 portales gubernamentales.	La gobernanza de datos fortalece la seguridad digital.	OG, OE2
18	Ribeiro, Fonte & Silva	2025	Portugal	Analizar la seguridad en servicios públicos digitales.	67 plataformas gubernamentales.	Persisten vulnerabilidades en servicios digitales públicos.	OE1
19	Yang & Yang	2025	China	Analizar alianzas estratégicas en gobierno digital.	82 asociaciones públicas y tecnológicas.	La gobernanza colaborativa mejora la seguridad institucional.	OG, OE2
20	Busetti & Scanni	2025	Italia	Evaluar mecanismos de reporte de incidentes de ciberseguridad.	110 reportes institucionales.	El reporte temprano mejora la respuesta frente a amenazas digitales.	OE2

Tabla 5 – Artículos seleccionados para el análisis

3.2. Riesgos de ciberseguridad identificados en el gobierno digital

Los resultados evidencian que los riesgos de ciberseguridad en gobiernos digitales pueden agruparse en tres categorías principales: riesgos tecnológicos, riesgos humanos y riesgos de infraestructura. Los riesgos tecnológicos fueron los más recurrentes en la literatura analizada, seguidos por los riesgos humanos y los riesgos de infraestructura.

Los riesgos tecnológicos incluyen software desactualizado, vulnerabilidades en portales institucionales, accesos no autorizados y deficiencias en sistemas de cifrado. En este sentido, Chen & Li (2024) reportaron la persistencia de vulnerabilidades críticas en plataformas gubernamentales digitales, mientras que Altamirano-de-la-Borda (2021) identificó debilidades en los estándares de protección de datos en instituciones públicas peruanas.

Los riesgos humanos estuvieron asociados principalmente con la limitada capacitación del personal y la exposición a ataques de ingeniería social. Pethig, Kroenung & Nöltner (2021) señalaron que la desconfianza en la seguridad digital reduce la adopción de servicios de gobierno electrónico, evidenciando la importancia de fortalecer la cultura de ciberseguridad en los usuarios.

Los riesgos de infraestructura estuvieron vinculados con limitaciones tecnológicas, deficiencias en conectividad y ausencia de mecanismos de monitoreo continuo. León-Gutiérrez et al. (2022) identificaron brechas en infraestructura y capacitación tecnológica dentro del gobierno digital peruano, como se observa en la Tabla 6, la clasificación de riesgos se realizó mediante un análisis temático de los estudios seleccionados, agrupando los hallazgos según la naturaleza predominante de las amenazas identificadas en la literatura científica.

Tipo de riesgo	Frecuencia	Porcentaje
<i>Tecnológicos</i>	8	40 %
<i>Humanos</i>	7	35 %
<i>Infraestructura</i>	5	25 %
<i>Total</i>	20	100 %

Tabla 6 – Clasificación de riesgos identificados

3.3.Estrategias de mitigación frente a riesgos de ciberseguridad

La evidencia revisada permitió identificar tres estrategias predominantes para reducir los riesgos de ciberseguridad en el gobierno digital: fortalecimiento del talento humano, modernización tecnológica y monitoreo continuo.

En relación con el fortalecimiento del talento humano y las capacidades organizacionales, diversos estudios destacan la importancia de contar con personal preparado para gestionar riesgos de ciberseguridad. Norris y Mateczun (2023), mediante el análisis de 14 gobiernos locales de Estados Unidos (3 condados y 11 ciudades) representados por responsables de tecnología de información y seguridad informática, identificaron desafíos relacionados con la preparación institucional, las capacidades internas y los mecanismos de respuesta frente a incidentes cibernéticos. Estos resultados evidencian que la disponibilidad de personal especializado y estructuras organizacionales adecuadas constituye un elemento fundamental para fortalecer la gestión de la ciberseguridad en gobiernos digitales.

Respecto a la modernización tecnológica, se observó la implementación de autenticación multifactor, cifrado de datos y actualización de sistemas. En Italia, Rossi y Bianchi (2022) señalaron que el 68 % de las municipalidades incorporó mecanismos de cifrado para proteger la información institucional.

En cuanto al monitoreo continuo, Kim y Park (2022) evidenciaron que el seguimiento permanente de amenazas digitales permitió reducir incidentes en el 68 % de las agencias gubernamentales analizadas en Corea del Sur.

Según la Tabla 7, las estrategias de mitigación fueron clasificadas según la frecuencia de aparición en los estudios analizados.

Estrategia	Frecuencia	Porcentaje
Modernización tecnológica	8	40 %
Capacitación del personal	7	35 %
Monitoreo continuo	5	25 %
Total	20	100 %

Fuente: Elaboración propia.

Tabla 7 –Estrategias de mitigación identificadas

3.4.Relación entre ciberseguridad y confianza ciudadana

Los estudios analizados coinciden en que la implementación de políticas de ciberseguridad fortalece la confianza ciudadana en los servicios digitales del Estado. Pethig, Kroenung & Nöltner (2021) evidenciaron que la percepción de inseguridad reduce la adopción de plataformas de gobierno digital. Asimismo, Sadik-Zada, Aliyev & Mammadov (2024) concluyeron que la integración entre digitalización y ciberseguridad contribuye a reducir el fraude administrativo y fortalecer la legitimidad institucional. De manera similar, Henman, Graham & Lata (2021) sostienen que una adecuada gobernanza de datos fortalece la seguridad digital y mejora la relación entre Estado y ciudadanía.

4. Discusión

Los resultados obtenidos en la presente revisión sistemática confirman que la ciberseguridad constituye un componente esencial para la consolidación del Gobierno Digital, debido a que garantiza la protección de la información pública, la continuidad de los servicios electrónicos y la confianza de la ciudadanía en las plataformas estatales. La evidencia analizada demuestra que la transformación digital de las instituciones públicas ha generado avances significativos en la eficiencia administrativa; sin embargo, también ha incrementado la exposición a amenazas que comprometen la integridad, disponibilidad y confidencialidad de los datos gubernamentales. Estos hallazgos son consistentes con el NIST Cybersecurity Framework (NIST, 2018), particularmente con las funciones Identify y Protect, las cuales destacan la importancia de identificar los activos críticos, evaluar las vulnerabilidades e implementar controles de seguridad que reduzcan la probabilidad de incidentes cibernéticos desde las etapas iniciales de la gestión del riesgo. Asimismo, desde la perspectiva de la norma ISO/IEC 27005:2018, los riesgos identificados evidencian la necesidad de desarrollar procesos sistemáticos de identificación, análisis, evaluación y tratamiento de riesgos como parte de una gestión integral de la seguridad de la información en las instituciones públicas. En esta línea, Pethig, Kroenung y Nöltner (2021) sostienen que la percepción de inseguridad digital reduce la disposición de los ciudadanos a utilizar plataformas de gobierno electrónico, mientras que Sadik-Zada, Aliyev y Mammadov (2024) afirman que la ciberseguridad

fortalece la legitimidad institucional y contribuye a reducir prácticas de fraude administrativo.

En relación con el objetivo general del estudio, orientado a identificar los riesgos de ciberseguridad presentes en los gobiernos digitales y las estrategias implementadas para su mitigación, se observa que la ciberseguridad debe comprenderse no solo como una dimensión tecnológica, sino como un elemento estratégico dentro de la gestión pública contemporánea. En este sentido, Caldarulo, Welch y Feeney (2022) señalan que la planificación institucional y la capacitación fortalecen la resiliencia digital en gobiernos locales, mientras que Norris y Mateczun (2023), a partir del análisis de 14 gobiernos locales estadounidenses (3 condados y 11 ciudades), evidenciaron que la capacidad de respuesta frente a incidentes cibernéticos depende de la preparación organizacional y de la disponibilidad de capacidades internas especializadas. Asimismo, Hassan y Duarte (2024) sostienen que la evolución de la ciberseguridad en la administración pública requiere mecanismos permanentes de actualización tecnológica y fortalecimiento organizacional. De manera complementaria, Yang y Yang (2025) destacan que las alianzas estratégicas entre instituciones públicas y sectores tecnológicos favorecen la gobernanza colaborativa y fortalecen la seguridad institucional. Estos resultados evidencian que la gestión de la ciberseguridad debe abordarse como un proceso continuo de gestión del riesgo y mejora organizacional, en concordancia con los principios establecidos por la norma ISO/IEC 27005:2018, la cual promueve la identificación, evaluación, tratamiento y monitoreo permanente de los riesgos para fortalecer la resiliencia institucional y garantizar la continuidad de los servicios públicos digitales.

