19 FONDO EDITORIAL Revista de la Carrera de Ingeniería de Sistemas Julio 2024

UNIVERSIDAD DE LIMA

# INTERFASES

# INTERFASES

*Interfases*
Revista de la Carrera de Ingeniería de Sistemas de la Facultad de Ingeniería
de la Universidad de Lima
N.º 19, julio, 2024
doi: https://doi.org/10.26439/interfases2024.n19
Lima, Perú


Fondo Editorial
Av. Javier Prado Este 4600
Urb. Fundo Monterrico Chico
Santiago de Surco, Lima, Perú
Código postal 15023
Teléfono (511) 437-6767, anexo 30131
fondoeditorial@ulima.edu.pe
www.ulima.edu.pe

Edición, diseño y carátula: Fondo Editorial de la Universidad de Lima.

Correspondencia:
interfases@ulima.edu.pe



Periodicidad: semestral
Arbitraje editorial: revisión por pares de doble ciego
Directorios y catálogos: Redalyc, CrossRef, Dialnet, Latindex y DOAJ

ISSN: 1993-4912 (en línea)
Hecho el depósito legal en la Biblioteca Nacional del Perú n.º 2020-09967

**REVISORES CIENTÍFICOS DE EDICIONES PASADAS**

# ÍNDICE

# PRESENTACIÓN

doi: https://doi.org/10.26439/interfases2024.n19.7294

La revista *Interfases* se enorgullece en presentar su edición número 19, correspondiente al periodo enero-julio del 2024. Esta edición incluye once artículos científicos que abordan temas de vanguardia en ciencias de la computación, ingeniería de *software*, sistemas de información, tecnologías de la información, ciberseguridad, ciencia de datos y áreas afines. Nuestro firme compromiso de promover la difusión del conocimiento científico se refleja en esta publicación, que se ofrece en acceso abierto para que nuestros lectores puedan aprovechar plenamente los avances generados por nuestros colaboradores, apasionados por la investigación y el desarrollo tecnológico.

Fiel a su tradición, la revista *Interfases* continúa contribuyendo de manera ininterrumpida al progreso de la ciencia y la ingeniería. En esta edición, hemos recibido manuscritos de autores provenientes de diversas universidades, países e idiomas. Tras un riguroso proceso de revisión por pares ciegos, se seleccionaron once manuscritos de alta calidad. Estos artículos no solo reflejan la diversidad y el alcance global de nuestra comunidad académica, sino también el compromiso de *Interfases* con la excelencia y la innovación en la investigación científica.

En primer lugar, presentamos el trabajo de Alejandra Morales Ramírez, Rodolfo García Lozano, Juan de Jesús Amador Reyes y Cuauhtémoc Hidalgo Cortés, todos ellos de la Universidad Autónoma del Estado de México, quienes proponen la metodología de minería de procesos para el desarrollo de proyectos empresariales y científico-académicos. En segundo lugar, Rosa Flor Gomez Risco de la Universidad Nacional de Piura, Perú, presenta un estudio estadístico que compara la presencia de mujeres con la de hombres en términos de niveles educativos, grupos de edad y años de inicio en carreras STEM, lo que puede servir como base para desarrollar soluciones dirigidas a cerrar brechas. En tercer lugar, Madeleine Gillian Rabines Floreano de la Universidad Nacional de Trujillo, Perú, y Lourdes Ramírez Cerna de la Universidad de Lima, Perú, realizan un estudio sobre el panorama de las mujeres peruanas en carreras de STEM.

En cuarto lugar, Zoraida Mamani Rodríguez de la Universidad Nacional Mayor de San Marcos, Lima, Perú, propone el diseño e implementación de una aplicación *cloud native* en

una perspectiva de ingeniería de *software* continua, aplicada al caso de estudio SIGCON. En quinto lugar, Michael Dorin de la University of St. Thomas, Minnesota, Estados Unidos, y Sergio Montenegro de la Julius-Maximilians-Universität de Würzburg, Alemania, muestran en su artículo que es posible crear de manera económica y sostenible un laboratorio que pruebe eficientemente y de forma correcta los algoritmos y protocolos de encriptación utilizando tabletas desechadas y computadoras de placa única económicas.

En sexto lugar, Marcelo López-Nocera y María F. Pollo-Cattaneo de la Universidad Tecnológica Nacional de Buenos Aires, Argentina, y Francisco Redelico del Instituto de Medicina Traslacional e Ingeniería Biomédica de Buenos Aires, Argentina, realizaron un análisis de la brecha entre la universidad y la industria del *software* en la República Argentina: una perspectiva docente y posibles soluciones. En séptimo lugar, Erly Galia Villarroel Enriquez y Juan Gutiérrez-Cárdenas de la Universidad de Lima, Perú, presentan un trabajo que utiliza la frecuencia de llamadas al sistema para detectar y clasificar *malware* utilizando los algoritmos XGBoost, LightGBM y random forest.

En octavo lugar, Erick Leonel García Ibáñez de la Peter the Great St. Petersburg Polytechnic University, Rusia, desarrolla un análisis de datos para determinar la escala de las empresas gestoras de activos más prominentes. En noveno lugar, Eduardo Rodrigo Wong León y Marco Antonio Coral Ygnacio de la Universidad Católica Sedes Sapientiae, Perú, llevan a cabo una revisión sistemática de la literatura sobre implementaciones de sistemas de control de tráfico. En décimo lugar, Elaine Rodrigues Koller, Paulo de Moura y Patrícia de Sá Freire de la Universidade Federal de Santa Catarina (UFSC), Brasil, realizan una revisión integrativa de la gestión del conocimiento como herramienta estratégica de innovación en las organizaciones.

Por último, Aron Lo Li y Juan Gutiérrez-Cárdenas de la Universidad de Lima y Víctor H. Ayma de la Universidad del Pacífico proponen la creación de un conjunto de datos, así como una metodología que permita a los usuarios capturar patrones de tecleo de estudiantes pertenecientes a una universidad en Lima, Perú, a través de un entorno en la nube y desde sus propios dispositivos.

Finalmente, expresamos nuestro más sincero agradecimiento a todos los investigadores que enviaron sus manuscritos para esta edición de *Interfases*. Apreciamos el esfuerzo y el tiempo invertidos en desarrollar y compartir sus investigaciones. Las contribuciones de nuestros autores nos proporcionan valiosas enseñanzas y abren prometedoras líneas de trabajo futuro. Extendemos también nuestro agradecimiento a los revisores de *Interfases*, quienes garantizan la calidad de nuestra revista a través de un riguroso proceso de evaluación por pares, en consonancia con las mejores prácticas de investigación y ética en la publicación científica.

Dra. Nadia Katherine Rodríguez Rodríguez
Directora de *Interfases*

# Diseño de una metodología de minería de procesos para el desarrollo de proyectos de tipo empresarial y científico-académico

Alejandra Morales Ramírez
amoralesr@uaemex.mx
http://orcid.org/0000-0002-8737-5985
Universidad Autónoma del Estado de México, México

Rodolfo García Lozano
rzgarcial@uaemex.mx
http://orcid.org/0000-0003-1087-6156
Universidad Autónoma del Estado de México, México

Juan de Jesús Amador Reyes
jjamadorr@uaemex.mx
https://orcid.org/0000-0003-1925-2710
Universidad Autónoma del Estado de México, México

Cuauhtémoc Hidalgo Cortés
chidalgoc@uaemex.mx
http://orcid.org/0000-0001-6324-7180
Universidad Autónoma del Estado de México, México

Recibido: 31 de octubre de 2023 / Aceptado: 26 enero de 2024
doi: https://doi.org/10.26439/interfases2024.n19.6732

RESUMEN. La minería de procesos es una especialidad que permite a las organizaciones descubrir, analizar y mejorar sus procesos reales de negocio mediante la extracción de conocimiento de los registros de eventos que se encuentran en los sistemas de información actuales. La presente investigación tuvo como objetivo proponer una metodología de minería de procesos para el desarrollo de proyectos de tipo empresarial y científico-académico que consta de nueve fases: planeación y alcance, preprocesamiento de datos, procesamiento de datos, análisis de control de flujo, análisis de rendimiento, análisis de roles, presentación de resultados, publicación de resultados, y transferencia y seguimiento. La metodología propuesta es el resultado de la experiencia adquirida a través de la revisión de la literatura, del análisis de las metodologías publicadas y de los conocimientos y experiencias de los investigadores. Es conveniente resaltar que, en investigaciones futuras, se aplicará la metodología propuesta de minería de procesos a diversos casos de estudio, tanto del área empresarial como del ámbito científico-académico. Por ello, el objetivo de este trabajo de investigación está enfocado en analizar

la eficiencia de la metodología, la facilidad y pertinencia de su aplicación, así como la congruencia entre las fases que la integran.

PALABRAS CLAVE: minería de procesos / metodología de desarrollo / registro de eventos / conocimiento / proyectos empresariales / proyectos científico-académicos

## DESIGN OF A PROCESS MINING METHODOLOGY FOR THE DEVELOPMENT OF BUSINESS AND SCIENTIFIC-ACADEMIC PROJECTS

ABSTRACT. Process mining enables organizations to discover, analyze, and enhance their business processes by extracting knowledge from event logs available in current information systems. This research proposes a process mining methodology for developing business and scientific-academic projects, consisting of nine phases: planning and scope, data preprocessing, data processing, flow control analysis, performance analysis, role analysis, results presentation, results publication, and transfer and monitoring. This methodology results from the researcher's literature review, analysis of published methodologies, and their knowledge and experiences. Future research will apply this process mining methodology to various case studies in business and scientific-academic domains. This research aims to analyze the methodology's efficiency, its application's ease and relevance, and the congruence between its phases.

KEYWORDS: process mining / development methodology / event log / knowledge / business projects / scientific-academic projects

## INTRODUCCIÓN

Hoy en día muchas organizaciones destinan recursos para gestionar y mejorar sus procesos de negocio mediante el uso de diversos sistemas de información (SI). Estos sistemas permiten registrar datos importantes relacionados con sus recursos ejecutores, eventos, tiempos de inicio y finalización de cada actividad, y otras variables asociadas a la ejecución de los procesos (Silva Osses et al., 2016; Aguirre & Rincón, 2015).

Los avances tecnológicos de los últimos años han permitido que la capacidad de procesamiento de la información crezca de manera espectacular. Esto hace que los SI puedan almacenar información histórica sobre la ejecución real de los procesos de negocio y, de este modo, contar con grandes volúmenes de datos que las organizaciones pueden usar para saber qué sucede dentro de los procesos de negocio, para diagnosticar problemas y para sugerir el tratamiento adecuado.

No obstante, cuando se incrementa la cantidad de datos almacenados, la capacidad para entenderlos se reduce (Merchán et al., 2021). Analizar los datos de manera tradicional requiere de mucho tiempo, involucra a muchas personas y es costoso. Por ello, se hace fundamental el uso de herramientas de análisis de datos que permitan obtener conocimiento útil de manera automatizada a partir de grandes volúmenes de datos, como la inteligencia artificial, la minería de datos y la minería de procesos - MP (Checoli et al., 2020).

La MP es una disciplina de investigación relativamente nueva y novedosa, que se localiza, por un lado, entre la inteligencia computacional y la minería de datos y, por otro, entre la modelación y el análisis de procesos de negocio (González Gonzáles et al., 2019; Van der Aalst, 2016; Van der Aalst, 2011). Es una disciplina que está evolucionando (Merchán et al., 2021) y que permite que las organizaciones obtengan información sobre los procesos de negocio y entiendan qué está pasando *de facto*, a partir de la extracción del conocimiento de los registros de eventos que se encuentran disponibles en los SI actuales (Van der Aalst, 2016; Aguirre & Rincón, 2015).

Hay tres técnicas de MP que se pueden realizar utilizando el registro de eventos: descubrimiento, conformidad y mejora. El descubrimiento de procesos tiene la finalidad de generar un modelo de proceso sin utilizar previamente información sobre cómo es o cómo debería de ser el flujo real del proceso. Por otra parte, el análisis de conformidad compara el modelo del proceso existente con el registro de eventos del mismo proceso, con el objetivo de verificar si la realidad es igual al modelo y viceversa. Esto puede ser utilizado para valorar si lo observado se ajusta al flujo ideal de trabajo. Finalmente, el mejoramiento busca corregir o rediseñar un modelo de proceso existente utilizando la información del proceso real guardada dentro del registro de eventos (Van der Aalst et al., 2012).

En los últimos años, la MP ha adquirido un interés creciente, tanto en la práctica a nivel mundial dentro del ámbito industrial y empresarial (Badakhshan et al., 2022;

Van der Aalst, 2012), como en la comunidad científica en la cual se observa un incremento en las investigaciones relacionadas (Fuentes et al., 2019). Este interés se debe a que, a través de su uso, se obtiene conocimiento basado en el registro de eventos y no en opiniones subjetivas o experiencias obsoletas. Esto ayuda a las organizaciones a proponer acciones de mejora o rediseño para que, al ser implementadas, se alcancen procesos de negocio más eficientes (Dos Santos et al., 2019; Van der Aalst, 2013; Van der Aalst, 2012).

Para llevar a cabo proyectos de MP prácticos y científicos en distintos contextos, se han utilizado diversos enfoques metodológicos que han surgido desde la aparición de esta disciplina. Algunos de estos enfoques inician directamente con la preparación de datos hasta llegar a la presentación de resultados: por ejemplo, el método de diagnóstico de procesos o PDM (Bozkaya et al., 2009) y el marco metodológico de minería de procesos o PMMF, por sus siglas en inglés (De Weerdt et al., 2013). Otros enfoques metodológicos abarcan desde la planeación hasta la implementación de mejoras y soporte del proceso: el ciclo de vida L* (Van der Aalst et al., 2012), la metodología de proyectos de minería de procesos o PMPM (Van der Heijden, 2012), la metodología de proyectos de minería de procesos ($PM^2$) (Van Eck et al., 2015), la extensión de la metodología $PM^2$ (Silva Osses et al., 2016) y la guía de análisis para la MP enfocada en el usuario (Céspedes et al., 2018).

Sin embargo, debido a que esta disciplina es reciente, el desarrollo de varias de estas metodologías ha dependido de la información disponible, de la experiencia de los autores, del proyecto a realizar y del objetivo que se persigue. Por estas razones, es posible observar que en los diferentes enfoques se presentan etapas con objetivos y actividades muy similares, pero organizadas de diferente manera. Otra excepción se presenta en los estudios de caso. Como consecuencia de su enfoque académico, este tipo de proyectos regularmente no contempla la fase de implementación de mejoras y seguimiento, como se observa en diversas investigaciones (Butt et al., 2023; Martínez-Escobar et al., 2021; Sangil, 2020; Emamjome et al., 2019; Terragni & Hassani, 2018; Silva Osses, 2017; Park & Sik, 2016).

La presente investigación tiene como objetivo proponer una metodología para el desarrollo de proyectos de MP que contemple las siguientes características:

- Considere todas las etapas que se podrían presentar en el desarrollo de un proyecto de MP: entendimiento del proceso, tratamiento de datos, aplicación de las técnicas y herramientas de la MP, comunicación de resultados e implementación de recomendaciones o acciones de mejora.
- Sirva de base para el desarrollo de proyectos, tanto de tipo empresarial como científico-académico.
- Dentro de sus objetivos contemple la opción de hacer divulgación científica a través de la publicación de artículos.

La visión es proponer una metodología de desarrollo de proyectos de MP que pueda aplicarse, con sus propias características, en diferentes ámbitos.

## 2. METODOLOGÍA

La investigación se abordó a través de las siguientes dos fases:

1. **Revisión de la literatura.** La finalidad de esta fase fue recopilar información relevante sobre el desarrollo de las metodologías de MP publicadas entre los años 2009 y 2020. La búsqueda se realizó en tres bases de datos: *Springer Link, EBSCOhost* y *ScienceDirect*. El criterio de búsqueda se basó en la combinación del concepto "minería de procesos" con las palabras "proyecto", "metodología", "marco" y "guía" que pudieran aparecer en el título, el resumen o las palabras clave de los artículos y ponencias de congreso que se consultaron.

   La identificación basada en el criterio de búsqueda antes descrito dio como resultado 19 estudios elegibles.

   Sin embargo, doce de ellos se excluyeron, principalmente porque las metodologías propuestas se encontraban dirigidas a proyectos de negocio o áreas específicas de conocimiento (por ejemplo, las metodologías creadas para realizar proyectos de MP en el área de la salud). Se consideró que las metodologías propuestas en las investigaciones descartadas no podrían ser utilizadas en su totalidad para desarrollar otros proyectos de MP en escenarios con diferentes características o que estén relacionados con otras áreas del conocimiento (por ejemplo, el área científico-académica).

   Esto resultó en siete artículos seleccionados que fueron revisados y analizados por los autores para ser incluidos y descritos en la investigación que se presenta a continuación.

2. **Análisis comparativo de las metodologías encontradas.** En esta fase, en primer lugar, se definió una metodología general para agrupar las actividades descritas en cada una de las metodologías analizadas en el paso anterior. Posteriormente, se identificaron y describieron las áreas de oportunidad como punto de partida para realizar la propuesta de desarrollo.

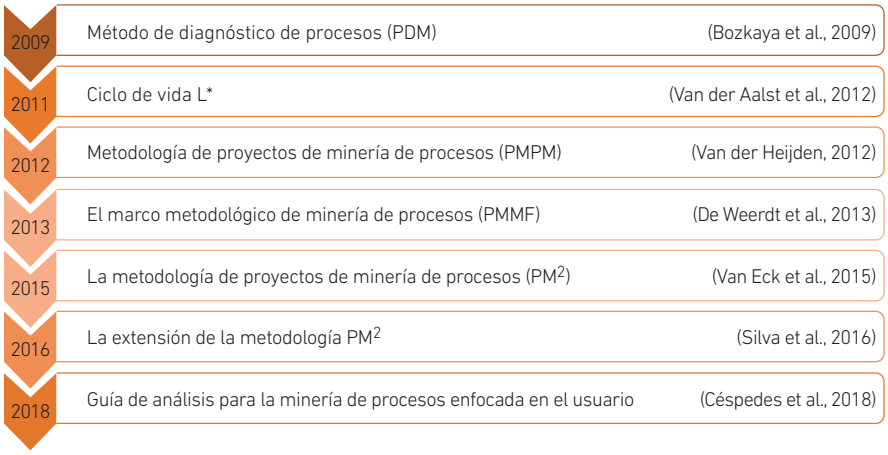
### *2.1 Revisión de la literatura*

Desde que surgió la MP, diversos autores han desarrollado enfoques metodológicos con el objetivo de aplicar esta disciplina en diferentes escenarios. Estos desarrollos

se encuentran compuestos por diversas fases con objetivos particulares, que van desde la planificación hasta la propuesta de mejoras y el soporte del proceso. Los trabajos más relevantes que cumplieron con los criterios de selección de esta investigación son los que se muestran en la Figura 1 y se describen en los siguientes apartados:

**Figura 1**

*Enfoques metodológicos de desarrollo de proyectos de MP*




## PDM

El PDM comprende seis etapas de trabajo y ha sido utilizado para dar una visión general del proceso de emisión de documentos —implementado en Oracle 9i— de una organización gubernamental holandesa (Bozkaya et al., 2009) y para descubrir la información relevante respecto de la ejecución y desempeño real del proceso de atención de servicios básicos al cliente de una empresa dedicada al suministro del servicio eléctrico (Morales et al., 2022).

Las fases de este método son las siguientes:

- *Preparación del registro.* En esta etapa se extraen los datos de los SI y se adecuan al formato necesario para que puedan ser utilizados por la MP. Todos estos datos se organizan en casos (ID únicos), actividades, marcas de tiempo, recursos y otros campos que puedan aportar información importante del proceso.
- *Inspección de registros.* En esta etapa se realiza la estadística necesaria para entender la estructura del registro de eventos. En este análisis, se reconocen los posibles procesos, las actividades principales, se obtiene el número de casos, el total de eventos, el mínimo, máximo y promedio de eventos por casos, y se identifican los eventos iniciales y finales. Como resultado de la estadística,

en algunas ocasiones es necesario eliminar casos incompletos, con la finalidad de no generar ruido en la siguiente etapa de análisis.

- *Análisis de control de flujo.* Esta etapa tiene como objetivo analizar el registro de eventos para conocer el proceso real de la empresa y verificar la conformidad; es decir, identificar si cada uno de los casos puede ser reproducido en el proceso descubierto. Es importante considerar que, para obtener modelos de proceso estructurados, en muchas ocasiones es necesario eliminar los eventos poco frecuentes, a través de la utilización de filtros, lo que evitará generar modelos tipo espagueti (no estructurados).
- *Análisis de rendimiento.* En esta fase se estudian los modelos obtenidos en el análisis de control de flujo, con el objetivo de conocer, por ejemplo, el desempeño del proceso e identificar los cuellos de botella, el tiempo de rendimiento de las actividades y del proceso en general. Lo anterior permite obtener información para reconocer las áreas que se pueden mejorar en el proceso de la empresa.
- *Análisis de roles.* Si el registro de eventos contiene información suficiente de quién ejecutó un evento, en esta fase se analizan los roles (persona o recurso) del proceso. Para esta actividad, Bozkaya et al. (2009) proponen realizar una tabla de frecuencias, en la que las columnas representan cada evento del registro y los renglones los roles. Así, en cada celda se coloca el número de veces que ese rol ejecutó cierto evento. Posteriormente, se elabora un perfil para cada rol. Si los roles tienen perfiles afines, forman un grupo. Lo anterior permite que la organización compare los grupos descubiertos con el modelo organizativo utilizado, con la finalidad de detectar si las personas realizan actividades similares y observar cómo se encuentra la comunicación y la división del trabajo dentro de los departamentos.
- *Transferencia de resultados.* El objetivo de esta etapa es crear un documento para notificar a los interesados del proyecto los resultados de los hallazgos de la investigación y, en caso de ser necesario, para reconfigurar los procesos y hacerlos más eficientes.

**Ciclo de vida L***

Esta metodología, que fue desarrollada para ser aplicada en proyectos cuyos procesos se encuentran estables y bien estructurados, se encuentra conformada por cinco etapas (Van der Aalst et al., 2012):

- *Planificación y justificación.* Antes de comenzar con la aplicación de la metodología, en esta etapa (0) se define una planeación y justificación, con la finalidad de identificar los beneficios que se podrán lograr a partir del proyecto.

- *Extracción.* En esta etapa (1), a partir de los SI y de los expertos en el dominio y la gestión, se extraen los datos de los eventos, modelos y preguntas. De esta fase resultan los objetivos, preguntas, modelos hechos a mano, datos históricos, etcétera.
- *Creación de modelo de control de flujo.* Esta etapa (2) tiene como objetivo construir el modelo de control de flujo y relacionarlo con el registro de eventos a través de las técnicas de descubrimiento de la MP. Con la información obtenida, se podrían contestar las preguntas planteadas, disparar acciones de rediseño o ajustar el proceso analizado.
- *Creación de modelo de proceso integrado.* Esta etapa (3) es utilizada para extender los modelos de proceso a través del análisis de otras perspectivas como tiempo, recursos y datos. Por ejemplo, se utilizan los tiempos marcados en los eventos para calcular los tiempos de espera de cada actividad. El resultado es un modelo de proceso más completo, que permite apoyar la toma de acciones.
- *Soporte operacional.* En esta etapa (4), el conocimiento descubierto de los datos del registro de eventos históricos se une con la información de los casos que se encuentran en ejecución y es utilizado para realizar las siguientes tres actividades de apoyo: intervención, predicción y recomendación.

Es importante resaltar que esta metodología no puede ser aplicada en su totalidad cuando los procesos no se encuentran estructurados, o —de ser el caso—, solo se podrán llevar a cabo las tres primeras etapas.

**PMPM**

Esta metodología está conformada por seis etapas que permiten conducir proyectos orientados al negocio. Fue probada en el departamento de servicios financieros holandés Rabobank, que engloba 141 bancos locales (Van der Heijden, 2012). Sus etapas son:

- *Alcance.* En esta fase se identifica y comprende el funcionamiento del proceso, incluyendo las actividades, recursos, restricciones, SI, etcétera. Además, se formulan los objetivos y metas que ayudarán a definir las preguntas del proyecto y determinar qué herramientas y técnicas serán utilizadas para crear el registro de eventos y realizar las actividades de la MP.
- *Comprensión de datos.* La finalidad de esta fase es ubicar dónde se encuentran los datos almacenados y descubrir cómo están organizados y relacionados. Además, se verifica la calidad de estos datos; es decir, que sean confiables, íntegros y seguros.
- *Creación del registro de eventos*. En esta fase se crea el registro de eventos a través de (a) la selección de los datos, los cuales pueden ser casos iniciados,

finalizados, actuales, ocurridos en un periodo de tiempo, etcétera; (b) la exportación de los datos seleccionados de los SI a un formato adecuado; y (c) la preparación de los datos a través de la creación de atributos derivados o transformación de los datos a otros valores o la inserción de variables predeterminadas, según sea conveniente.

- *Minería de procesos.* En esta fase se lleva a cabo la familiarización con el registro de eventos, para asegurar que tal registro se encuentre lo suficientemente estructurado como para aplicar las técnicas de MP requeridas, para realizar un análisis real de los datos y para generar conocimiento suficiente que responda las preguntas que ayuden a mejorar el proceso.
- *Evaluación.* En esta etapa se evalúan los modelos que se construyeron en la fase anterior, mediante la verificación, validación y acreditación de los resultados obtenidos con los objetivos planteados. Cuando resulte necesario se pueden realizar otros análisis —más específicos— que agreguen valor a los objetivos. Además, podrían surgir nuevas preguntas que requieran la creación de otro registro de eventos.
- *Despliegue.* En esta última fase, la información adquirida se debe transferir a la organización a través de la redacción y presentación de un informe en el que se incluyan recomendaciones sobre acciones de mejora que podrían llevarse a cabo en el proceso.

**PMMF**

Este marco metodológico es adaptable tanto para los SI estructurados como para los no tan estructurados. Fue probado en una compañía de seguros belga que pertenece a la industria de servicios financieros y que incluye el manejo de seguros de ahorros para el retiro y seguros de vida. El sistema de información de la compañía es un sistema de gestión de documentos. Este marco se encuentra conformado por las siguientes cinco etapas (De Weerdt et al., 2013):

- *Preparación de los datos.* El objetivo de esta fase es extraer los datos almacenados en los SI, los cuales tienen que ser relevantes en base al alcance del proceso y al marco de tiempo definido, para poder realizar los análisis correspondientes.
- *Exploración de datos.* En esta fase se realiza la exploración de los datos a través de análisis estadísticos y de los diversos algoritmos de la MP, a fin de obtener diversas visualizaciones iniciales del proceso que se está analizando. De esta manera los expertos comerciales podrán trabajar iterativamente con la fase de preparación de los datos, el alcance del proceso y el marco de tiempo. Lo anterior, con la finalidad de garantizar que los datos de entrada realmente sean los adecuados para el análisis de las siguientes fases.

- *Perspectiva.* Con el conjunto de datos delimitado según el tiempo y el alcance, en esta fase se identifican diferentes perspectivas de análisis a través de la construcción de varios registros de eventos, siempre y cuando el entorno del proceso sea desconocido por el usuario.
- *Análisis.* Esta fase se divide en dos segmentos: el análisis de descubrimiento básico; y el análisis de cumplimiento y rendimiento.
  El primero incluye: (a) la perspectiva del flujo de control que, a su vez, incluye el análisis de las secuencias de actividad dentro del proceso; (b) la perspectiva desde un punto de vista organizacional; es decir, mediante los datos, se investigan los equipos o personas involucrados en el proceso; (c) la perspectiva de caso, que explora otros atributos de las ejecuciones de procesos contenidos en el registro de eventos para descubrir patrones particulares.
  En la fase del análisis de cumplimiento y rendimiento se realizan los análisis de evaluación: por ejemplo, los tiempos de procesamiento de casos y la validación de que la realidad es consistente con el comportamiento del proceso esperado o requerido.
- *Resultados.* En esta fase se exhibe el análisis de los resultados, considerado como un valioso punto de partida para la mejora de los procesos o incluso para realizar una reingeniería en ellos, con la finalidad de optimizarlos. La gerencia puede definir nuevos objetivos y medidas en función de los conocimientos obtenidos a través de la MP para resolver, por ejemplo, las ineficiencias identificadas en el proceso.