Dentro del contexto peruano, los hallazgos evidencian que las instituciones públicas enfrentan desafíos similares a los observados en otros países. La identificación de brechas de seguridad en plataformas institucionales, la persistencia de ataques de phishing y las limitaciones en infraestructura tecnológica muestran que la digitalización del sector público continúa avanzando en un entorno de vulnerabilidad operativa. León-Gutiérrez et al. (2022) identificaron importantes deficiencias relacionadas con infraestructura tecnológica y capacitación del personal en entidades públicas peruanas. De manera complementaria, Cornejo (2025) señala que el fortalecimiento de la madurez en ciberseguridad y el desarrollo de capacidades organizacionales constituyen factores determinantes para incrementar la resiliencia digital frente a amenazas emergentes. Esta situación coincide con lo señalado por Ribeiro, Fonte y Silva (2025), quienes reportaron vulnerabilidades persistentes en servicios digitales públicos de Portugal, demostrando que incluso países con mayores niveles de digitalización continúan enfrentando riesgos estructurales asociados con la seguridad informática. En conjunto, estos resultados ponen de manifiesto que la transformación digital no garantiza, por sí sola, un entorno seguro para la prestación de servicios públicos. La efectividad del Gobierno Digital depende de que las iniciativas de modernización tecnológica estén acompañadas por una gestión integral de riesgos, políticas institucionales de ciberseguridad y procesos permanentes de fortalecimiento de capacidades organizacionales.

Respecto al primer objetivo específico, relacionado con la clasificación de los riesgos de ciberseguridad más frecuentes, la evidencia permitió identificar tres dimensiones principales: riesgos tecnológicos, riesgos humanos y riesgos de infraestructura. Los riesgos tecnológicos se relacionan con software obsoleto, debilidades en los sistemas

de autenticación y fallas en el cifrado de información. Por su parte, los riesgos humanos se asocian con la limitada formación del personal y con errores derivados del desconocimiento de buenas prácticas digitales. Finalmente, los riesgos de infraestructura reflejan deficiencias en conectividad, equipamiento y capacidad tecnológica institucional. Esta clasificación confirma que la seguridad digital en el sector público depende de la interacción entre factores técnicos, organizacionales y humanos. Desde una perspectiva de gestión del riesgo, estos tres grupos de amenazas no deben abordarse de manera independiente, sino como componentes interrelacionados de un mismo sistema de seguridad. En este sentido, el NIST Cybersecurity Framework (NIST, 2018) plantea que la identificación de activos críticos, la protección de los sistemas, la detección temprana de incidentes, la respuesta coordinada y la recuperación de los servicios constituyen funciones complementarias que permiten fortalecer la resiliencia institucional frente a amenazas cada vez más complejas. Los resultados de esta revisión evidencian que las instituciones públicas que integran estas funciones de manera articulada presentan una mayor capacidad para prevenir incidentes y reducir su impacto sobre la continuidad de los servicios digitales.

Los hallazgos de Chen y Li (2024) coinciden con Altamirano-de-la-Borda (2021) al señalar que las vulnerabilidades tecnológicas representan una de las principales amenazas para los gobiernos digitales. Sin embargo, mientras Chen y Li enfatizan las deficiencias estructurales de las plataformas gubernamentales digitales en China, Altamirano-de-la-Borda centra su análisis en las limitaciones normativas y organizacionales presentes en las instituciones públicas peruanas. Estas diferencias sugieren que la naturaleza de los riesgos de ciberseguridad está condicionada por el nivel de madurez digital, el desarrollo institucional y la capacidad de gobernanza tecnológica de cada país. En consecuencia, las estrategias de mitigación no deberían implementarse mediante modelos estandarizados, sino adaptarse al contexto organizacional, regulatorio y tecnológico de cada administración pública. Asimismo, Pethig, Kroenung y Nöltner (2021) sostienen que la desconfianza ciudadana en la seguridad de los sistemas digitales limita la adopción de servicios electrónicos, lo que demuestra que las consecuencias de las vulnerabilidades no solo afectan el funcionamiento tecnológico, sino también la relación entre ciudadanía y Estado. En este sentido, la gestión de la ciberseguridad debe concebirse como un factor estratégico para fortalecer tanto la resiliencia institucional como la confianza ciudadana en los servicios públicos digitales.

En relación con el segundo objetivo específico, enfocado en analizar las estrategias implementadas para mitigar los riesgos identificados, los estudios revisados muestran que las medidas más efectivas se concentran en tres líneas de acción complementarias: fortalecimiento del talento humano, modernización tecnológica y monitoreo continuo. La primera corresponde al fortalecimiento del talento humano mediante programas permanentes de capacitación en seguridad digital. En este aspecto, Ahn y Chen (2022) evidenciaron que la formación tecnológica mejora el uso seguro de herramientas digitales en funcionarios públicos, mientras que Martínez y Gómez (2023) concluyeron que la capacitación institucional fortalece la protección digital en gobiernos municipales españoles. De manera similar, Hassan y Duarte (2024) destacan que la capacitación continua constituye uno de los principales mecanismos para reducir incidentes asociados con errores humanos. Estos hallazgos evidencian que el factor humano continúa siendo

uno de los principales elementos de vulnerabilidad dentro de las organizaciones públicas, por lo que la capacitación debe entenderse como una estrategia preventiva de carácter permanente y no como una acción aislada. Desde esta perspectiva, el fortalecimiento de las competencias digitales del personal incrementa la capacidad institucional para identificar amenazas, aplicar buenas prácticas de seguridad y responder de manera oportuna ante incidentes cibernéticos.

La segunda línea de acción se relaciona con la modernización tecnológica mediante el uso de autenticación multifactor, cifrado de datos y actualización de sistemas institucionales. Rossi y Bianchi (2022) señalaron que la implementación de mecanismos de cifrado mejoró significativamente la protección de la información en municipalidades italianas, mientras que Frandell y Feeney (2022) evidenciaron que la integración de factores tecnológicos y organizacionales fortalece la capacidad institucional frente a incidentes cibernéticos en gobiernos locales. Asimismo, Chang y Huang (2023) destacan que el intercambio de información sobre amenazas cibernéticas fortalece la resiliencia institucional y mejora la capacidad de respuesta frente a incidentes. En conjunto, estas medidas reflejan un enfoque preventivo orientado a reducir la superficie de exposición frente a amenazas cibernéticas y fortalecer la resiliencia de las plataformas gubernamentales. Su efectividad aumenta cuando se implementan de manera integrada con procesos de gestión del riesgo y monitoreo continuo, permitiendo a las instituciones públicas anticipar vulnerabilidades y responder de forma más eficiente ante incidentes de seguridad.

La tercera estrategia identificada corresponde al monitoreo continuo de amenazas digitales mediante sistemas de supervisión permanente y mecanismos de alerta temprana. Kim y Park (2022) evidenciaron que el monitoreo constante permitió reducir incidentes cibernéticos en agencias gubernamentales de Corea del Sur, mientras que Oliveira y Santos (2023) concluyeron que la supervisión continua de redes digitales disminuye los riesgos tecnológicos en gobiernos municipales brasileños. De manera complementaria, Busetti y Scanni (2025) sostienen que los sistemas de reporte temprano fortalecen la capacidad institucional de respuesta ante amenazas de ciberseguridad. Estos resultados ponen de manifiesto que el monitoreo continuo constituye un elemento esencial para una gestión proactiva de la ciberseguridad, al facilitar la detección temprana de vulnerabilidades, la respuesta oportuna frente a incidentes y la mejora continua de los controles de seguridad. En consecuencia, las instituciones públicas que incorporan mecanismos permanentes de supervisión fortalecen su capacidad de anticipación, reducen el impacto de los ataques cibernéticos y aumentan la continuidad operativa de los servicios digitales.

Asimismo, la evidencia indica que la implementación de estrategias de ciberseguridad genera efectos positivos que trascienden la protección de los activos tecnológicos, al fortalecer la gobernanza digital, la transparencia institucional y la confianza ciudadana en los servicios públicos. Henman, Graham y Lata (2021) sostienen que una adecuada gobernanza de datos mejora la seguridad digital y favorece la interacción entre el Estado y la ciudadanía. En la misma línea, Sadik-Zada, Aliyev y Mammadov (2024) concluyen que la integración entre digitalización y ciberseguridad contribuye al fortalecimiento de la legitimidad institucional y a la reducción del fraude administrativo. A partir de la evidencia sintetizada en esta revisión, se propone un marco integrador para la gestión

de la ciberseguridad en el Gobierno Digital, conformado por cuatro componentes interdependientes: (i) gestión continua de riesgos, orientada a la identificación, evaluación y tratamiento de amenazas; (ii) fortalecimiento del talento humano mediante programas permanentes de capacitación y cultura de ciberseguridad; (iii) modernización tecnológica basada en mecanismos de autenticación, cifrado y monitoreo continuo; y (iv) gobernanza institucional sustentada en políticas, coordinación interinstitucional y mejora continua. La integración de estos componentes permite fortalecer la resiliencia organizacional, garantizar la continuidad de los servicios públicos digitales y consolidar la confianza de la ciudadanía en los procesos de transformación digital del Estado.