**PM²**

Esta metodología fue aplicada en el proceso de compras de repuestos del servicio de *hardware* de la empresa IBM, multinacional líder en tecnología y consultoría. Se encuentra conformada por seis etapas que pueden emplearse en procesos estructurados y no estructurados (Van Eck et al., 2015), y que se describen a continuación:

- *Planificación*. En esta etapa se elige el proceso de negocio y se establecen los objetivos, que serán traducidos a preguntas de investigación con la finalidad de orientar el desarrollo del proyecto. Estas preguntas se responderán a través de los datos que se encuentran en el registro de eventos.
  En esta etapa también se conforma el equipo de trabajo del proyecto, considerando los siguientes perfiles: propietario del negocio (encargado de los procesos de negocios), experto del negocio (con conocimientos en aspectos comerciales y de ejecución de los procesos), experto en sistemas (familiarizado con el rubro de TI de los procesos y los sistemas que los soportan) y analistas del proceso (expertos en análisis de procesos y en aplicación de técnicas de MP).

- *Extracción.* En esta etapa se establece el límite de la extracción de los datos, que serán obtenidos de los SI que soportan la ejecución de los procesos de negocio seleccionados para ser analizados. Además, se realiza la transferencia de conocimiento entre los expertos comerciales y los analistas de proceso, con el objetivo de que estos últimos sean certeros en las etapas de procesamiento y análisis.
- *Procesamiento de datos.* En esta etapa se obtiene el registro de eventos para poder aplicar las diferentes técnicas de MP, utilizando —de ser necesario— las opciones para crear vistas, agregar eventos, enriquecer registros y filtrar registros.
- *Minería y análisis.* En esta etapa se aplican las técnicas de MP al registro de eventos, con el objetivo de responder las preguntas de investigación formuladas en la etapa de planificación y de obtener información sobre el rendimiento y el cumplimiento de los procesos. Además de las tres actividades de MP, en esta etapa también puede utilizarse información adicional (por ejemplo, de los recursos o de los tiempos) y aplicar otras técnicas de análisis al registro de eventos y a los modelos de procesos: por ejemplo, las técnicas de minería de datos o análisis visuales (histogramas de eventos por caso), cuyos resultados pueden servir para mejorar los modelos de procesos con información adicional.
- *Evaluación*. En esta etapa se relacionan los hallazgos obtenidos del análisis con ideas de mejora que permitan cumplir con los objetivos iniciales del proyecto. Es fundamental que los expertos del proceso participen en la verificación y validación de los resultados, para asegurar que sean útiles para la organización. Dentro de los resultados obtenidos se pueden diseñar ideas de mejora o nuevas preguntas de investigación.
- *Mejora y soporte de proceso.* El objetivo de esta etapa es utilizar los conocimientos adquiridos para modificar la ejecución real del proceso. Por tal motivo, las entradas son las propuestas de mejora de la etapa de evaluación y las salidas son las modificaciones o mejoras al proceso analizado.

**Extensión de la metodología PM²**

La extensión de metodología PM² fue probada con el proceso de contratación de servicios de una empresa publicitaria, con la finalidad de analizar por qué existen anulaciones del servicio por parte de los clientes cuando ya lo tienen contratado. PM² se encuentra conformada por las siguientes seis etapas (Silva Osses et al., 2016):

- *Planificar*. Esta etapa tiene como finalidad entender el problema y el proceso seleccionado de la organización, para lo cual hay que estudiar el negocio y los datos necesarios para crear el registro de eventos. Además, en esta fase se

identifican las preguntas de investigación y el equipo de trabajo que sea capaz de responder las interrogantes planteadas (analistas de proceso y expertos del negocio).

- *Extraer.* En esta etapa, se determina el alcance de los datos y, posteriormente, estos se extraen de los diferentes SI ya filtrados.
- *Procesar.* En la tercera etapa se crean las vistas necesarias de los registros de eventos para responder las preguntas de investigación. En muchas ocasiones se pueden agregar los datos para enriquecer el registro de eventos y, con ello, generar más información cuando se realiza el análisis.
- *Analizar.* En esta etapa se puede realizar el análisis de los datos a través del enfoque de la minería de datos y de la MP de manera no simultánea. Ello para que cada resultado obtenido pueda servir para mejorar el análisis de la otra disciplina propuesta.
  Cualquier técnica de minería de datos utilizada será válida cuando aporte valor al análisis. En cambio, todas las técnicas de MP tendrán que ser aplicadas con la finalidad de descubrir el proceso, hacer el análisis de conformidad adecuado y mejorar dicho proceso.
- *Evaluar*. Esta fase consiste en verificar y validar los resultados obtenidos con la información original y con los elementos clave del proceso. Con base en ello, será posible vislumbrar qué ideas son potenciales para mejorar el proceso.
- *Mejorar*. En esta fase se implementan las mejoras detectadas para el proceso y se da soporte.

**Guía de análisis para la MP enfocada en el usuario**

Esta guía de análisis para la MP propuesta en el trabajo de investigación de Céspedes et al. (2018) no muestra información de su aplicación en algún caso de estudio, pero fue desarrollada bajo los principios del diseño centrado en el usuario y la norma ISO 9241-210:2019. Esta guía se encuentra dividida en cinco etapas iterativas, que permiten —cuando sea necesario— repetir pasos hasta obtener el resultado deseado. El usuario interviene como actor principal en cada una de sus fases, las que se describen a continuación:

- *Análisis del contexto.* Esta etapa inicia estableciendo los objetivos, el contexto de uso y la planificación del proyecto. Por ello, es importante identificar a los usuarios (por ejemplo, proveedores de datos, especialistas del negocio, propietarios del negocio, etcétera) y conocer su relación con el proyecto, así como el entorno tecnológico a nivel de *hardware* y *software* en que se desarrollará. Lo anterior es, pues, el punto de inicio para establecer las necesidades de los usuarios y concretar los elementos de la solución.

- *Análisis de eventos.* En esta fase se identifican los datos (significativos y de calidad) del proceso a analizar, a través de la recolección de las fuentes de datos disponibles, debido a que en muchas ocasiones los datos se encuentran almacenados en varios SI.
- *Preparación de eventos.* La finalidad de esta fase es realizar el procesamiento, limpieza y depuración de los datos para conseguir un registro de eventos apropiado y poder aplicar las técnicas de MP. Conjuntamente, se determinan las variables significativas en relación con los objetivos de la investigación.
- *Identificación de patrones*. La primera actividad que se lleva a cabo en esta fase es elegir las técnicas y herramientas de análisis de procesos que ayuden a cumplir los objetivos planteados. La segunda actividad es identificar la finalidad del análisis, que podría ser: (a) descubrir el modelo de control de flujo; (b) verificar la conformidad; (c) realizar un análisis de rendimiento; (d) hacer un análisis organizacional del proceso; y (e) construir el modelo completo del proceso.
- *Resultados y despliegue.* En esta fase se transfieren a la organización los resultados de interés a través de una presentación clara y útil, de acuerdo con las perspectivas y necesidades de los usuarios. Los resultados deberán incluir la propuesta de acciones de mejora y el rediseño del proceso.

### *2.2 Análisis comparativo de las metodologías encontradas*

Con el propósito de agrupar las actividades descritas en las metodologías mencionadas en el apartado anterior, se decidió definir una metodología general de seis etapas. Se observó que, en muchos casos, las actividades de varias fases en una metodología son agrupadas en una misma fase en otra metodología, o, en otras ocasiones, una metodología considera una sola fase, que en otra es divida en varias etapas (véase la Tabla 1).

A continuación, se enumeran las seis etapas de la metodología general, junto con las actividades involucradas en cada una de ellas:

- **Etapa A. Planeación y alcance.** Involucra el alcance, la justificación y planeación del proyecto, así como la identificación de usuarios, objetivos, metas, preguntas de investigación y recursos (humanos y técnicos), especificando la relación que tienen estos últimos con el proyecto.
- **Etapa B. Ubicación y preprocesamiento de datos.** En esta etapa se ubican, se describen y se exportan los datos que se encuentran dentro de los SI, incluyendo su definición, alcance y calidad.
- **Etapa C. Procesamiento de datos.** En esta etapa se realiza la selección, limpieza y depuración de los datos. Además, cuando es preciso, la inserción y creación de atributos o variables al registro de eventos.

- **Etapa D. Minería y análisis de procesos.** En esta etapa se eligen las técnicas y herramientas de la MP y se aplican para obtener los modelos y la información de la conformidad, el rendimiento y el cumplimiento del proceso, para responder las preguntas de investigación. En caso sea necesario, en esta fase se extienden los modelos o la información a través de otras perspectivas o herramientas (por ejemplo, análisis visuales y minería de datos).
- **Etapa E. Presentación de resultados.** En esta etapa se redacta y se presenta el informe a los diferentes usuarios, con la finalidad de comunicar los resultados y las recomendaciones, modificaciones o acciones de mejora que se podrían llevar a cabo en el proceso.
- **Etapa F. Transferencia y seguimiento.** En esta etapa se implementan las recomendaciones o mejoras a la ejecución real del proceso (presentadas en la etapa anterior) y se soportan sus operaciones.

**Tabla 1**

*Ubicación de cada fase de las metodologías analizadas dentro de la metodología genérica*

| Nombre de la metodología | Etapas de la metodología genérica | | | | | |
|---|---|---|---|---|---|---|
| | Etapa A | Etapa B | Etapa C | Etapa D | Etapa E | Etapa F |
| | Ubicación de las fases de cada metodología analizada | | | | | |
| PDM | ---- | Fase 1 Preparación del registro | Fase 2 Inspección del registro | Fases 3, 4 y 5 Análisis de control de flujo; análisis de rendimiento; y análisis de roles | Fase 6 Transferencia de resultados | --- |
| Ciclo de vida L* | Fase 0 Justificación y planeación | Fase 1 Extracción | Fase 1 Extracción | Fases 2 y 3 Creación del modelo de control de flujo; y creación del modelo del proceso integrado | Fase 4 Soporte operacional | ---- |
| PMPM | Fase 1 Alcance | Fase 2 Comprensión de datos | Fase 3 Creación del registro de eventos | Fases 4 y 5 Minería de procesos; y evaluación | Fase 6 Despliegue | ---- |
| PMMF | ---- | Fase 1 Preparación de datos | Fase 2 Exploración de los datos | Fases 3 y 4 Perspectiva y análisis | Fase 5 Resultados | ---- |
| PM$^2$ | Fase 1 Alcance | Fase 2 Extracción | Fase 3 Procesamiento de datos | Fases 4 y 5 Minería y análisis; y evaluación | Fase 5 Evaluación | Fase 6 Mejoras y soporte de proceso |

(*continúa*)

(*continuación*)

| Nombre de la metodología | Etapas de la metodología genérica | | | | | |
|---|---|---|---|---|---|---|
| | Etapa A | Etapa B | Etapa C | Etapa D | Etapa E | Etapa F |
| | Ubicación de las fases de cada metodología analizada | | | | | |
| Extensión de la metodología PM² | Fase 1 Planificar | Fase 2 Extraer | Fase 3 Procesar | Fase 4 Analizar con minería de datos y procesos | Fase 5 Evaluar | Fase 6 Mejorar |
| Guía de análisis para la MP enfocada en el usuario | Fase 1 Análisis del contexto | Fase 2 Análisis de eventos | Fase 3 Procesamiento de eventos | Fase 4 Identificación de patrones | Fase 5 Resultados y despliegue | ---- |

En base a la información presentada en cada metodología y en la Tabla 1, se puede mencionar lo siguiente:

- Las metodologías PDM y PMMF no incluyen la fase de planeación, por lo que, para iniciar un proyecto de MP con estas metodologías, no es necesario tener un entendimiento total del proceso ni del negocio. Estas metodologías pueden ser empleadas en proyectos basados en datos.
- A pesar de que en la metodología PMPM se especifica la definición de un alcance, no se sugiere realizar detalladamente una planeación, lo cual podría ser una limitante en comparación con las metodologías que inician con esta fase.
- Solo las metodologías PM² y la extensión PM² incluyen la fase de implementación de las recomendaciones o mejoras y el servicio de soporte al proceso.
- En la guía de análisis para la MP enfocada en el usuario se plantea involucrar a los diversos usuarios de la empresa (por ejemplo, proveedores de datos, especialistas del negocio, propietarios del negocio, especialista en el proceso, científicos de datos, analistas de datos, etcétera) en todas las fases. Sin embargo, no se explica cómo intervienen en las actividades y cuáles son sus funciones específicas.
- En ninguna de las metodologías analizadas se muestran actividades que puedan ayudar a los usuarios a elegir y utilizar, según sea el caso, las herramientas (redes Petri, árboles de proceso, modelos BMPN, etcétera) y técnicas (minado Alpha, heurístico, *fuzzy*, etcétera) existentes de la MP.
- Las metodologías PDM y PMMF son de carácter exploratorio, pues se centran en los datos y no involucran a los usuarios.
- La mayoría de las metodologías analizadas fueron probadas por sus propios autores a través de un caso de estudio dentro del mismo proyecto de

investigación (Tabla 2), por lo que es conveniente analizar y evaluar su pertinencia a través de otros casos de estudio externos.

- Solo se recomienda utilizar el ciclo de vida L* cuando se desarrollan proyectos de MP estructurados. De lo contrario, solo se podrán llevar a cabo las tres primeras fases de esta metodología.
- Las metodologías no incluyen una fase cuyo objetivo sea hacer divulgación científica a través de alguna publicación que dé a conocer los hallazgos tecnológicos y conocimientos encontrados.

**Tabla 2**

*Aplicación y validez de las diferentes metodologías dentro del mismo proyecto de investigación*

| Nombre de la metodología | Aplicada en un caso de estudio | Validada por sus propios autores | Hasta qué fase de la metodología se desarrolló el caso de estudio |
|---|---|---|---|
| PDM | Sí | Sí | Transferencia de resultados |
| Ciclo de vida L | No | No | --- |
| PMPM | Sí | Sí | Despliegue |
| PMMF | Sí | Sí | Resultados |
| $PM^2$ | Sí | Sí | Evaluación |
| Extensión de la metodología $PM^2$ | Sí | Sí | Evaluación |
| Guía de análisis para la MP enfocada en el usuario | No | No | ---- |

En base a las excepciones mencionadas anteriormente, se considera que existe la oportunidad de desarrollar una metodología más robusta, que involucre las principales fases del desarrollo de proyectos de MP definidas en las metodologías estudiadas y que pueda aplicarse a distintos tipos de proyectos (ya sea desde un punto de vista empresarial o uno científico-académico), a fin de mejorar el rendimiento de sus procesos o el cumplimiento de las normas y reglamentos. Es importante tener en cuenta que los objetivos de un proyecto de MP pueden ser muy concretos (lograr una reducción de costos del 10 % para un determinado proceso, por ejemplo) o muy genéricos (obtener información valiosa sobre el desempeño del proceso).