Desde una perspectiva comparativa, los resultados muestran que los países con mayores niveles de madurez digital presentan mejores resultados cuando la ciberseguridad es asumida como una política pública permanente y no como una respuesta reactiva frente a incidentes aislados. La evidencia revisada demuestra que las instituciones con mejores capacidades de prevención integran simultáneamente inversión tecnológica, capacitación continua y mecanismos colaborativos de gobernanza digital. En consecuencia, la sostenibilidad del Gobierno Digital depende de la articulación entre infraestructura tecnológica segura, capital humano especializado y compromiso institucional orientado a la prevención continua de riesgos. Desde una perspectiva práctica, estos hallazgos sugieren que las instituciones públicas deberían priorizar el desarrollo de programas permanentes de gestión del riesgo, fortalecer la formación continua del personal, establecer mecanismos de monitoreo y respuesta ante incidentes, y consolidar políticas de gobernanza digital que integren la ciberseguridad como un eje transversal de la transformación digital. Estas acciones contribuyen no solo a reducir la exposición frente a amenazas cibernéticas, sino también a mejorar la continuidad operativa, la calidad de los servicios públicos digitales y la confianza de la ciudadanía.

En síntesis, la evidencia analizada permite sostener que la ciberseguridad se ha convertido en un pilar estratégico del Gobierno Digital, al contribuir simultáneamente a la protección de la información pública, al fortalecimiento de la confianza ciudadana y a la mejora de la eficiencia administrativa. La principal contribución de esta revisión sistemática consiste en integrar la evidencia científica reciente sobre los riesgos de ciberseguridad y las estrategias de mitigación en un marco analítico que relaciona la gestión continua del riesgo, el fortalecimiento del talento humano, la modernización tecnológica y la gobernanza institucional como componentes esenciales para la consolidación del Gobierno Digital. En este contexto, los hallazgos obtenidos proporcionan una base útil para orientar la formulación de políticas públicas, fortalecer la toma de decisiones en las instituciones gubernamentales y servir como referencia para futuras investigaciones sobre ciberseguridad en el sector público.

5. Conclusiones

La presente revisión sistemática permitió confirmar que la ciberseguridad constituye un componente estratégico para la consolidación del Gobierno Digital, al contribuir a la protección de la información pública, la continuidad de los servicios digitales y el fortalecimiento de la confianza ciudadana. La evidencia analizada demuestra que la transformación digital del Estado no depende exclusivamente de la incorporación de nuevas tecnologías, sino también de la implementación de políticas integrales de gestión

del riesgo, mecanismos de gobernanza y estrategias permanentes de protección de la información.

En relación con el objetivo general, se concluye que la ciberseguridad debe integrarse como una política transversal dentro de los procesos de transformación digital del sector público. Su efectividad depende de la articulación entre infraestructura tecnológica segura, gestión institucional, talento humano capacitado y mecanismos permanentes de monitoreo y mejora continua. La ausencia de cualquiera de estos componentes incrementa la exposición frente a amenazas cibernéticas y compromete la continuidad de los servicios públicos digitales.

Respecto al primer objetivo específico, la evidencia permitió establecer que los riesgos de ciberseguridad más frecuentes en el Gobierno Digital pueden agruparse en riesgos tecnológicos, humanos y de infraestructura. Entre ellos destacan los ataques de phishing, los accesos no autorizados, el uso de software desactualizado, las deficiencias en el cifrado de datos y las limitaciones en la infraestructura tecnológica. Asimismo, se identificó que una proporción importante de estos riesgos está asociada con factores humanos, particularmente con la insuficiente capacitación del personal y la limitada cultura organizacional orientada a la ciberseguridad.

En relación con el segundo objetivo específico, enfocado en analizar las estrategias implementadas para mitigar estos riesgos, la evidencia mostró que las medidas más efectivas integran acciones técnicas y organizacionales. Entre las estrategias más relevantes destacan la capacitación continua del personal, la implementación de autenticación multifactor, el uso de cifrado de información, el monitoreo permanente de amenazas y el fortalecimiento de políticas institucionales de gobernanza digital. La implementación coordinada de estas medidas fortalece la resiliencia institucional y favorece una gestión proactiva de los riesgos de ciberseguridad en el sector público.

Los resultados también permitieron identificar que los países con mayores niveles de madurez digital presentan mejores resultados cuando la ciberseguridad es considerada una estrategia permanente y no una respuesta aislada frente a incidentes. En este sentido, las experiencias internacionales muestran que la incorporación de tecnologías emergentes puede fortalecer la resiliencia institucional, siempre que su implementación se adapte al contexto administrativo y normativo de cada país.

Como principal aporte de esta revisión sistemática, se propone una visión integradora de la ciberseguridad basada en cuatro componentes complementarios: gestión continua del riesgo, fortalecimiento del talento humano, modernización tecnológica y gobernanza institucional. La articulación de estos componentes proporciona una base para orientar la formulación de políticas públicas y fortalecer los procesos de transformación digital en las instituciones del sector público.

Entre las limitaciones del estudio, se reconoce que la revisión se centró en investigaciones publicadas entre 2020 y 2025 en bases de datos seleccionadas, por lo que algunos estudios recientes o literatura gris podrían no haber sido considerados. Futuras investigaciones podrían ampliar el periodo de análisis e incorporar estudios comparativos entre regiones para profundizar en la evolución de la ciberseguridad en el Gobierno Digital. Asimismo,

se recomienda desarrollar investigaciones empíricas que permitan validar el marco integrador propuesto en distintos contextos institucionales y niveles de gobierno.

En consecuencia, la consolidación del Gobierno Digital dependerá de la capacidad de los Estados para incorporar la ciberseguridad como un componente estratégico de sus políticas de transformación digital. Los hallazgos de esta revisión ofrecen una base de referencia para la formulación de estrategias institucionales orientadas a fortalecer la resiliencia organizacional, proteger los servicios públicos digitales y promover una transformación digital segura y sostenible.

Referencias

- Ahn, M., & Chen, Y. C. (2022). *Artificial intelligence and digital transformation in public administration*. *Public Performance & Management Review*, 45(4), 789–810. <https://doi.org/10.1080/15309576.2022.2041396>
- Altamirano-de-la-Borda, K. J. (2021). *La seguridad de la información en la administración pública*. En *Actas del Congreso Internacional de Ingeniería de Sistemas 2020* (pp. 77–95). Universidad de Lima. <https://doi.org/10.26439/ciis2020.5480>
- Busetti, S., & Scanni, G. (2025). *Cybersecurity incident reporting in digital public administration*. *Government Information Quarterly*, 42(1), 101912. <https://doi.org/10.1016/j.giq.2024.101912>
- Caldarulo, M., Welch, E. W., & Feeney, M. K. (2022). *Cybersecurity capacity in local governments*. *Government Information Quarterly*, 39(3), 101688. <https://doi.org/10.1016/j.giq.2022.101688>
- Castillo, A., Bruzza, M., & Tupia, M. (2021). *Estado del arte sobre la identificación de amenazas no intencionales de ciberseguridad de parte de personal interno en instituciones públicas*. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, (E41), 70–82.
- Chang, Y., & Huang, K. (2023). *Cybersecurity information-sharing networks and digital resilience*. *Information Polity*, 28(2), 177–193. <https://doi.org/10.3233/IP-220038>
- Chen, Y., & Li, X. (2024). *Cybersecurity risks in Chinese digital government platforms*. *Telecommunications Policy*, 48(2), 102671. <https://doi.org/10.1016/j.telpol.2023.102671>
- Cornejo, H. (2025). *Madurez en ciberseguridad y resiliencia digital en PYMES iberoamericanas: Diagnóstico y desafíos estratégicos*. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, (60), 127–141. <https://doi.org/10.17013/risti.60.127-141>
- European Union Agency for Cybersecurity. (2021). *ENISA Threat Landscape 2021*. <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2021>

- European Union Agency for Cybersecurity. (2020). *Cybersecurity risk management: Guidance for SMEs*. <https://www.enisa.europa.eu>
- Frاندell, A., & Feeney, M. (2022). *Cybersecurity threats in local government: A sociotechnical perspective*. *The American Review of Public Administration*, 52(8), 558–572. <https://doi.org/10.1177/02750740221125432>
- Hassan, W., & Duarte, A. E. (2024). *Cybersecurity evolution in public administration systems*. *Journal of Information Security and Applications*, 78, 103602. <https://doi.org/10.1016/j.jisa.2023.103602>
- Henman, P., Graham, T., & Lata, L. (2021). *Digital government and data governance challenges*. *Information Polity*, 26(4), 455–470. <https://doi.org/10.3233/IP-200280>
- Instituto Nacional de Estadística y Geografía. (2021). *Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares (ENDUTIH) 2021*. <https://www.inegi.org.mx/programas/dutih/2021/>
- International Organization for Standardization. (2018). *ISO/IEC 27005:2018 Information technology—Security techniques—Information security risk management*. <https://www.iso.org>
- Kim, H., & Park, J. (2022). *Digital risk mitigation strategies in public agencies*. *Government Information Quarterly*, 39(4), 101720. <https://doi.org/10.1016/j.giq.2022.101720>
- León-Gutiérrez, E. D., Tesillo-Gómez, C. M., Escobar-Arcaya, Y. A., & Godoy-Montoya, L. A. (2022). *Revisión de los avances y cambios en ciberseguridad en el Perú, para una transformación digital*. *Innovación y Software*, 3(2), 109–120. <https://doi.org/10.48168/innosoft.s9.a62>
- Martínez, P., & Gómez, A. (2023). *Riesgos y estrategias digitales en municipios españoles*. *Revista Española de Ciencia Política*, 61, 145–163. <https://doi.org/10.21308/recp.61.06>
- National Institute of Standards and Technology. (2018). *Framework for Improving Critical Infrastructure Cybersecurity (Version 1.1)*. <https://doi.org/10.6028/NIST.CSWP.04162018>
- Norris, D. F., & Mateczun, L. K. (2023). *Cyberattacks on local governments 2020: Findings from a key informant survey*. *Journal of Cyber Policy*, 8(3), 294–317. <https://doi.org/10.1080/23738871.2023.2178319>
- Olaya Moncada, D. P. (2024). *Impacto de las tecnologías emergentes aplicadas a la gerencia de los procesos logísticos y redes de valor: Una revisión bibliográfica hacia la cadena farmacéutica* (Trabajo de especialización, Universidad Nacional Abierta y a Distancia). Repositorio Institucional UNAD. <https://repository.unad.edu.co/bitstream/handle/10596/62941/dpolayam.pdf>

- Oliveira, P., & Santos, R. (2023). *Segurança digital em governos municipais brasileiros*. Revista de Administração Pública, 57(5), 1–18. <https://doi.org/10.1590/0034-761220220155>
- Organisation for Economic Co-operation and Development. (2022). *Digital Government Review*. <https://www.oecd.org/digital>
- Presidencia del Consejo de Ministros. (2020). *Política Nacional de Transformación Digital*. Gobierno del Perú. <https://www.gob.pe/institucion/pcm>
- Pethig, F., Kroenung, J., & Nöltner, K. (2021). *Understanding citizens' avoidance of e-government services*. *Government Information Quarterly*, 38(4), 101589. <https://doi.org/10.1016/j.giq.2021.101589>
- Ribeiro, L., Fonte, A., & Silva, P. (2025). *Cybersecurity challenges in digital public services*. *International Journal of Information Management Data Insights*, 5(1), 100289. <https://doi.org/10.1016/j.jjime.2024.100289>
- Rossi, G., & Bianchi, L. (2022). *Security in e-government platforms: Evidence from Italy*. *Information Systems Frontiers*, 24(6), 1845–1860. <https://doi.org/10.1007/s10796-021-10145-5>
- Sadik-Zada, E. R., Aliyev, T., & Mammadov, R. (2024). *Digital government and fraud mitigation in emerging economies*. *Digital Policy, Regulation and Governance*, 26(3), 250–268. <https://doi.org/10.1108/DPRG-01-2024-0008>
- Yang, L., & Yang, K. (2025). *Digital governance alliances and institutional cybersecurity*. *Government Information Quarterly*, 42(2), 101955. <https://doi.org/10.1016/j.giq.2025.101955>
- Yepes-Nuñez, J. J. (2021). *PRISMA 2020: actualización de la guía para revisiones sistemáticas*. *Revista Panamericana de Salud Pública*, 45, e83. <https://doi.org/10.26633/RPSP.2021.83>

Propuesta de un modelo de madurez para evaluar la ciberresiliencia de las pymes frente al ransomware

Domingo Ruiz Arroyo¹

domingoruizarroyo@gmail.com

¹ Grado en Ingeniería Informática, Universitat Oberta de Catalunya – UOC, Barcelona, España

DOI: 10.17013/risti.62.86–99

Resumen: Este artículo propone un modelo de madurez para evaluar la ciberresiliencia de las pymes frente al ransomware desde una perspectiva integrada de seguridad de la información, operación y continuidad de negocio. Mediante revisión de literatura y análisis comparado de ISO/IEC 27001, ISO/IEC 27002, NIST Cybersecurity Framework, Esquema Nacional de Seguridad, ENISA e INCIBE, se estructuran diez dimensiones y cinco niveles de madurez. La propuesta incorpora una lógica de autoevaluación basada en evidencias, pensada para organizaciones con recursos limitados. Su finalidad es identificar brechas, priorizar mejoras y orientar decisiones de ciberresiliencia sin presentar resultados empíricos ni sustituir auditorías normativas.

Palabras-clave: ciberresiliencia; ransomware; pymes; modelo de madurez; continuidad de negocio.

Proposal for a Maturity Model to Assess SME Cyber Resilience Against Ransomware

Abstract: This article proposes a maturity model for assessing SME cyber resilience against ransomware from an integrated perspective that combines information security, operations, and business continuity. Through a literature review and comparative analysis of ISO/IEC 27001, ISO/IEC 27002, the NIST Cybersecurity Framework, the Spanish National Security Scheme, ENISA, and INCIBE guidance, the model structures ten dimensions and five maturity levels. It includes an evidence-based self-assessment approach for organizations with limited resources. Its purpose is to identify gaps, prioritize improvements, and support cyber resilience decisions without presenting empirical results or replacing formal audits.

Keywords: cyber resilience; ransomware; SMEs; maturity model; business continuity.

1. Introducción

El ransomware ha pasado de ser una amenaza técnica asociada a estaciones de trabajo aisladas a convertirse en un riesgo de interrupción operativa para organizaciones de

cualquier tamaño. Su impacto no se limita al cifrado de información, sino que puede afectar a la disponibilidad de sistemas, la continuidad de procesos críticos, la relación con clientes y proveedores, la reputación y la capacidad de recuperación de la organización. Esta característica obliga a tratarlo desde una perspectiva de ciberresiliencia, entendida como la capacidad de anticipar, resistir, responder y recuperarse de incidentes que comprometen los sistemas de información y los procesos de negocio (Barker et al., 2022; Dupont et al., 2023; Lella et al., 2022).

Las pymes afrontan este problema con una dificultad adicional. Aunque dependen cada vez más de servicios digitales, plataformas en la nube, correo electrónico, sistemas de gestión, herramientas de colaboración y conexiones remotas, suelen disponer de recursos limitados para implantar controles avanzados de seguridad. En este tipo de organizaciones, la protección frente al ransomware no puede plantearse como una acumulación de herramientas técnicas, sino como un conjunto progresivo de capacidades de gobierno, gestión de activos, control de accesos, copias de seguridad, detección, respuesta, recuperación y formación del personal. En línea con trabajos recientes publicados en RISTI sobre integración tecnológica, la adopción efectiva de capacidades digitales depende también de formación, recursos y barreras organizativas (Cuenca Rodríguez & Serrano Betored, 2026; Eliot, 2024; European Union Agency for Cybersecurity [ENISA], 2021).

Los marcos de referencia existentes ofrecen una base amplia y sólida. ISO/IEC 27001 e ISO/IEC 27002 proporcionan criterios para establecer y mantener un sistema de gestión de seguridad de la información y controles asociados; ISO/IEC 27005 y NIST SP 800-30 orientan la gestión de riesgos; ISO 22301 y NIST SP 800-34 estructuran la continuidad de negocio; el NIST Cybersecurity Framework 2.0 facilita una taxonomía de resultados de ciberseguridad aplicable a organizaciones de distinto tamaño; y el Esquema Nacional de Seguridad recoge principios y medidas relevantes para sistemas de información en el contexto español (International Organization for Standardization, 2019; International Organization for Standardization & International Electrotechnical Commission, 2022a, 2022b, 2022c; Ministerio de Asuntos Económicos y Transformación Digital, 2022; National Institute of Standards and Technology, 2012; National Institute of Standards and Technology, 2024; Swanson et al., 2010).