## 3. RESULTADOS

El resultado obtenido después de tomar en cuenta las actividades involucradas en las diferentes metodologías encontradas, el análisis realizado y la experiencia de los investigadores, fue el diseño de una metodología de MP que pueda ser aplicada en el desarrollo de proyectos tanto del ámbito empresarial como del ámbito científico-académico, con sus propias características particulares (ver figuras 2 y 3).

**Metodología propuesta**

La metodología propuesta se encuentra conformada por las siguientes fases:

***Fase 1. Planeación y alcance.*** Esta fase debe estar a cargo de los usuarios especialistas del negocio (encargados de los procesos de negocios), del experto en sistemas (familiarizado con los SI) y del analista de procesos (experto en aplicar la MP), quienes conforman el equipo de trabajo. Este equipo responderá las interrogantes planteadas, las cuales podrían estar enfocadas a mejorar el rendimiento de un proceso comercial o a verificar su cumplimiento con respecto a ciertas reglas y regulaciones. Esta fase incluye las siguientes actividades:

- Identificar el o los procesos que se quieren analizar y cuyo funcionamiento se quiere comprender.
- Identificar los principales problemas del o los procesos a analizar.
- Definir el alcance del proyecto a través de los objetivos y, posteriormente, convertirlos en preguntas de investigación, que pueden estar relacionadas, por ejemplo, con tiempo, recursos, costos, etcétera.
- Planificar el análisis.
- Identificar las herramientas y técnicas de la MP disponibles, para realizar de manera adecuada los análisis correspondientes.

En los casos en que se quiera realizar un proyecto de MP de tipo académico o de investigación y se tenga acceso a los datos del proceso, pero no el contacto directo con los expertos en el negocio, esta fase no será necesaria, ya que se estaría llevando a cabo un proyecto de carácter exploratorio.

***Fase 2. Preprocesamiento de datos.*** El objetivo de esta fase es ubicar los datos en los SI, comprenderlos (tal como se encuentran almacenados) y extraer suficientes (representativos en base al rendimiento esperado y balanceados dentro del periodo de tiempo proporcionado), conforme a los objetivos definidos, para llevar a cabo el análisis y los diagnósticos significativos a través de la MP.

Al extraer los datos, se debe evitar que falten, que sean incorrectos, imprecisos o irrelevantes, porque pueden afectar los resultados. Por ejemplo, la ausencia de identificadores únicos que vinculan todos los eventos relacionados dificulta la creación del registro de eventos, o los datos de tiempo imprecisos para cada evento afectan los resultados de las mediciones de rendimiento.

Es recomendable, por tanto, buscar y hacer uso de estrategias y herramientas tecnológicas que permitan extraer los datos de los SI de manera adecuada, ya que existen muchas formas de almacenamiento (los datos, por ejemplo, se pueden encontrar en bases de datos relacionales, orientadas a objetos, documentales, etcétera). Posteriormente, en la siguiente fase, los datos obtenidos serán convertidos en un registro de eventos.

La metodología divide la extracción de datos, y la creación y el procesamiento de registros en dos etapas (preprocesamiento de datos y procesamiento de datos, respectivamente), debido a que la extracción de datos de eventos requiere mucho tiempo y se repite con menos frecuencia que las actividades de procesamiento de datos, tales como el filtrado y la creación de diferentes vistas sobre los mismos datos.

***Fase 3. Procesamiento de datos.*** En esta fase se crean los registros de eventos a través de la limpieza y depuración de los datos. Para ello, se hace la identificación y creación de eventos (por ejemplo, unir los atributos hora y fecha en un solo evento, para obtener la marca de tiempo) y, cuando es preciso, se eliminan casos incompletos, actividades repetidas o se agrupan casos similares con la utilización de filtros. Lo anterior, con la finalidad de obtener registros de eventos relevantes para las siguientes fases.

Para que pueda ser utilizado por la MP, el archivo debe contener como mínimo el ID del caso, la actividad y la marca de tiempo del evento. Este archivo, en primera instancia, puede tener las extensiones CSV o XLS, y a través de PROM, Disco, Apromore u otra aplicación de MP, puede ser transformado al formato de registro de eventos XES o MXML.

Posteriormente, se analiza el registro de eventos para identificar las características de las actividades, el número de casos y eventos, los eventos iniciales y finales, las visualizaciones iniciales del proceso, etcétera. Este análisis se puede llevar a cabo con la MP y con la ayuda de otras técnicas estadísticas o visuales.

En la práctica, en este punto se pueden presentar situaciones en las que se observa la falta de datos para realizar el análisis completo según el alcance del proyecto, por lo que será necesario regresar a la fase anterior, complementar los datos faltantes y así tener un conjunto de datos satisfactorio.

En las siguientes tres etapas se aplicarán las técnicas y herramientas de la MP en los registros de eventos, con el propósito de obtener información —sobre el rendimiento y el cumplimiento de los procesos— que ayudará a responder las preguntas de investigación. Si no existieran preguntas de investigación definidas porque es un proyecto orientado a datos, se pueden aplicar técnicas exploratorias combinadas con el descubrimiento de procesos para obtener una visión general del proceso. Para los casos en donde sí se definieron preguntas concretas de investigación, los análisis pueden centrarse en responder dichas preguntas.

***Fase 4. Análisis de control de flujo.*** En esta fase se aplican las técnicas de descubrimiento de procesos para conocer cómo es el proceso real, para luego compararlo con la documentación de la organización y verificar que los datos incluidos en el registro de eventos puedan ser representados en el proceso descubierto de manera adecuada, y con ello juzgar la calidad del modelo, identificar casos divergentes, encontrar desviaciones en el proceso y dar respuestas a algunas preguntas planteadas en la fase 1.

Algunos de los minados que se pueden utilizar para descubrir los modelos reales de los procesos que se están analizando son: Alpha, Alpha ++, Heurístico, ETM, Fuzzy e inductivo. Estos pueden ser usados a través de diversos *plugins* contenidos en el *framework* de ProM.

***Fase 5. Análisis de rendimiento.*** En esta fase se utilizan los modelos de los procesos obtenidos en la fase anterior para examinar el desempeño de proceso, el tiempo de rendimiento de las actividades y los cuellos de botella, con el fin de encontrar áreas de mejora que apoyen en la toma de acciones.

Un *plugin* del *framework* ProM, bastante completo, que puede ser utilizado para obtener el rendimiento del proceso es *multi-perspective process explorer*. Sin embargo, existen otras aplicaciones más amigables (como Disco y Apromore) que también pueden usarse para obtener modelos en los que se muestra el rendimiento de los procesos.

***Fase 6. Análisis de roles.*** Esta fase se lleva a cabo siempre que el registro de eventos contenga datos sobre los recursos que ejecutan tales eventos. Se analizan los roles que desempeñan los miembros del equipo de trabajo involucrados en el proceso, para detectar quién ejecuta cada actividad y para explorar su productividad.

Para el caso en que no se tenga información acerca de los roles de los empleados, pero sí, por ejemplo, los nombres de quienes intervienen en cada actividad del proceso, esta fase se podría llevar a cabo con otros tipos de análisis: por ejemplo, un análisis de intervenciones del personal en el proceso o un mapa de delegación de trabajo, en los que se pueda observar —aparte del número de casos atendidos por el personal— la colaboración que existe entre los recursos dentro del proceso.

Los análisis mencionados se pueden llevar a cabo, por ejemplo, con los *plugins* de *social network* contenidos en el *framework* de ProM o con las aplicaciones Disco y Apromore.

Lo anterior puede ayudar a los especialistas del negocio a decidir sobre la asignación de los trabajos de cada integrante del equipo.

***Fase 7. Presentación de resultados.*** En esta fase se realiza la descripción de las actividades llevadas a cabo en el proyecto de MP. El objetivo es presentar el informe final a los usuarios interesados. Es de particular interés la presentación de la metodología utilizada para llevar a cabo la MP, la descripción de las diferentes etapas y los hallazgos encontrados (por ejemplo, el modelo de proceso descubierto, las ejecuciones anormales, los tiempos de rendimiento, etcétera). Finalmente, es fundamental que se presenten las recomendaciones o acciones de mejora que se pueden incluir en el proceso.

Antes de presentar los resultados, es conveniente que los expertos en los procesos participen en su revisión y validación, ya que los mineros regularmente no son expertos en el dominio de los procesos que se están analizando.

***Fase 8. Publicación de resultados.*** En esta etapa están consideradas todas las acciones necesarias para publicar los resultados en algún medio de difusión. Es recomendable aplicarla cuando se realizan proyectos exploratorios basados en datos, cuando los usuarios son nuevos en la MP o cuando tienen un perfil académico. En este tipo de proyectos

se investigan las posibilidades de aplicación de la MP, por lo que resulta un ejercicio interesante la difusión de los resultados, de las experiencias en la aplicación de esta tecnología, así como la comparación con otras formas de trabajo reportadas o conocidas. Esta es una etapa pensada para cumplir con el objetivo de divulgación científica; es decir, promover, entre la sociedad, los hallazgos tecnológicos y conocimientos obtenidos. Aun cuando, desde el punto de vista de las organizaciones, estas acciones pueden ser consideradas opcionales, el llevarlas a cabo podría fomentar la vinculación y colaboración entre el sector empresarial y la comunidad científica, así como la creación de nuevas oportunidades de negocio basadas en el conocimiento.

***Fase 9. Transferencia y seguimiento.*** En esta fase se implementan las recomendaciones o acciones de mejora a la ejecución real del proceso y se soportan sus operaciones; esto es, continuamente se debe medir y evaluar, con la finalidad de detectar cualquier falla o comportamiento anormal.

**Figura 2**

*Diagrama de bloques de la metodología de MP para el desarrollo de proyectos de tipo empresarial y científico-académico*

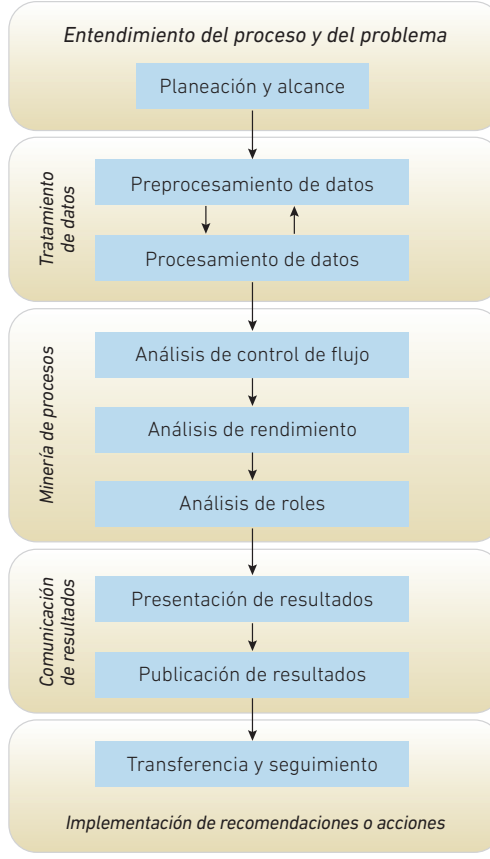
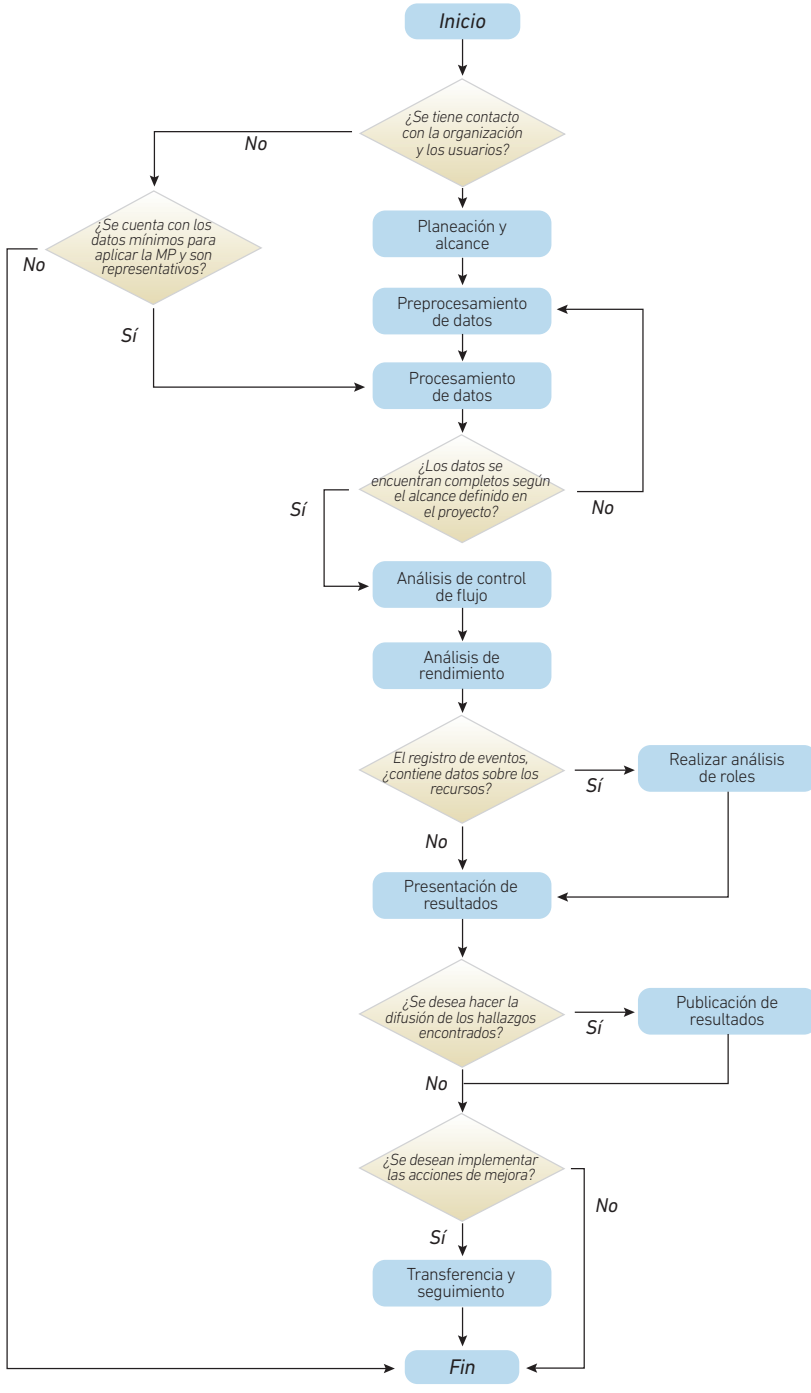


**Figura 3**

*Diagrama de flujo de la metodología de MP para el desarrollo de proyectos de tipo empresarial y científico-académico*



## 4. CONCLUSIONES

La MP, a través de sus técnicas, permite a las organizaciones realizar un escaneo de sus procesos de negocio para comprobar con mayor rigurosidad su cumplimiento, diagnosticar problemas e identificar soluciones que faciliten acciones de mejora o rediseño y que, al ser implementadas, logren procesos de negocio más eficaces. Por esos motivos, se considera importante que todo proyecto de MP se lleve a cabo con un enfoque metodológico que sirva de guía para que se alcancen los objetivos planteados y se garanticen resultados.