Sin embargo, estos marcos no siempre son directamente operables para una pyme que necesita saber por dónde empezar, qué evidencias mínimas debe reunir y cómo priorizar una mejora gradual. La literatura sobre modelos de madurez permite ordenar esa transición mediante niveles discretos, comparables y orientados a la mejora continua (Becker et al., 2009; de Bruin et al., 2005). A partir de esta lógica, el presente artículo propone un modelo de madurez específico para evaluar la ciberresiliencia de las pymes frente al ransomware.

La pregunta de investigación que guía el trabajo es la siguiente: ¿cómo puede estructurarse un modelo de madurez que permita a las pymes evaluar de forma práctica su ciberresiliencia frente a ataques de ransomware, integrando controles técnicos, organizativos y de continuidad de negocio? El objetivo general es diseñar una propuesta de modelo de madurez aplicable a pymes, sin pretensión de certificar cumplimiento normativo ni de presentar resultados empíricos. Los objetivos específicos son: analizar

marcos de referencia relevantes, identificar dimensiones críticas, definir niveles de madurez, proponer una lógica de autoevaluación y delimitar sus condiciones de uso.

2. Fundamentación teórica

2.1. Ciberresiliencia y ransomware en pymes

La ciberresiliencia amplía el enfoque tradicional de la seguridad de la información porque no se centra únicamente en prevenir incidentes. Incluye también la capacidad de mantener operaciones esenciales, contener daños, recuperar servicios y aprender de eventos adversos. En el caso del ransomware, esta ampliación resulta especialmente necesaria, ya que incluso controles preventivos razonables pueden fallar ante errores humanos, credenciales comprometidas, vulnerabilidades no corregidas o dependencias de terceros.

El ransomware combina, normalmente, tres mecanismos de presión: cifrado de activos digitales, interrupción de procesos y, en determinados casos, amenaza de filtración de datos. Esto convierte el incidente en un problema técnico, jurídico, operativo y directivo. Por ello, un modelo útil para pymes debe incorporar controles de higiene digital, pero también decisiones sobre continuidad de negocio, tiempos de recuperación, responsabilidades internas, comunicación y priorización de procesos críticos (Barker et al., 2022; Ilca et al., 2023; Nelson et al., 2025; Souppaya & Scarfone, 2013).

Las recomendaciones específicas sobre ransomware coinciden en algunos puntos básicos: reducción de superficie de exposición, uso de autenticación multifactor en accesos críticos, aplicación del principio de mínimo privilegio, actualización de sistemas, protección antimalware, copias de seguridad aisladas y probadas, preparación de respuesta a incidentes y formación del personal (Barker et al., 2022; Instituto Nacional de Ciberseguridad, 2019, 2021). Estas prácticas son conocidas, pero su aplicación real depende de la capacidad de la organización para incorporarlas de forma estable a sus rutinas de gestión.

2.2. Modelos de madurez en seguridad y continuidad

Un modelo de madurez describe una evolución progresiva desde prácticas iniciales, reactivas o informales hasta prácticas definidas, medidas y optimizadas. Su utilidad no reside en producir una cifra aislada, sino en facilitar un diagnóstico estructurado, hacer visibles las brechas y ordenar las prioridades de mejora. En sistemas de información, los modelos de madurez se han utilizado para evaluar capacidades de gestión TI, seguridad, procesos, gobierno y transformación digital (Becker et al., 2009; de Bruin et al., 2005; Perozzo et al., 2022; Yigit Ozkan et al., 2020; Yigit Ozkan et al., 2021).

Para una pyme, la madurez debe formularse con criterios observables y comprensibles. Un nivel no debería asignarse solo por percepción del responsable de TI, sino por evidencias como políticas aprobadas, inventarios actualizados, registros de revisión de accesos, pruebas de restauración, procedimientos de incidentes, simulacros o indicadores de recuperación. Este enfoque reduce el riesgo de autoevaluaciones excesivamente optimistas y favorece conversaciones prácticas entre dirección, tecnología y responsables de procesos.

La escala propuesta en este artículo adopta cinco niveles: reactivo, básico, gestionado, medido y optimizado. Esta escala no debe interpretarse como una traducción oficial de los niveles del NIST CSF ni como una equivalencia directa con otros modelos. Se trata de una síntesis propia diseñada para ofrecer suficiente granularidad a pymes y para reflejar cinco grados de evolución: ausencia de control, control mínimo, formalización, medición y mejora continua.

2.3. Marcos de referencia utilizados

El modelo se apoya en una lectura integrada de marcos de seguridad, riesgo, continuidad y respuesta a incidentes. El NIST CSF 2.0 organiza la gestión de la ciberseguridad mediante funciones, categorías y subcategorías que permiten describir resultados esperados. NISTIR 8374 adapta esa lógica al ransomware y refuerza la importancia de identificar, proteger, detectar, responder y recuperar. La guía de inicio rápido para pequeñas empresas del NIST aporta una aproximación especialmente adecuada para organizaciones con capacidades limitadas (Barker et al., 2022; Eliot, 2024; National Institute of Standards and Technology, 2024).

ISO/IEC 27001, ISO/IEC 27002 e ISO/IEC 27005 aportan el marco de gestión, los controles de seguridad y la gestión del riesgo de seguridad de la información. ISO 22301, junto con NIST SP 800-34, incorpora la perspectiva de continuidad de negocio y recuperación frente a interrupciones. El ENS, por su parte, refuerza la necesidad de principios básicos, medidas de seguridad, categorización y gestión continua de la protección de los sistemas (International Organization for Standardization, 2019; International Organization for Standardization & International Electrotechnical Commission, 2022a, 2022b, 2022c; Ministerio de Asuntos Económicos y Transformación Digital, 2022; Swanson et al., 2010).

Finalmente, las guías de ENISA e INCIBE permiten aterrizar estas prácticas al lenguaje de las pymes. Su aportación es relevante porque traduce recomendaciones técnicas a medidas comprensibles sobre copias de seguridad, concienciación, respuesta ante ransomware, recuperación y decisión de no pagar rescates como solución operativa. El modelo propuesto toma estos marcos como referencias conceptuales, no como un sustituto de una auditoría normativa.

Marco o fuente	Aportación principal al modelo
ISO/IEC 27001 e ISO/IEC 27002	Sistema de gestión de seguridad de la información, políticas, controles, gestión de accesos, operación segura y mejora continua.
ISO/IEC 27005	Enfoque de gestión de riesgos de seguridad de la información y priorización según impacto y probabilidad.
ISO 22301 y NIST SP 800-34	Continuidad de negocio, análisis de impacto, planes de recuperación, pruebas y mantenimiento de capacidades.
NIST CSF 2.0 y NISTIR 8374	Funciones de ciberseguridad, perfiles de riesgo, prevención, respuesta y recuperación frente al ransomware.
ENS, ENISA e INCIBE	Principios, medidas y recomendaciones prácticas para seguridad, concienciación, backups, recuperación y gestión de incidentes.

Tabla 1 – Marcos de referencia considerados y aportación principal al modelo

3. Metodología de construcción del modelo

El artículo adopta una metodología de diseño conceptual y síntesis aplicada. No se ha realizado trabajo de campo, encuesta, entrevista ni análisis de datos empíricos. En consecuencia, el resultado no se presenta como una validación estadística, sino como una propuesta teórica/aplicada fundamentada en literatura académica, normas y guías técnicas. Esta delimitación es importante para evitar atribuir al modelo un alcance que todavía no posee.

La construcción se desarrolló en cuatro fases. En primer lugar, se identificaron marcos de referencia relacionados con seguridad de la información, gestión de riesgos, continuidad de negocio, respuesta a incidentes y ransomware. En segundo lugar, se extrajeron capacidades recurrentes que aparecían de manera transversal en esos marcos. En tercer lugar, dichas capacidades se agruparon en dimensiones comprensibles para una pyme. En cuarto lugar, se definieron niveles de madurez y evidencias mínimas que permiten orientar la autoevaluación.

Los criterios de selección de dimensiones fueron cuatro. La dimensión debía estar relacionada con el ciclo de vida del riesgo de ransomware; debía ser observable mediante evidencias razonables; debía poder aplicarse sin exigir estructuras propias de grandes empresas; y debía contribuir a una mejora gradual. Con estos criterios se descartaron dimensiones excesivamente técnicas o dependientes de herramientas concretas, y se priorizaron capacidades organizativas, operativas y de continuidad.

La definición de niveles se realizó siguiendo una progresión de formalización, cobertura, repetibilidad, medición y mejora continua. Así, el nivel 1 representa prácticas reactivas o inexistentes; el nivel 2, controles básicos pero parciales; el nivel 3, procesos documentados y mantenidos; el nivel 4, prácticas medidas y probadas; y el nivel 5, capacidades optimizadas e integradas en la gestión del negocio. Esta estructura permite que la pyme identifique tanto su punto de partida como el nivel objetivo razonable.

El mapeo entre marcos y dimensiones se realizó por correspondencia funcional: las capacidades de gobierno y riesgo se asociaron a ISO/IEC 27001, ISO/IEC 27005, NIST CSF y NIST SP 800-30; las de protección, copias y detección a ISO/IEC 27002, NISTIR 8374, ENISA e INCIBE; y las de respuesta, recuperación y continuidad a ISO 22301, NIST SP 800-34 y recomendaciones específicas sobre ransomware. Solo se incorporaron como dimensiones aquellas capacidades que aparecían de forma recurrente o que resultaban críticas para la recuperación operativa de una pyme.

4. Propuesta del modelo de madurez

4.1. Finalidad, alcance y principios de diseño

El modelo tiene como finalidad evaluar la preparación de una pyme frente al ransomware desde una perspectiva de ciberresiliencia. No pretende sustituir una auditoría de ISO/IEC 27001, una certificación ENS, un análisis técnico de vulnerabilidades ni una prueba de recuperación ante desastres. Su utilidad principal es diagnóstica: ayuda a identificar brechas, priorizar acciones y explicar a la dirección la relación entre seguridad, continuidad y riesgo operativo.

El alcance se limita a organizaciones pequeñas y medianas que disponen de sistemas de información propios o externalizados, utilizan servicios digitales para procesos de negocio y necesitan una herramienta de autoevaluación manejable. El modelo puede aplicarse a la organización completa o a una unidad concreta, siempre que se mantenga la trazabilidad de las evidencias revisadas y de los procesos incluidos.

El diseño se basa en cinco principios: simplicidad, para que pueda ser usado sin equipos especializados; evidencia, para evitar valoraciones puramente subjetivas; gradualidad, para facilitar planes de mejora realistas; integración, para conectar controles técnicos y continuidad de negocio; y neutralidad tecnológica, para no depender de un fabricante, producto o arquitectura concreta.

4.2. Dimensiones del modelo

El modelo se organiza en diez dimensiones. Las tres primeras se relacionan con el gobierno del riesgo: políticas, activos y accesos. Las cuatro siguientes se orientan a la protección y detección: copias de seguridad, protección frente a malware, segmentación y monitorización. Las tres últimas abordan la capacidad de respuesta y recuperación: gestión de incidentes, continuidad de negocio y formación del personal.

La lógica del modelo es acumulativa, aunque no estrictamente lineal. Una pyme puede tener un buen nivel en copias de seguridad y, al mismo tiempo, un nivel bajo en monitorización. Por ello, el resultado global debe interpretarse junto con el perfil por dimensión. En ransomware, una única dimensión débil puede condicionar la resiliencia completa; por ejemplo, copias no probadas o privilegios excesivos pueden anular parcialmente otras medidas de protección.

Nivel	Denominación	Descripción general
1	Reactivo	No existen prácticas formales o dependen de respuestas improvisadas ante problemas concretos. La organización actúa después del incidente.
2	Básico	Existen controles mínimos, pero son parciales, poco documentados y dependen de personas concretas. La cobertura no está asegurada.
3	Gestionado	Los procesos están documentados, asignados y mantenidos. Hay responsabilidades definidas y evidencias básicas de ejecución.
4	Medido	La organización prueba, revisa y mide sus capacidades. Existen indicadores, revisiones periódicas y acciones correctivas.
5	Optimizado	La ciberresiliencia está integrada en la gestión del negocio, se mejora de forma continua y se adapta a cambios tecnológicos y de amenaza.

Tabla 2 – Niveles generales de madurez propuestos

4.3. Criterios por dimensión

La Tabla 3 resume los criterios de madurez por dimensión. Para mantener la legibilidad, los niveles se agrupan en tres estados de avance: inicial, intermedio y avanzado. En una evaluación completa, cada dimensión debe asignarse a un nivel específico del 1 al 5 según la evidencia disponible. La tabla no agota todos los controles posibles, sino que define los criterios mínimos para una autoevaluación inicial.

Dimensión	Nivel 1-2: inicial	Nivel 3: gestionado	Nivel 4-5: avanzado
Gobierno y políticas de seguridad	No hay política formal o existe de forma genérica y no comunicada.	Política aprobada, responsabilidades definidas y comunicación interna básica.	Revisiones periódicas, indicadores y alineación con gestión de riesgos y dirección.
Gestión de activos	Inventario inexistente, parcial o mantenido manualmente sin criticidad.	Inventario actualizado de equipos, usuarios, servicios y activos críticos.	Inventario verificado, trazable, clasificado por criticidad y actualizado con cambios.
Gestión de accesos e identidades	Cuentas compartidas, privilegios excesivos o bajas no controladas.	Usuarios nominales, roles definidos, MFA en accesos críticos y revisiones básicas.	Mínimo privilegio sostenido, revisiones periódicas, registros y control de accesos externos.
Copias de seguridad y recuperación	Copias irregulares, no aisladas o sin pruebas de restauración.	Política de copias, cobertura de activos críticos y restauraciones básicas probadas.	Estrategia 3-2-1, pruebas planificadas, objetivos RTO/RPO y documentación de resultados.
Protección frente a malware y ransomware	Protección reactiva o dependiente de antivirus no gestionado.	Controles antimalware, actualización de sistemas y endurecimiento básico.	Defensa multicapa, bloqueo de macros o scripts de riesgo y revisión continua de exposición.
Segmentación de red y seguridad perimetral	Red plana, accesos remotos poco controlados o reglas no revisadas.	Segmentación funcional, cortafuegos gestionado y control de accesos remotos.	Segmentación según criticidad, reducción de superficie y revisión periódica de reglas.
Monitorización y detección	Sin registros centralizados o alertas dispersas.	Registros básicos de sistemas críticos y revisión definida.	Correlación de eventos, umbrales, seguimiento sistemático y mejora de reglas de detección.
Respuesta a incidentes	Respuesta improvisada y dependiente de conocimiento individual.	Plan documentado con roles, contactos, escalado y pasos ante ransomware.	Simulacros, lecciones aprendidas, tiempos medidos y coordinación con terceros.
Continuidad de negocio	No hay análisis de procesos críticos ni prioridades de recuperación.	Procesos críticos identificados y continuidad básica documentada.	BIA, prioridades de recuperación, pruebas de continuidad y coordinación con dirección.
Formación y concienciación	Formación inexistente o puntual sin seguimiento.	Plan anual de concienciación y comunicación de buenas prácticas.	Formación segmentada, simulaciones, medición de eficacia y cultura de reporte temprano.

Tabla 3 – Dimensiones del modelo y criterios de avance

4.4. Evidencias mínimas recomendadas

El modelo debe evaluarse con evidencias. Para una pyme, la evidencia no tiene que ser compleja, pero sí verificable. Un documento breve, un registro de restauración, una captura de configuración, una lista de usuarios revisada o un acta de simulacro pueden ser suficientes si reflejan una práctica real y repetible. La Tabla 4 propone ejemplos de evidencias mínimas por bloque de capacidades.

Bloque	Evidencias recomendadas
Gobierno y riesgo	Política de seguridad, responsable designado, registro de riesgos, revisión por dirección, lista de proveedores críticos.
Activos y accesos	Inventario de equipos y servicios, relación de usuarios, revisiones de privilegios, registro de altas y bajas, MFA en accesos críticos.
Protección y detección	Estado de parches, configuración antimalware, reglas de cortafuegos, registros revisados, alertas relevantes y acciones correctivas.
Copias y recuperación	Política de copias, alcance de datos protegidos, ubicación de copias, pruebas de restauración, RTO/RPO definidos para procesos críticos.
Respuesta y continuidad	Plan de respuesta a incidentes, contactos de emergencia, plan de comunicación, BIA, pruebas de continuidad y lecciones aprendidas.
Personas y cultura	Plan de formación, campañas de phishing simulado o ejercicios internos, comunicaciones al personal y canal de reporte de incidentes.