La metodología propuesta en esta investigación es una opción que permite orientar al usuario a desarrollar proyectos de MP, tanto de tipo empresarial como científico-académico. Esta tiene nueve fases: planeación y alcance, preprocesamiento de datos, procesamiento de datos, análisis de control de flujo, análisis de rendimiento, análisis de roles, presentación de resultados, publicación de resultados, y transferencia y seguimiento.

La metodología diseñada incluye todas las actividades principales de la MP. Inicia con la identificación y comprensión del proceso de negocio que se desea estudiar. Con esta base, se realiza la planificación y definición de los objetivos y preguntas de investigación, para que después los datos se puedan ubicar, extraer, limpiar y convertir al formato adecuado, antes de desarrollar las fases que involucran la MP. Posteriormente, se realiza la presentación de los hallazgos encontrados, junto con las recomendaciones o acciones de mejora que se pueden llevar a cabo en el proceso.

Finalmente, la etapa de publicación de resultados se halla junto a la etapa de transferencia y seguimiento.

La publicación de resultados es una fase pensada principalmente para los usuarios nuevos en la MP o de perfil académico. Tiene como finalidad difundir, a través de una publicación, los hallazgos tecnológicos y conocimientos obtenidos. Aun cuando desde el punto de vista de las organizaciones estas acciones son opcionales, el llevarlas a cabo podría fomentar la vinculación y colaboración entre el sector empresarial y la comunidad científica, así como la creación de nuevas oportunidades de negocio basadas en el conocimiento.

Finalmente, la fase de transferencia y seguimiento tiene como actividad principal la implementación de las recomendaciones o acciones de mejora, así como su seguimiento.

## 5. FUTURAS LÍNEAS DE INVESTIGACIÓN

En trabajos futuros se recomienda aplicar la metodología propuesta de MP a diversos casos de estudio, tanto del área empresarial como del ámbito científico-académico. El objetivo de estos trabajos de investigación deberá estar enfocado en analizar la eficiencia

de la metodología, la facilidad y pertinencia de su aplicación, así como la congruencia entre las fases que la integran.

## REFERENCIAS

Aguirre, M. H. & Rincón, G. N. (2015). Minería de procesos: desarrollo, aplicaciones y factores críticos. *Cuadernos de administración, 28*(50), 137-157. https://doi.org/10.11144/Javeriana.cao28-50.mpda

Badakhshan, P., Wurm, B., Grisold, T., Geyer-Klingeberg, J., Mendling, J., & Brocke, J. (2022) Creating business value with process mining. *The Journal of Strategic Information Systems*, *31*(4), 101745. https://doi.org/10.1016/j.jsis.2022.101745

Bozkaya, M., Gabriels, J., & Van der Werf, J. M. (2009, febrero). *Process diagnostics: a method based on Process Mining.* [Ponencia]. International conference on information, process, and knowledge management, Cancun, México, 22-27. https://doi.org/10.1109/eKNOW.2009.29

Butt, N., Mahmood, Z., Sana, M.U., De La Torre, I., Castanedo, J., Brie, S., & Ashraf, I. (2023) Behavioral and performance analysis of a real-time case study event log: a process mining approach. *Applied sciences*, *13*(7), 4145, 1-21. https://doi.org/10.3390/app13074145

Céspedes, Y., Molero, G., & Arieta, P. (2018). Diseño de una guía de análisis para la minería de procesos enfocada en el usuario. *Ciencias de la información, 49*(3), 26-33. https://biblat.unam.mx/hevila/Cienciasdelainformacion/2018/vol49/no3/4.pdf

Checoli, A., Vecino, D., Scalabrin, E., & Portela, E. (2020). An extended model for remaining time prediction in manufacturing systems using process mining. *Journal of manufacturing systems, 56*, 188-201. https://doi.org/10.1016/j.jmsy.2020.06.003

De Weerdt, J., Schupp, A., Vanderloock, A., & Baesens, B. (2013). Process mining for the multi-faceted analysis of business processes. A case study in a financial services organization. *Computers in industry, 64*(1), 57-67. https://doi.org/10.1016/j.compind.2012.09.010

Dos Santos, C., Meincheim, A., Faria, E., Rosano, M., Vecino, D., Ribeiro, D., Portela, E., & Scalabrin, E. (2019). Process mining techniques and applications. A systematic mapping study. *Expert system with Applications*, *133*, 260-295. https://doi.org/10.1016/j.eswa.2019.05.003

Emamjome, F., Andrews, R., & ter Hofstede, A. (2019, octubre). A case study lens on Process Mining in practice. In H. Panetto, C. Debruyne, M. Hepp, D. Lewis, C. Ardagna & R. Meersman (Eds.) *On the move to meaningful Internet systems: OTM 2019 Conferences*, *11877.* Springer. https://doi.org/10.1007/978-3-030-33246-4_8

Fuentes, S., Domínguez, A., García, W., Romero, P., & Leyva, L. (2019) Caracterización de la producción científica en el área disciplinar de la minería de proceso. *Investigación bibliotecológica, 33*(78), 193-216. http://dx.doi.org/10.22201/iibi.24488321xe.2019.78.57925

González González, A., Leal Rodríguez, L., Martínez Caballero, D., & Morales Fonte, D. (2019). Herramientas para la gestión por procesos. *Cuadernos Latinoamericanos de Administración*, 15(28). https://www.redalyc.org/articulo.oa?id=409659500003

Martínez-Escobar, A., Silega, N. & Noguera-García, M. (2021). Aplicación de minería de procesos para la mejora de los servicios públicos. *Revista cubana de transformación digital, 2*(4), 92-103. https://rctd.uic.cu/rctd/article/view/150

Merchán E., Mero K., & Mero C. (2021). Técnicas aplicadas a la minería de proceso: revisión sistemática. *Serie Científica de la Universidad de las Ciencias Informáticas, 14* (9), 148-162. https://publicaciones.uci.cu/index.php/serie/article/view/950

Morales, A., Martínez, C., Amador, J., Hidalgo, C., & García R. (2022). Minería de procesos aplicada a un sistema de solicitudes de servicios al cliente: un caso de estudio basado en datos. *Revista Ibérica de Sistemas e Tecnologias de Informação,* (E52), 117-132. Https://www.proquest.com/openview/e70e87c918af4fcefcdf912504e6fd67/1?pq-origsite=gscholar&cbl=1006393

Park, S., & Sik, Y. (2016). A study of Process Mining-based business process innovation. *Procedia computer science, 9*, 734-743. https://doi.org/10.1016/j.procs.2016.07.066

Sangil, M. (2020, noviembre). *Heuristics-based process mining on extracted Philippine public procurement event logs* [Ponencia]. 7th International conference on behavioral and social computing (BESC), Bournemouth, Reino Unido, pp.1-4. https://doi.org/10.1109/BESC51023.2020.9348306

Silva Osses, A., Arias, M., Quelves, L., Rojas, E., Fernández, B., Muñoz-Gama, J., & Sepúlveda, M. (2016, octubre). Business process analysis in advertising: an extension to a methodology based on process mining projects. [Ponencia]. 35th International Conference of the Chilean Computer Science Society (SCCC), Valparaiso, Chile, 1-12. https://doi.org/10.1109/SCCC.2016.7836000

Silva Osses, A. (2017). *Metodología para el análisis de proceso de negocio basada en minería de procesos y datos*. [Memoria de Titulación, Universidad Técnica Federico Santa María de Chile]. Repositorio institucional de la UTFSM. Https://repositorio.usm.cl/bitstream/handle/11673/14074/3560902038214UTFSM.pdf?sequence=1&isAllowed=y

Terragni, A., & Hassani, M. (2018, agosto). Analyzing customer journey with process mining: from discovery to recommendations. [Ponencia]. IEEE 6th International

Conference on Future Internet of Things and Cloud (Ficloud), Barcelona, España, 224-229. https://doi.org/10.1109/FiCloud.2018.00040

Van der Aalst, W. (2011). *Process mining: data science in action*. Springer.

Van der Aalst, W. (2012). Process mining: overview and opportunities. *ACM Transactions on management information systems, 3*(2), 1-17. https://doi.org/10.1145/2229156.2229157

Van der Aalst, W., Adriansyah, A., Alves de Medeiros, A., Arcieri, F., Baier, T., Blickle, T., Chandra, J., Van den Brand, P., Brandtjen, R., Buijs, J., Burattin, A., Carmona, J., Castellanos, M., Claes, J., Cook, J., Costantini, N., Curbera, F., Damiani, E., De Leoni, M., ...Wynn, M. (2012). Process mining manifesto. En F. Daniel, K. Barkaoui y S. Dustdar (Eds.) Business process management workshops (BPM 2011). *Lecture notes in business information processing, 99.* Springer. https://doi.org/10.1007/978-3-642-28108-2_19

Van der Aalst, W. (2013). Service mining: using process mining to discover, check, and improve service behavior. *IEEE Transactions on services computing*, *6*(4), 525-535. https://doi.org/10.1109/TSC.2012.25

Van der Aalst, W. (2016). *Process mining: data science in action* (2nd ed.). Springer

Van der Heijden, T. (2012). *Process mining project methodology: developing a general approach to apply process mining in practice* [Tesis de maestría, Eindhoven University of Technology]. https://research.tue.nl/en/studentTheses/process-mining-project-methodology

Van Eck, M. L., Lu, X., Leemans, S. J., & Van der Aalst, W. (2015, junio). *PM²: a process mining project methodology*. [Ponencia]. International conference on advanced information systems engineering, Estocolmo, Suecia. https://doi.org/10.1007/978-3-319-19069-3_19

# RENACYT y las brechas de género en carreras STEM en el Perú

Rosa Flor Gomez Risco
rgomezr@unp.edu.pe
http://orcid.org/0000-0003-3738-9729
Universidad Nacional de Piura, Perú

Recibido: 28 de septiembre del 2023 / Aceptado: 1 de junio del 2024
doi: https://doi.org/10.26439/interfases2024.n19.6685

RESUMEN. Este trabajo pretende determinar las brechas de género en las carreras STEM (las de ciencias, tecnología, ingeniería y matemáticas) en el Perú, para contribuir a fomentar la igualdad de oportunidades en la sociedad. El estudio presenta estadísticas que comparan la presencia de mujeres con la de hombres tomando en cuenta sus niveles educativos, grupos de edad y años de inicio en estas carreras, lo que sirve como base para desarrollar maneras de cerrar estas brechas. Se realizó un análisis descriptivo con información recopilada de las bases de datos del Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC) con el fin de comprender mejor la situación de los profesionales que forman parte del Registro Nacional Científico, Tecnológico y de Innovación Tecnológica (RENACYT) en el Perú. A pesar de los notables avances logrados en las últimas décadas, persisten numerosos obstáculos que dificultan la plena integración de las mujeres en estos campos. La medición de estas brechas de género en ciencia, tecnología e innovación es un desafío, pues existen escasos datos e indicadores disponibles a nivel internacional para analizar estos fenómenos.

PALABRAS CLAVE: brechas / género / carreras STEM

## RENACYT AND GENDER GAPS IN PERU'S STEM CAREERS

ABSTRACT. This study aims to identify gender gaps in Peru's STEM careers (science, technology, engineering, and mathematics) to promote equal opportunities in society. It presents statistics comparing the presence of women and men, considering their educational levels, age groups, and years of entry into these careers. This data serves as a basis for developing strategies to close these gaps. The research involved a descriptive analysis of data from the National Council for Science, Technology, and Technological Innovation (CONCYTEC) to understand the situation of professionals in the National Registry of Scientific, Technological, and Innovation Personnel (RENACYT) in Peru. Despite significant progress in recent decades, many obstacles still hinder the full integration of women into these fields. Measuring these gender gaps in science,

technology, and innovation poses a challenge due to the limited data and indicators available internationally.

KEYWORDS: gaps / gender / STEM careers

## 1. INTRODUCCIÓN

La presente investigación tuvo como objetivo identificar las brechas de género que existen en carreras STEM entre los profesionales RENACYT en el Perú, puesto que la reducción de esas brechas es esencial para promover la igualdad de oportunidades en la sociedad. El estudio identificó disparidades en las cifras de mujeres en comparación con sus contrapartes masculinas en términos de niveles, grupo, rangos de edad y año de inicio como profesionales RENACYT. Esta identificación puede ayudar a implementar soluciones dirigidas a abordar estas brechas.

El CONCYTEC es la entidad líder dentro del Sistema Nacional de Ciencia y Tecnología e Innovación Tecnológica (SINACYT) y tiene la responsabilidad de liderar, promover, coordinar, supervisar y evaluar las actividades relacionadas con la ciencia, tecnología e innovación tecnológica en todo el territorio peruano. Además, orienta las iniciativas del sector privado y pone en marcha acciones de apoyo destinadas a estimular el avance en este campo (Carrasco & Valenzuela, 2021). Un total de 6514 profesionales (94,61 %) se han incorporado al RENACYT bajo el reglamento de 2018, mientras que solo 371 investigadores (5,39 %) fueron admitidos según el nuevo reglamento de 2021. La proporción de investigadoras es de 1 a 2, en comparación con los investigadores masculinos. Esta diferencia está relacionada con la cantidad de profesionales en cada género (De La Cruz-Cerrón et al., 2022).

A continuación, se abordará la metodología utilizada en el estudio, seguida por la presentación de resultados y un análisis detallado de dichos resultados. Finalmente, se extraerán conclusiones significativas y se propondrán áreas para futuras investigaciones.