Tabla 4 – Evidencias mínimas para la autoevaluación

5. Sistema de evaluación y priorización

5.1. Unidad de evaluación y escala

La unidad básica de evaluación es la dimensión. Cada dimensión recibe un valor de 1 a 5 según el nivel más alto que pueda sostenerse con evidencias. La asignación debe ser conservadora: si una práctica existe solo de forma informal, no debería situarse en niveles gestionados o medidos. Cuando una dimensión no sea aplicable, la organización debe justificarlo, aunque en ransomware la mayoría de dimensiones suelen ser relevantes para cualquier pyme digitalizada.

El resultado global se calcula como la media aritmética de las dimensiones evaluadas. Para evitar una falsa sensación de seguridad, se recomienda acompañar la media con dos datos adicionales: el nivel mínimo obtenido en cualquier dimensión crítica y el número de dimensiones por debajo del nivel 3. Así, una organización con media aceptable pero copias no probadas o respuesta inexistente no debería considerarse suficientemente resiliente.

La clasificación orientativa puede definirse así: 1,00-1,49, reactivo; 1,50-2,49, básico; 2,50-3,49, gestionado; 3,50-4,49, medido; y 4,50-5,00, optimizado. Estos umbrales son analíticos, no empíricos. Su función es facilitar la interpretación interna y priorizar mejoras, no establecer una certificación externa ni comparar organizaciones sin contexto sectorial.

5.2. Reglas de puntuación propuestas

Para reforzar la coherencia de la evaluación se proponen tres reglas. La primera es la regla de evidencia: toda puntuación debe apoyarse en un documento, registro, configuración o prueba verificable. La segunda es la regla de cobertura: no basta con que la práctica exista en un sistema aislado si los activos críticos quedan fuera. La tercera es la regla de prueba: en copias, respuesta y continuidad, el nivel 4 requiere evidencias de ensayo, no solo documentación.

La pyme puede utilizar el modelo de dos maneras. En una evaluación rápida, puede asignar niveles mediante entrevista interna y revisión documental básica. En una evaluación ampliada, puede incorporar pruebas de restauración, simulacros de incidente, revisión de accesos y análisis de procesos críticos. En ambos casos, el resultado debe convertirse en un plan de mejora con acciones, responsables, fecha objetivo y evidencia esperada.

La priorización no debe basarse solo en el nivel más bajo. Debe considerar la criticidad del proceso afectado, la exposición al exterior, la dependencia de proveedores, la existencia de datos sensibles y el impacto de una parada. En ransomware, las mejoras con mayor retorno inicial suelen concentrarse en MFA, revisión de privilegios, copias probadas, actualización de sistemas, formación y plan de respuesta.

6. Aplicabilidad en pymes

La principal aportación práctica del modelo es que permite iniciar conversaciones de seguridad con lenguaje de negocio. En lugar de presentar una lista extensa de controles, ordena la discusión en diez dimensiones y cinco niveles. Esto facilita que la dirección entienda que la protección frente al ransomware no depende solo de comprar tecnología, sino de mantener capacidades mínimas y probadas.

En pymes sin departamento de seguridad, el modelo puede ser utilizado por el responsable de informática, un proveedor gestionado, la dirección o un equipo mixto. La evaluación puede completarse de forma razonable con entrevistas internas, revisión de configuración, inventarios, registros de backup y documentación de continuidad. El resultado no debe verse como una calificación estática, sino como una fotografía para definir el siguiente ciclo de mejora.

La utilidad es especialmente clara cuando existen restricciones presupuestarias. Un modelo de madurez ayuda a distinguir entre mejoras urgentes y mejoras deseables. Por ejemplo, antes de desplegar herramientas avanzadas de correlación de eventos, una pyme con copias no probadas debería asegurar restauraciones fiables. Del mismo modo, antes de invertir en tecnologías complejas, puede ser más eficiente cerrar cuentas obsoletas, aplicar MFA y reducir accesos administrativos.

También puede apoyar la relación con proveedores. Muchas pymes externalizan correo, hosting, ERP, copias, soporte o ciberseguridad. El modelo permite formular preguntas concretas: quién restaura, en cuánto tiempo, qué evidencias se entregan, cómo se protegen credenciales privilegiadas, qué logs se conservan, qué se hace ante un cifrado masivo y cómo se comunica un incidente. Estas preguntas convierten la resiliencia en un requisito operativo y contractual.

Como ejemplo ilustrativo, una pyme de 40 empleados con ERP en servidor local, correo en la nube y soporte externo podría obtener nivel 2 en activos si mantiene un inventario parcial, nivel 3 en accesos si aplica MFA en correo y VPN, nivel 2 en copias si realiza backups pero no prueba restauraciones, nivel 1 en respuesta si no dispone de procedimiento documentado y nivel 2 en formación si solo imparte comunicaciones puntuales. Con este diagnóstico, el plan inicial no consistiría en implantar herramientas avanzadas, sino en probar restauraciones, cerrar cuentas obsoletas, documentar roles de incidente y fijar RTO/RPO mínimos para el ERP y la facturación.

7. Discusión

7.1. Aportaciones teóricas

Desde el punto de vista teórico, el artículo integra tres tradiciones que a menudo se abordan por separado: la seguridad de la información, la continuidad de negocio y los modelos de madurez. La propuesta considera el ransomware como un riesgo transversal que afecta a activos, procesos, personas y decisiones directivas. Esta lectura es coherente con la evolución de la ciberseguridad hacia enfoques de riesgo y resiliencia, en los que prevenir es necesario pero no suficiente (Dupont et al., 2023; Sepúlveda Estay et al., 2020).

La escala de cinco niveles permite traducir marcos extensos a un lenguaje de diagnóstico. Frente a modelos puramente normativos, la propuesta no pregunta únicamente si existe un control, sino en qué grado está formalizado, cubierto, probado y mejorado. Esta progresión facilita estudiar futuras relaciones entre madurez, capacidad de recuperación, continuidad operativa y cultura de seguridad en pymes.

7.2. Utilidad práctica para la gestión TI

Para la gestión TI, el modelo ofrece una herramienta de priorización. Permite ordenar iniciativas que suelen competir por recursos: actualización de sistemas, segmentación, revisión de cuentas, copias, formación, respuesta a incidentes o monitorización. En lugar de tratar cada medida como un proyecto aislado, el modelo las vincula a una capacidad de resiliencia.

El modelo también puede servir como base para cuadros de mando sencillos. Una pyme puede revisar trimestral o semestralmente el nivel de cada dimensión, registrar evidencias y actualizar el plan de mejora. Esta práctica ayuda a que la ciberseguridad deje de depender de acciones puntuales y pase a incorporarse a la gestión ordinaria.

7.3. Comparación con otros enfoques

El NIST CSF 2.0 utiliza perfiles y niveles de implementación para describir la gestión del riesgo de ciberseguridad, pero no prescribe una escala única de madurez para pymes. ISO/IEC 27001 permite certificar un sistema de gestión, pero puede resultar exigente para organizaciones que todavía necesitan ordenar capacidades básicas. El ENS establece un marco normativo aplicable a determinados sistemas, pero su finalidad tampoco es ofrecer una autoevaluación específica frente al ransomware.

La propuesta se sitúa en un punto intermedio. No compite con esos marcos ni los sustituye, sino que los sintetiza en una herramienta preliminar. Su valor reside en facilitar el tránsito desde una postura reactiva hacia prácticas gestionadas y medidas. Una pyme con nivel bajo puede usar el modelo como hoja de ruta; una pyme con nivel medio puede usarlo para preparar auditorías, mejorar contratos con proveedores o reforzar continuidad.

Su diferenciación principal está en combinar, en una misma matriz, evidencia verificable, continuidad operativa y priorización para ransomware. A diferencia de modelos más generales, no busca cubrir toda la ciberseguridad, sino identificar el nivel mínimo de capacidades que condicionan la recuperación de una pyme ante cifrado, interrupción y pérdida de disponibilidad.

8. Limitaciones y futuras líneas de investigación

La principal limitación del trabajo es su carácter conceptual. El modelo no ha sido validado mediante encuestas, entrevistas, estudios de caso ni aplicación en empresas reales. Por tanto, no se pueden inferir niveles medios de madurez, relaciones causales ni eficacia empírica. Las dimensiones y niveles se justifican por revisión y síntesis de marcos, pero requieren contraste en contextos organizativos diversos.

Una segunda limitación es la heterogeneidad de las pymes. No todas presentan la misma dependencia digital, exposición al exterior, regulación, sector, tamaño o capacidad presupuestaria. El modelo debe adaptarse al contexto y evitar comparaciones simplistas entre organizaciones. Un comercio pequeño, una industria con sistemas operacionales y una empresa de servicios digitales no tienen el mismo perfil de riesgo.