## 2. METODOLOGÍA

En la presente contribución se propusieron las siguientes interrogantes: ¿cuáles son las brechas de género presentes en las carreras STEM en el Perú?, ¿cuáles son las brechas de género por rango de edad, por niveles de reconocimiento y por grupo de edad en cada nivel de reconocimiento? Para dar respuesta a ello, se analizaron los datos de 6885 profesionales RENACYT, de los cuales 2166 son mujeres y 4719 varones. Para la obtención de los datos, se accedió a la información del propio RENACYT, que está gestionado por el Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica del Perú (RENACYT, 2022). Conforme a sus políticas de privacidad y seguridad, todos los procedimientos se llevaron a cabo asegurando el cumplimiento normativo y la protección de datos personales durante todo el proceso de investigación. Se comparó a los profesionales de RENACYT por género, rangos de edad y niveles de reconocimiento. Estos niveles son categorías que distinguen a los investigadores peruanos según su trayectoria y contribuciones en el campo de la ciencia, tecnología e innovación; se establecen para reconocer y clasificar a los profesionales en función de sus logros, publicaciones,

proyectos y aportes al conocimiento. Los datos porcentuales muestran, en las siguientes figuras, las características mencionadas por género, y tales diferencias se respaldan a través de la prueba Chi cuadrado, utilizada para determinar si existe una diferencia significativa entre ambos géneros.

## 3. RESULTADOS

En la Figura 1 se observa que el grupo de profesionales inscritos en RENACYT más abundante es el de 40 a 49 años, seguido por los de 30 a 39 y, luego, por el grupo de 50 a 59 años. Se observa que siempre es mayor la cantidad de varones que de mujeres. El p-valor significativo (0,007) encontrado muestra las diferencias observadas entre varones y mujeres en la distribución por grupos de edad. Esto sugiere que ser profesional RENACYT podría influir en la participación y trayectoria en el ámbito científico y tecnológico, afectando de manera diferencial a varones y mujeres en distintos rangos de edad. Según Millones-Gómez et al. (2021), las universidades peruanas con políticas efectivas para fomentar la investigación y la producción científica tienden a atraer a un mayor número de profesionales hacia el campo de la ciencia y la tecnología. Estas políticas incluyen la financiación de la investigación y el reconocimiento del capital humano, lo que aumenta la posibilidad de participación en actividades científicas y la inscripción en RENACYT. Además, la presencia de investigadores con publicaciones en revistas indexadas indica la calidad y el compromiso de la institución con la investigación, e influye también en la inscripción de profesionales en RENACYT.

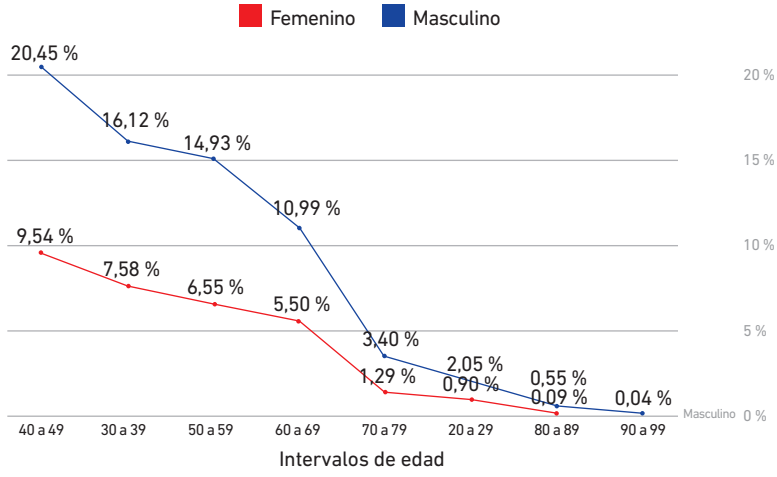
Rivera-Lozada et al. (2022) manifiestan que la edad promedio en la que los investigadores alcanzan su máximo nivel de productividad científica suele situarse entre los 40 y 50 años, momento en el que acumulan experiencia y liderazgo en proyectos de investigación. Esta tendencia podría estar reflejada en la mayor cantidad de inscripciones en el RENACYT en el grupo de 40 a 49 años. Además, los hallazgos en el presente estudio sugieren que las cohortes de edades más jóvenes, de 30 a 39 años, también muestran una presencia significativa en el RENACYT. Esto puede estar relacionado con políticas de incentivo a la investigación y el desarrollo profesional temprano en la carrera académica. La explicación para esta disparidad en la distribución por edades está relacionada con el momento en que los profesionales alcanzan su máximo potencial investigador y con la variabilidad en las oportunidades de investigación y el apoyo institucional ofrecido por las universidades peruanas (Carrasco & Valenzuela, 2021).

La presencia de profesionales del estudio en el grupo de 50 a 59 años se debe a una combinación de factores, como la estabilidad laboral y la dedicación continua a la investigación después de haber alcanzado cierto reconocimiento en sus campos respectivos, tanto en mujeres como en varones (Supo-Condori et al., 2020). Millones-Gómez et al. (2021) señalan que las políticas de investigación y financiamiento de proyectos pueden influir en la participación de investigadores jóvenes, principalmente varones,

en programas de registro como el RENACYT. Es importante destacar que la disparidad en la distribución por edades también puede reflejar diferencias en las oportunidades de carrera y en el acceso a recursos para la investigación entre diferentes grupos de edad. Según los hallazgos de Rivera-Lozada et al. (2022), los investigadores más jóvenes pueden enfrentar desafíos en términos de financiamiento y acceso a infraestructura de investigación, lo que podría influir en su participación en programas de registro como el RENACYT. Estas conclusiones resaltan la importancia de considerar el contexto más amplio en el que se desarrolla la actividad científica y tecnológica al interpretar los resultados de estudios sobre la distribución por edades de los investigadores en el RENACYT.

**Figura 1**

*Profesionales RENACYT por rango de edad*




*Nota*. Elaborado con datos tomados de: RENACYT, 2022, *Registro Nacional de Investigadores RENACYT*. Plataforma Nacional de Datos Abiertos. https://www.datosabiertos.gob.pe/dataset/registro-nacional-de-investigadores-renacyt/resource/629a1bde-a47f-4d6c-a383-dd70cb6fafc7

En la Figura 2 se aprecia que la distribución más amplia se registra en el primer y tercer nivel de reconocimiento (33,89 % y 32,65 %, respectivamente), en contraste con los niveles VII (0,77 %) y el de investigador distinguido (0,25 %). El primer nivel predominante está compuesto por investigadores con edades entre 40 y 49 años. Los resultados muestran una diferencia significativa en los niveles de reconocimiento RENACYT: un porcentaje notablemente bajo de investigadores, especialmente mujeres, alcanzan niveles superiores. Además, en cada nivel de reconocimiento, son siempre los varones quienes destacan en cantidad respecto de las mujeres. El p-valor encontrado (0,006) muestra las diferencias observadas entre varones y mujeres en la distribución por niveles de reconocimiento. Este fenómeno puede atribuirse a una serie de factores identificados en el estudio de Rivera-Lozada et al. (2022), por ejemplo, la producción científica está

estrechamente asociada con la inscripción en RENACYT, la obtención de un doctorado, la experiencia como asesor de tesis y la recepción de formación en investigación por parte de la universidad. No obstante, la presencia de barreras, como la carga de trabajo y las limitaciones económicas, particularmente significativas para las mujeres, actúan como obstáculos para la investigación. Estos hallazgos sugieren que las desigualdades de género y las limitaciones estructurales pueden incidir en la distribución desigual de los niveles de reconocimiento en RENACYT, contribuyendo a la brecha observada entre los sexos y a la representación desigual en los estratos más altos del sistema de reconocimiento académico.

Los niveles más altos, como el primer y tercer nivel, así como los niveles VII e investigador distinguido, están predominantemente compuestos por hombres, mientras que las mujeres están subrepresentadas en estos niveles. Estos hallazgos refuerzan la preocupación por la desigualdad de género en la ciencia y la tecnología; preocupación que ha sido abordada en estudios anteriores (Millones-Gómez et al., 2021; Supo-Condori et al., 2020). La explicación para esta brecha de género radica en las barreras sistémicas y culturales que enfrentan las mujeres en la academia y la investigación. Estas barreras pueden incluir sesgos de género en la evaluación y promoción académica, la ausencia de modelos femeninos a seguir en posiciones de liderazgo y la falta de políticas institucionales que apoyen la equidad de género en la investigación (Rivera-Lozada et al., 2022). Estas condiciones podrían contribuir a la menor representación de mujeres en los niveles más altos del RENACYT.

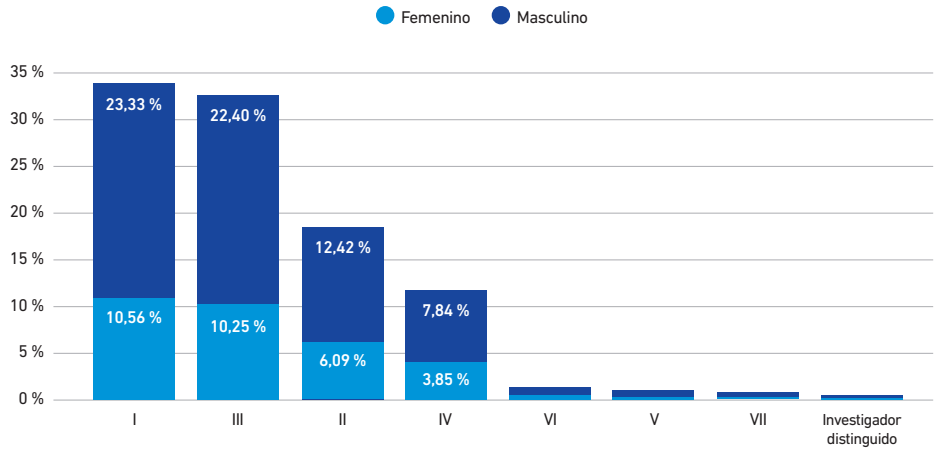
La predominancia de hombres en los niveles más altos del RENACYT, por tanto, se relaciona con factores socioeconómicos y culturales más amplios que perpetúan la desigualdad de género en la sociedad, como la división tradicional del trabajo en el hogar y las expectativas sociales de género, que pueden limitar las oportunidades de las mujeres para dedicarse plenamente a la investigación y la academia, afectando su progreso profesional y sus logros en el campo científico y tecnológico (Supo Condori et al., 2020).

Es esencial abordar estas desigualdades de género en la investigación y la academia mediante la implementación de políticas y programas que fomenten la igualdad de oportunidades y el acceso equitativo a recursos y reconocimiento profesional para mujeres investigadoras. Esto podría incluir iniciativas para promover la equidad de género en la evaluación y promoción académica, el apoyo a la maternidad y la paternidad equitativas, y el fortalecimiento de redes de apoyo y mentoría para mujeres en la ciencia y la tecnología (Millones-Gómez et al., 2021).

Se destaca la importancia de abordar las desigualdades de género en la investigación y la academia para promover un ambiente inclusivo y equitativo para todos los investigadores. Estas conclusiones subrayan la necesidad de acciones concertadas a nivel institucional y político para eliminar las barreras que impiden la plena participación y contribución de las mujeres en la ciencia y la tecnología en el Perú.

**Figura 2**

*Niveles de reconocimiento RENACYT*




*Nota.* Elaborado con datos tomados de: RENACYT, 2022, *Registro Nacional de Investigadores RENACYT*. Plataforma Nacional de Datos Abiertos. https://www.datosabiertos.gob.pe/dataset/registro-nacional-de-investigadores-renacyt/resource/629a1bde-a47f-4d6c-a383-dd70cb6fafc7

En la Figura 3 puede observarse que en el nivel I las mujeres representan 2,9 % y los varones 7,02 % del total. En el nivel III, en el que la mayoría de profesionales se encuentra en el rango de edades de 30 a 39 años, 3,24 % son mujeres y 6,03 % varones. En el nivel de investigador distinguido, la mayor parte tiene entre 40 a 49 años, con 0,01 % de mujeres y 0,15 % varones. Los datos adicionales revelan una disparidad significativa en la representación de género en diferentes niveles del RENACYT. Específicamente en el nivel I, en el que se espera que se encuentren los investigadores más destacados, las mujeres representan solo el 2,9 %, mientras que los hombres constituyen el 7,02 % del total. Esta brecha de género en el nivel inicial del RENACYT es preocupante y refleja desafíos persistentes en la inclusión y el reconocimiento de las mujeres en la investigación científica y tecnológica.

En el nivel III, en el que se espera que los profesionales se encuentren más establecidos en sus carreras, la mayoría tiene entre 30 y 39 años. Se observa que, también en este nivel, la representación de las mujeres es significativamente menor que la de los hombres: 3,24 % frente a 6,03 %. Estos hallazgos sugieren que persisten barreras para el avance profesional de las mujeres en la ciencia y la tecnología, a pesar de su experiencia y competencia (Millones-Gómez et al., 2021).

En el nivel de investigador distinguido, en el que se esperaría encontrar a los investigadores más destacados y experimentados, la mayoría se encuentra en el rango de edades de 40 a 49 años. Sin embargo, la representación de mujeres en este nivel es mínima, con solo un 0,01 %, en comparación con el 0,15 % de hombres. Esto evidencia la

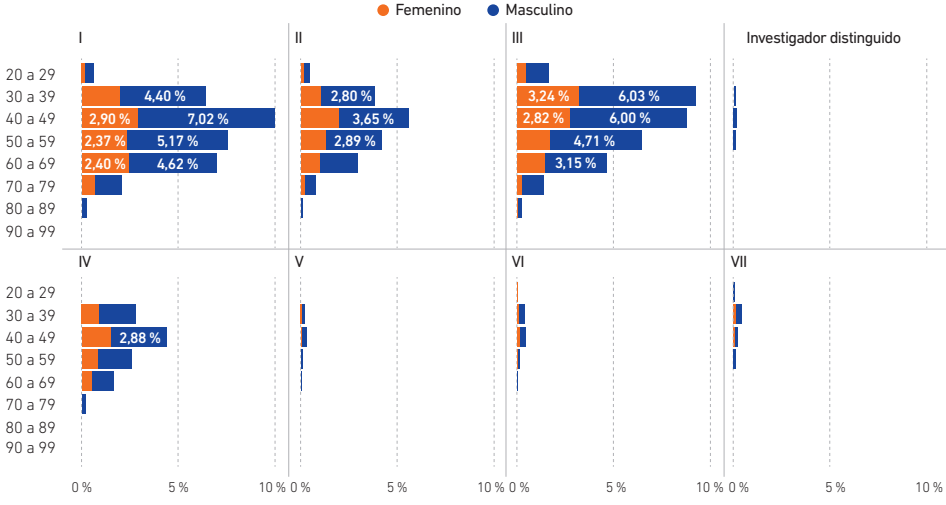
brecha de género en los niveles más altos del RENACYT y plantea interrogantes sobre las barreras que enfrentan las mujeres para alcanzar el reconocimiento y la promoción en la investigación científica y tecnológica (Rivera-Lozada et al., 2022).