También debe tenerse en cuenta que la amenaza ransomware evoluciona. Nuevas técnicas de extorsión, abuso de credenciales, explotación de proveedores o ataques a servicios en la nube pueden exigir revisar las dimensiones y evidencias del modelo. Por ello, el instrumento debe entenderse como una propuesta actualizable.

Las futuras líneas de investigación deberían orientarse a tres tareas. Primero, validar el modelo con expertos en seguridad, continuidad y gestión TI. Segundo, aplicarlo en una muestra de pymes para analizar comprensión, esfuerzo de uso y utilidad percibida. Tercero, estudiar si las mejoras de madurez se relacionan con mejores tiempos de recuperación, menor impacto operativo y mayor capacidad de aprendizaje tras incidentes.

9. Conclusiones

El ransomware exige a las pymes una visión de ciberresiliencia que combine prevención, detección, respuesta y recuperación. La madurez no puede reducirse a disponer de herramientas de seguridad; depende también de políticas, inventarios, accesos, copias probadas, continuidad, formación y capacidad de aprendizaje. Esta integración es especialmente relevante en organizaciones con recursos limitados, donde las decisiones deben priorizarse con criterios claros.

El artículo ha propuesto un modelo de madurez compuesto por diez dimensiones y cinco niveles. Su objetivo es facilitar una autoevaluación práctica, basada en evidencias y orientada a la mejora gradual. La propuesta distingue entre diagnóstico y certificación: no pretende sustituir auditorías normativas ni validar cumplimiento de estándares, sino ofrecer una guía inicial para identificar brechas y orientar planes de mejora frente al ransomware.

La aportación principal reside en traducir marcos extensos de seguridad y continuidad a una herramienta comprensible para pymes. El modelo permite visualizar el estado actual, definir un nivel objetivo y priorizar acciones con impacto operativo. Su aplicación futura deberá validarse empíricamente, pero su estructura ofrece una base razonable para avanzar desde respuestas reactivas hacia una gestión medible y sostenida de la ciberresiliencia.

Referencias

- Barker, W. C., Fisher, W. F., Scarfone, K., & Souppaya, M. (2022). Ransomware risk management: A cybersecurity framework profile (NIST IR 8374). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.IR.8374>
- Becker, J., Knackstedt, R., & Pöppelbuß, J. (2009). Developing maturity models for IT management. *Business & Information Systems Engineering*, 1(3), 213-222. <https://doi.org/10.1007/s12599-009-0044-5>
- Cuenca Rodríguez, M. E., & Serrano Betored, P. (2026). Integración de la tecnología en la enseñanza musical: formación, uso y barreras desde la perspectiva docente. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, (61), 51-65. <https://doi.org/10.17013/risti.61.51-65>
- de Bruin, T., Rosemann, M., Freeze, R., & Kulkarni, U. (2005). Understanding the main phases of developing a maturity assessment model. In *Proceedings of the 16th Australasian Conference on Information Systems*, (pp. 8-19).
- Dupont, B., Shearing, C., Bernier, M., & Leukfeldt, R. (2023). The tensions of cyber-resilience: From sensemaking to practice. *Computers & Security*, 132, 103372. <https://doi.org/10.1016/j.cose.2023.103372>
- Eliot, D. (2024). NIST Cybersecurity Framework 2.0: Small Business Quick-Start Guide (NIST SP 1300). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.1300>
- European Union Agency for Cybersecurity. (2021). Cybersecurity for SMEs: Challenges and recommendations. ENISA.
- Ilca, L. F., Lucian, O. P., & Balan, T. C. (2023). Enhancing cyber-resilience for small and medium-sized organizations with prescriptive malware analysis, detection and response. *Sensors*, 23(15), 6757. <https://doi.org/10.3390/s23156757>
- Instituto Nacional de Ciberseguridad. (2019). Recomendaciones para hacer copias de seguridad en la nube. INCIBE.

- Instituto Nacional de Ciberseguridad. (2021). Ransomware: Una guía de aproximación para el empresario. INCIBE.
- International Organization for Standardization. (2019). ISO 22301:2019 Security and resilience - Business continuity management systems - Requirements. ISO.
- International Organization for Standardization, & International Electrotechnical Commission. (2022a). ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection - Information security management systems - Requirements. ISO.
- International Organization for Standardization, & International Electrotechnical Commission. (2022b). ISO/IEC 27002:2022 Information security, cybersecurity and privacy protection - Information security controls. ISO.
- International Organization for Standardization, & International Electrotechnical Commission. (2022c). ISO/IEC 27005:2022 Information security, cybersecurity and privacy protection - Guidance on managing information security risks. ISO.
- Lella, I., Tsekmezoglou, E., Naydenov, R., Malatras, A., Garcia, S., Valeros, V., & Gomaa, A. (Eds.). (2022). Threat landscape for ransomware attacks. European Union Agency for Cybersecurity. <https://doi.org/10.2824/456263>
- Ministerio de Asuntos Económicos y Transformación Digital. (2022). Real Decreto 311/2022, de 3 de mayo, por el que se regula el Esquema Nacional de Seguridad. Boletín Oficial del Estado, 106, 61715-61804.
- National Institute of Standards and Technology. (2012). Guide for conducting risk assessments (NIST SP 800-30 Rev. 1). <https://doi.org/10.6028/NIST.SP.800-30r1>
- National Institute of Standards and Technology. (2024). The NIST Cybersecurity Framework (CSF) 2.0 (NIST CSWP 29). <https://doi.org/10.6028/NIST.CSWP.29>
- Nelson, A., Rekhi, S., Scarfone, K., & Souppaya, M. (2025). Incident response recommendations and considerations for cybersecurity risk management: A CSF 2.0 community profile (NIST SP 800-61 Rev. 3). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-61r3>
- Perozzo, H., Zaghloul, F., & Ravarini, A. (2022). CyberSecurity Readiness: A model for SMEs based on the socio-technical perspective. *Complex Systems Informatics and Modeling Quarterly*, 33, 53-66. <https://doi.org/10.7250/csimq.2022-33.04>
- Sepúlveda Estay, D. A., Sahay, R., Barfod, M. B., & Jensen, C. D. (2020). A systematic review of cyber-resilience assessment frameworks. *Computers & Security*, 97, 101996. <https://doi.org/10.1016/j.cose.2020.101996>
- Souppaya, M., & Scarfone, K. (2013). Guide to malware incident prevention and handling for desktops and laptops (NIST SP 800-83 Rev. 1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-83r1>

- Swanson, M., Bowen, P., Phillips, A. W., Gallup, D., & Lynes, D. (2010). Contingency planning guide for federal information systems (NIST SP 800-34 Rev. 1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-34r1>
- Yigit Ozkan, B., Spruit, M., Wondolleck, R., & Burriel Coll, V. (2020). Modelling adaptive information security for SMEs in a cluster. *Journal of Intellectual Capital*, 21(2), 235-256. <https://doi.org/10.1108/JIC-05-2019-0128>
- Yigit Ozkan, B., van Lingen, S., & Spruit, M. (2021). The Cybersecurity Focus Area Maturity (CYSFAM) model. *Journal of Cybersecurity and Privacy*, 1(1), 119-139. <https://doi.org/10.3390/jcp1010007>

Critérios Editoriais

A RISTI (Revista Ibérica de Sistemas e Tecnologias de Informação) é um periódico científico, que foca a investigação e a aplicação prática inovadora no domínio dos sistemas e tecnologias de informação.

O Conselho Editorial da RISTI incentiva potenciais autores a submeterem artigos originais e inovadores para avaliação pelo Conselho Científico.

A submissão de artigos para publicação na RISTI deve realizar-se de acordo com as chamadas de artigos e as instruções e normas disponibilizadas no sítio Web da revista (<http://www.risti.xyz/>).

Todos os artigos submetidos são avaliados por um conjunto de membros do Conselho Científico, não inferior a três elementos.

Em cada número da revista são publicados entre cinco a oito dos melhores artigos submetidos.

Criterios Editoriales

La RISTI (Revista Ibérica de Sistemas y Tecnologías de la Información) es un periódico científico, centrado en la investigación y en la aplicación práctica innovadora en el dominio de los sistemas y tecnologías de la información.

El Consejo Editorial de la RISTI incentiva autores potenciales a enviar sus artículos originales e innovadores para evaluación por el Consejo Científico.

El envío de artículos para publicación en la RISTI debe hacerse de conformidad con las llamadas de los artículos y las instrucciones y normas establecidas en el sitio Web de la revista (<http://www.risti.xyz/>).

Todos los trabajos enviados son evaluados por un número de miembros del Consejo Científico de no menos de tres elementos.

En cada número de la revista se publican cinco a ocho de los mejores artículos enviados.



Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

©ITMA 2026 <http://www.risti.xyz>