Estos hallazgos sugieren que existen múltiples factores que contribuyen a la brecha de género en la representación en el RENACYT. Estos pueden incluir sesgos de género en la evaluación y promoción académica de acuerdo con el reglamento (CONCYTEC, 2021), la falta de modelos femeninos a seguir en posiciones de liderazgo y la persistencia de estereotipos de género en la cultura científica y tecnológica. Además, las responsabilidades familiares desproporcionadas —que recaen sobre todo en las mujeres— y las limitadas oportunidades de mentoría y redes de apoyo también pueden influir en la representación de género en los niveles superiores del RENACYT (Supo-Condori et al., 2020).

La brecha de género en la representación en el RENACYT es un problema complejo que requiere enfoques integrales y medidas políticas para abordar las barreras sistémicas y culturales que limitan el avance profesional de las mujeres en la ciencia y la tecnología en el Perú. Es fundamental adoptar políticas y programas que promuevan la equidad de género en la investigación y la academia, así como crear un entorno inclusivo que reconozca y valore las contribuciones de todos los investigadores, independientemente de su género.

**Figura 3**

*Niveles por grupo de edad en los profesionales RENACYT*




*Nota.* Elaborado con datos tomados de: RENACYT, 2022, *Registro Nacional de Investigadores RENACYT*. Plataforma Nacional de Datos Abiertos. https://www.datosabiertos.gob.pe/dataset/registro-nacional-de-investigadores-renacyt/resource/629a1bde-a47f-4d6c-a383-dd70cb6fafc7

## 4. DISCUSIÓN DE LOS RESULTADOS

La investigación sobre la brecha de género en las disciplinas STEM ha ganado relevancia en los últimos años, debido a su potencial impacto en la equidad y el progreso en el ámbito científico y tecnológico. Vargas-Solar (2022) destaca la necesidad de un enfoque interseccional para comprender la complejidad de esta brecha. El término "multifacético" resalta cómo múltiples factores, como la etnia, el origen socioeconómico y otros, interactúan y contribuyen a las desigualdades de género en STEM. López-Bassols et al. (2018) presentan una perspectiva regional, resaltan la necesidad de abordar las desigualdades desde una perspectiva global y destacan cómo factores culturales, políticos y económicos pueden influir en la participación de mujeres en STEM. Saavedra y Camarena (2020), por su parte, destacan cómo las desigualdades de género en México no solo afectan el ámbito profesional, sino también el empoderamiento de las mujeres en la sociedad.

El marco normativo delineado por CONCYTEC (2021) establece los parámetros regulatorios para la clasificación y promoción de investigadores en el RENACYT en Perú. Este marco proporciona la infraestructura legal necesaria para comprender las disparidades de género en el ámbito de la investigación científica y tecnológica. El estudio realizado por Supo-Condori et al. (2020) identifica una brecha de género significativa en la representación de investigadores inscritos en el RENACYT, especialmente en los niveles superiores de clasificación. Esta observación sugiere la existencia de obstáculos adicionales que enfrentan las mujeres para ser reconocidas y promovidas en el campo de la investigación, lo que destaca la necesidad de abordar la inequidad de género en el sistema científico y tecnológico peruano. Millones-Gómez et al. (2021) subrayan la importancia fundamental de fortalecer las políticas de investigación y producción científica en el ámbito educativo para mejorar la calidad de la enseñanza superior y promover la equidad de género en la academia. El estudio enfatiza la necesidad de implementar programas específicos que apoyen la participación y el avance de las mujeres en la investigación científica, así como la importancia de fomentar un entorno inclusivo en las instituciones académicas. Rivera-Lozada et al. (2022) proponen enfoques integrales y medidas políticas para abordar las barreras sistémicas y culturales que limitan el avance profesional de las mujeres en la ciencia y la tecnología en el contexto peruano. Destacan la importancia de adoptar políticas que promuevan la igualdad de género y la creación de un entorno de trabajo inclusivo que fomente la participación plena y equitativa de las mujeres en la investigación y la academia. Por otro lado, Barrutia Barreto et al. (2020) examinan la desigualdad de género en la participación en la investigación científica en el Perú y subrayan la necesidad de abordar los estereotipos de género arraigados y mejorar el acceso de las mujeres a oportunidades de formación y financiamiento en ciencia y tecnología. El estudio destaca la importancia de implementar políticas y programas específicos diseñados para promover la inclusión de las mujeres en la investigación científica y tecnológica.

## 5. CONCLUSIONES

Los resultados del estudio sirven como base para el desarrollo de políticas y programas que fomenten la igualdad de género en la comunidad de profesionales RENACYT. En todos los niveles (desde el nivel I hasta el de investigador distinguido) las mujeres están subrepresentadas en comparación con los hombres. Incluso en el nivel I, en el que se espera encontrar investigadores en las etapas iniciales de sus carreras, la representación de mujeres es notablemente baja.

En el análisis se destaca una marcada disparidad en la distribución de profesionales según su edad y nivel de registro. El grupo de investigadores más numeroso se encuentra en la franja de 40 a 49 años, seguido por el que agrupa a los de 30 a 39 y a los de 50 a 59 años. Esta disparidad se acentúa al comparar los niveles de registro, siendo más notoria en los niveles inferiores (primer y tercer nivel) en contraposición a los niveles superiores (nivel VII y nivel de investigador distinguido).

Por otro lado, se evidencia una disparidad de género significativa en los diferentes niveles del RENACYT. En todos los niveles de registro se observa una notable subrepresentación de mujeres, siendo esta brecha más pronunciada en los estratos más altos. Por ejemplo, en el nivel I, solo el 2,9 % de los registrados son mujeres, mientras que, en el nivel de investigador distinguido, este porcentaje desciende a un meramente simbólico 0,01 %.

Además, se constata una asociación entre la edad de los profesionales y el nivel de registro en el RENACYT. Por ejemplo, el nivel I está predominantemente compuesto por investigadores con edades de entre 40 y 49 años, mientras que la mayoría de los profesionales en el nivel III se encuentra en el rango de edades de 30 a 39 años. Asimismo, en el nivel de investigador distinguido, la mayor parte de los registrados tienen edades comprendidas entre los 40 y 49 años. Esta asociación sugiere la existencia de patrones discernibles en el desarrollo profesional y académico dentro del campo de la investigación y el desarrollo.

Esta brecha de edad plantea la necesidad de explorar las causas subyacentes que podrían estar contribuyendo a estas diferencias. Factores como el acceso desigual a oportunidades de financiamiento y apoyo institucional, las barreras para la participación activa en la investigación en etapas tempranas de la carrera, y los posibles sesgos de género en la evaluación y reconocimiento de logros podrían estar influyendo en la representación desigual de mujeres, especialmente en el nivel I del RENACYT. Comprender estas causas es crucial para el desarrollo de políticas y programas efectivos que promuevan la igualdad de género y fomenten un ambiente inclusivo para todos los profesionales en el ámbito científico y tecnológico en Perú.

## 6. AGRADECIMIENTO

Se agradece el valioso apoyo brindado por el Proyecto ELLAS para la realización del póster presentado en el taller Equality in Leadership for Latin American STEM (2023) y

a la Universidad de Lima. Asimismo, se valoran enormemente las valiosas sugerencias recibidas, las cuales han permitido la extensión del contenido del póster en el presente manuscrito.

## REFERENCIAS

Barrutia Barreto I., Acosta Roa E., Quipuscoa Silvestre M., & Huaranga Rivera H. (2020). La difusión de la investigación científica en Perú: implicaciones en la educación superior. *Biblios Journal of librarianship and information science*, (77), 1-14. https://doi.org/10.5195/biblios.2019.748

Carrasco, E., & Valenzuela, D. (2021). Mujeres que eligen ciencias: autoeficacia, expectativas de resultado, barreras y apoyos percibidos para la elección de carrera universitaria. *Calidad en la Educación*, *54*, 271-302. https://www.mendeley.com/catalogue/6801d45d-d1f8-36e4-b04c-f07264014f3b/

CONCYTEC (2021). *Reglamento de calificación, clasificación y registro de los investigadores del sistema nacional de ciencia, tecnología e innovación tecnológica – Reglamento RENACYT.* https://www.gob.pe/institucion/concytec/informes-publicaciones/2131042-reglamento-de-calificacion-clasificacion-y-registro-de-los-investigadores-del-sistema-nacional-de-ciencia-tecnologia-e-innovacion-tecnologica-reglamento-renacyt

RENACYT (2022). *Registro nacional de investigadores RENACYT. Plataforma Nacional de Datos Abiertos.* https://www.datosabiertos.gob.pe/dataset/registro-nacional-de-investigadores-renacyt/resource/629a1bde-a47f-4d6c-a383-dd70cb6fafc7

De La Cruz-Cerrón, L. A., Ulloa-Ninahuamán, J., Suasnabar-Terrel, J., & Olivera-Meza, J. (2022). La investigación en el Perú: políticas, género y grupo etario. *Revista de Filosofía*, *39*(2), 610-623. http://dx.doi.org/10.5281/zenodo.7316908

López-Bassols, V., Grazzi, M., Guillard, C., & Salazar, M. (2018). *Las brechas de género en ciencia, tecnología e innovación en América Latina y el Caribe: resultados de una recolección piloto y propuesta metodológica para la medición.* Banco Interamericano de Desarrollo. http://dx.doi.org/10.18235/0001082

Millones-Gómez, P., Yangali-Vicente, J., Arispe-Alburqueque, C., Rivera-Lozada, O., Calla-Vásquez, K., Calla-Poma, R., Requena Mendizábal, M. & Minchón-Medina, C. (2021). Políticas de investigación y producción científica: un estudio de 94 universidades peruanas. *PLOS ONE 16*(5), e0252410. https://doi.org/10.1371/journal.pone.0252410

Rivera-Lozada, O., Rivera-Lozada, I., & Bonilla-Asalde, C. (2022). Factors associated with scientific production of professors working at a private university in

Peru: an analytical cross-sectional study. *F1000Research*, *11*, 1219. https://doi.org/10.12688/f1000research.126143.1

Saavedra García, M., & Camarena Adame, M. (2021). Las brechas de género y el empoderamiento femenino en México. *Géneroos, 27*(28), 219-246. https://revistasacademicas.ucol.mx/index.php/generos/article/view/71

Supo-Condori, F., Ríos Burga, J., Sucari León, R., Yabar-Miranda P., & Supo Quispe, L. (2020). Docentes investigadores RENACYT-CONCYTEC en la universidad peruana. *Controversias y concurrencias latinoamericanas*, *12*(21), 407-423. https://ojs.sociologia-alas.org/index.php/CyC/article/view/222

Vargas-Solar, G. (2022). Intersectional study of the gender gap in STEM through the identification of missing datasets about women: a multisided problem. *Applied Sciences*, *12*(12), 5813. https://doi.org/10.3390/app12125813

# Panorama de las mujeres peruanas en carreras STEM

Madeleine Gillian Rabines Floreano
mrabines@unitru.edu.pe
https://orcid.org/0009-0001-0581-7541
Universidad Nacional de Trujillo, Perú

Lourdes Ramírez Cerna
lramirec@ulima.edu.pe
https://orcid.org/0000-0002-7927-7875
Universidad de Lima, Perú

Recibido: 28 de setiembre del 2023 / Aceptado: 1 de junio del 2024
doi: https://doi.org/10.26439/interfases2024.n19.6686

RESUMEN. Este artículo ofrece un panorama de la participación de las mujeres peruanas en las diferentes carreras STEM (Science, Technology, Engineering and Mathematics) de las universidades licenciadas en el Perú. Presenta un análisis de los datos recopilados del Ministerio de Educación (Minedu) y de la Dirección General de Educación Superior Universitaria (Digesu), correspondientes al periodo 2017-2022, sobre la participación de las mujeres y los hombres en las diferentes carreras STEM en las regiones del Perú. La inclusión de las mujeres en estas carreras representa un desafío que comienza en la etapa escolar, en la que se busca promover el aprendizaje digital como herramienta que permita a niñas y adolescentes explorar de manera práctica las disciplinas relacionadas con ciencia, tecnología, ingeniería y matemática, y empoderarse de tal manera que puedan tomar la decisión de estudiar una carrera universitaria relacionada a STEM en el futuro. Según el análisis de los datos de Minedu y la Digesu, existe un aumento progresivo de la participación de las mujeres entre el 2017 y el 2022, lo cual resulta ciertamente alentador.

PALABRAS CLAVE: carreras STEM / mujeres en STEM / base de datos / Minedu y Digesu

## OVERVIEW OF PERUVIAN WOMEN IN STEM CAREERS

ABSTRACT. This article provides an overview of the participation of Peruvian women in various STEM (science, technology, engineering, and mathematics) careers at licensed universities in Peru. It presents an analysis of data collected from the Ministry of Education (Minedu) and the General Directorate of University Higher Education

(Digesu) for the period 2017-2022, focusing on the participation of women and men in different STEM careers across the regions of Peru. Including women in these careers represents a challenge that begins at the school level, where the aim is to promote digital learning as a tool that allows girls and adolescents to explore disciplines related to science, technology, engineering, and mathematics practically, thus empowering them to decide to pursue a university career related to STEM in the future. The analysis shows a progressive increase in the participation of women in STEM careers between 2017 and 2022, which is certainly encouraging.

KEYWORDS: STEM careers / women in STEM / Minedu and Digesu database

## 1. INTRODUCCIÓN

A inicios del siglo xxi se ha notado el gran impacto de las mujeres en carreras en las que tradicionalmente no se consideraba que fueran capaces de destacar (Razo Godínez, 2008). Esta evolución ha permitido que las mujeres desempeñen roles fundamentales y tengan una voz propia en la contribución a proyectos de investigación y desarrollo (I+D) destinados a impulsar avances tecnológicos (Kochhar & Dabla-Norris, 2018; Garduño & Reyes, 2022).

Según Bello (2020), a nivel global, solo el 35 % de los estudiantes en el nivel de educación superior de las carreras STEM está conformado por mujeres. Además, Hernández Herrera (2021) refiere que la educación superior es fundamental para el desarrollo de las diferentes competencias y conocimientos avanzados. En este contexto, el punto de partida es la etapa escolar, durante la cual es fundamental fomentar el acercamiento de las adolescentes a las disciplinas relacionadas con STEM. Por su parte, López Simó et al. (2020) mencionan la importancia de contribuir con la alfabetización digital de las futuras generaciones, a través de la cual pueden aprender a identificar, organizar y analizar información digital, así como crear y comunicar contenidos digitales. El desarrollo del pensamiento computacional en los estudiantes, basado en el pensamiento matemático y lógico, favorece que sean capaces de entender y modificar estas herramientas digitales, lo cual involucra al agente docente.

En el estudio realizado por Valero-Matas y Coca Jiménez (2021) se respalda que las materias de matemáticas, ciencias de la naturaleza y educación artística despiertan más interés cuando se presentan de manera práctica. Su muestra abarca alumnos (colegios privados y públicos) de 3°, 4°, 5° y 6° de educación primaria, de 2° y 4° de educación secundaria, así como estudiantes de bachillerato y formación profesional. En la educación primaria, el 32,7 % prefiere las matemáticas y el 14,4 % opta por ciencias de la naturaleza y educación artística. En educación secundaria, en cambio, los cursos con mayor preferencia son física y química (15,8 %), matemáticas (13,9 %), geografía e historia (12,2 %), y biología y geología (10,1 %). Los cursos con menos interés fueron lengua castellana y literatura, lengua extranjera, cultura clásica, entre otros. Por ello, es importante fomentar un ambiente educativo inclusivo y motivar a todos por igual para aumentar el interés en el aprendizaje de las materias STEM, de tal manera que tenga un impacto significativo en el futuro académico de los estudiantes.

La elección de una carrera profesional determina el camino que seguirá la vida de una persona. Esta decisión se fundamenta, en gran medida, en la educación que recibe en los primeros años de formación académica. Tomar esta decisión no es sencillo, dado que intervienen al menos tres factores: socioeconómico, cultural y el nivel de formación adquirido durante la educación media (Duke Escobar et al., 2021).

En el Perú, solo el 35 % de los profesionales de STEM son mujeres (Vivar, 2023). Ante ello, existen iniciativas para promover la participación de adolescentes mujeres en

carreras de STEM. Por ejemplo, el programa STEM para Todas (2023) reúne a adolescentes de 2° a 5° de secundaria y busca introducirlas al mundo STEM con la finalidad de reducir la brecha de género en este ámbito. Asimismo, Dávila et al. (2022) participan —en conjunto con investigadores de Brasil, Bolivia y Perú— en otras iniciativas, como el proyecto Meninas Digitais y el proyecto ELLAS, que consiste en recolectar datos abiertos por medio de políticas e iniciativas que enfrentan el desafío de la inclusión y liderazgo de las mujeres en STEM.

A partir de lo expuesto anteriormente, se ha propuesto realizar un análisis de los datos recopilados por el Ministerio de Educación (Minedu) y la Dirección General de Educación Superior y Universitaria (Digesu) (2023) correspondientes al periodo de 2017 al 2022. Estos datos indican un aumento progresivo en la participación de mujeres peruanas en carreras STEM en diversas universidades licenciadas en el Perú. Este crecimiento es alentador, ya que refleja un mayor interés y motivación por parte de las mujeres en la elección de estas carreras, desafiando así los estereotipos de la sociedad.

El presente trabajo incluye una metodología detallada (sección segunda), una presentación y análisis de los resultados obtenidos (sección tercera) y una discusión sobre ellos (sección cuarta), que servirán como base para exponer conclusiones relevantes (sección final).

## 2. METODOLOGÍA

Esta investigación adopta un enfoque descriptivo. A través de Minedu-Digesu (2023), se accedió a información sobre la participación de mujeres y hombres en la educación superior, durante el periodo comprendido entre el 2017 y el 2022, en las universidades licenciadas en el Perú. Las principales columnas de la base de datos utilizada comprenden: universidad, región (costa, sierra, selva alta y selva baja), gestión (pública o privada), estado de licenciamiento (licenciada o denegada), género (masculino o femenino), y grupos de carreras STEM (agropecuaria y veterinaria; ciencias de la salud; ciencias naturales, exactas y de la computación; y ingeniería, industria y construcción).

Las columnas mencionadas han sido usadas para filtrar los datos de cada una de las figuras. En todos los casos se consideraron las universidades licenciadas, ambos géneros, todos los grupos de carreras STEM, las regiones del Perú y los dos tipos de gestión (pública y privada), entre el 2017 y el 2022. A partir de esta tarea se realizó un análisis e interpretación de los datos recopilados con el fin de determinar en qué medida las mujeres se han incorporado progresivamente en las distintas carreras STEM en el Perú.

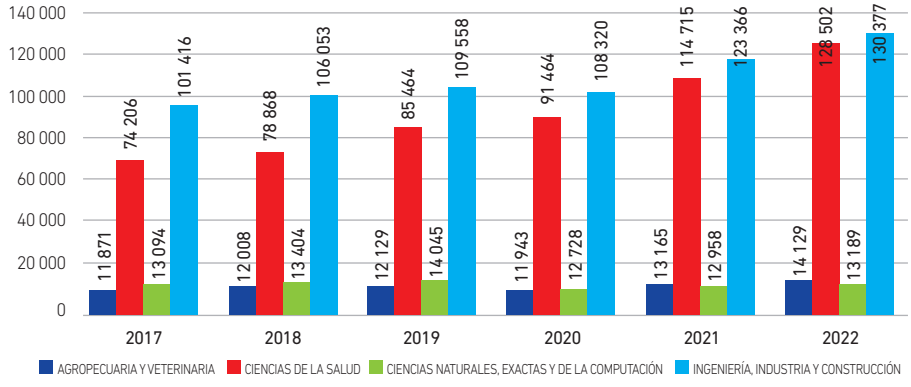
## 3. RESULTADOS

A continuación, se presenta el análisis de los grupos de las carreras STEM en el Perú, extraídos de la base de datos Minedu-Digesu (2023). Los grupos considerados en este estudio son cuatro: agropecuaria y veterinaria (medicina veterinaria, agronomía, agrícola, agroforestal, acuicultura, veterinaria y zootecnia); ciencias de la salud (medicina, odontología, enfermería, obstetricia, nutrición, estomatología, tecnología médica, farmacia y bioquímica); ciencias naturales, exactas y de la computación (biología, zootecnia, física, matemática, estadística, computación científica); y, finalmente, ingeniería, industria y construcción (sistemas, informática, arquitectura, civil, industrial, etcétera).

La Figura 1 muestra el incremento progresivo de la participación de las mujeres en los grupos de las carreras STEM de las universidades licenciadas en el Perú desde el año 2017 al año 2022.

**Figura 1**

*Cantidad de mujeres en los diferentes grupos de las carreras STEM en el Perú*



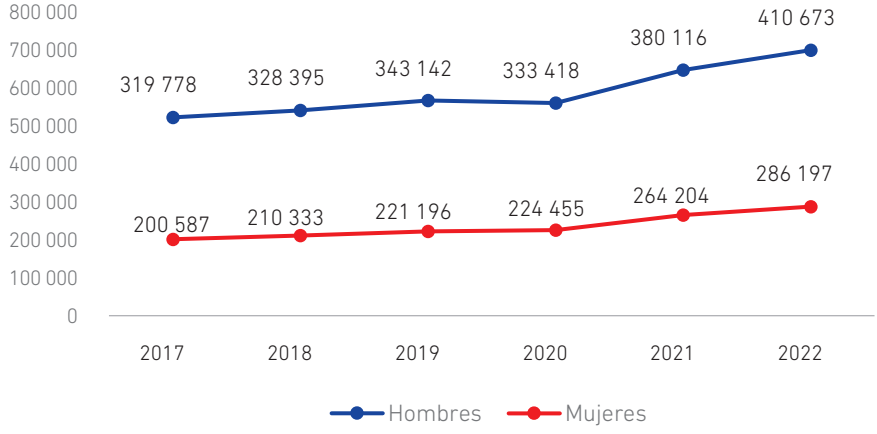

*Nota*. Elaborado con datos del Sistema de Recolección de Información para Educación Superior (SIRIES), por Minedu – Digesu, 2023. Datos al 6 de marzo del 2023.

Según el gráfico de barras que puede apreciarse en la Figura 2, del grupo agropecuaria y veterinaria, en particular en el año 2020, se observa una disminución en la participación de las mujeres: 11 943, frente a 12 129 el año previo. Sin embargo, en el año 2021 esta cifra aumentó a 13 165 mujeres. Algo similar sucede en el grupo de ciencias naturales, exactas y de la computación, en el que la participación de las mujeres disminuyó a 12 728 en el año 2020, en contraste con las 14 045 del año anterior, aunque para el año 2021 esta cifra aumentó a 12 958. Este fenómeno puede atribuirse a la pandemia del COVID-19, que afectó significativamente la situación económica de las personas, se presentaron problemas de salud y hubo dificultades en la adaptación al aprendizaje en línea.

La Figura 2 es una comparativa anual entre la participación de hombres y mujeres en carreras STEM en las universidades licenciadas en el Perú. Este resultado es positivo en el contexto peruano, debido a que muestra un continuo interés de las mujeres en estas carreras a lo largo de los años, al igual que ocurre con los hombres. La pandemia del COVID-19 también afectó ligeramente la participación de los hombres (hubo 333 418 en el año 2020), pero esta aumentó considerablemente el año 2021. Por otro lado, la participación de las mujeres sigue en aumento progresivo a lo largo de los años.

**Figura 2**

*Cantidad anual de personas en carreras STEM en el Perú (2017-2022)*



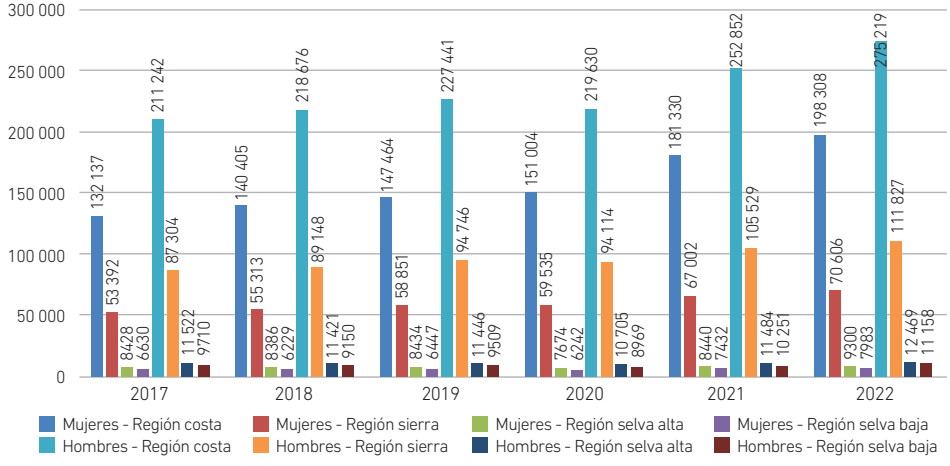

*Nota*. Elaborado con datos del Sistema de Recolección de Información para Educación Superior (SIRIES), por Minedu – Digesu, 2023. Datos al 6 de marzo del 2023.

La Figura 3 muestra la cantidad anual de mujeres y hombres de nivel superior en universidades licenciadas del Perú, agrupadas por regiones geográficas (costa, sierra y selva alta y selva baja). En términos generales, los datos del gráfico de barras revelan un aumento en la participación, tanto de mujeres como de hombres. Sin embargo, se observa que la cantidad de mujeres en la región selva alta disminuyó en el año 2020, al pasar de 8434, el año anterior, a 7674. No obstante, a partir del 2021, esta cifra aumentó a 8440 y continuó en aumento. La cantidad de mujeres en el nivel superior de la región selva baja también experimentó una disminución en el año 2020 (6242 estudiantes), frente a las 6447 del año previo, pero a partir del 2021 se incrementó a 7432 y continuó en aumento.

En cuanto a los hombres de nivel superior, se observa un decremento en el año 2020 en todas las regiones, en comparación con el año 2019. Sin embargo, en los años siguientes, la cantidad de hombres aumenta progresivamente.

**Figura 3**

*Cantidad anual de personas en carreras STEM en el Perú (2017-2022)*

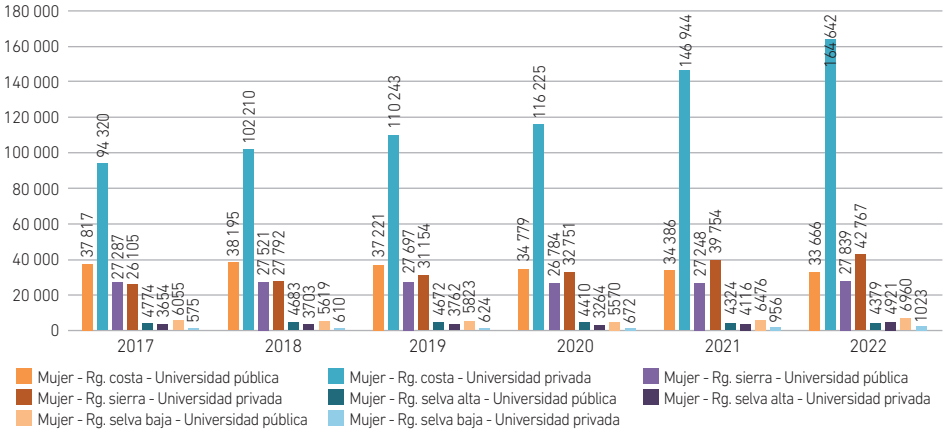



*Nota*. Elaborado con datos del Sistema de Recolección de Información para Educación Superior (SIRIES), por Minedu – Digesu, 2023. Datos al 6 de marzo del 2023.

La Figura 4 presenta la cantidad anual de participación de las mujeres en universidades públicas y privadas, agrupadas por regiones geográficas. En el gráfico se observa que las mujeres de la región costa muestran una preferencia por estudiar en universidades privadas. Por otro lado, en la región sierra se observa un cambio a lo largo de los años: en el 2019 la mayoría estudió en universidades públicas, pero a partir del 2018 en universidades privadas. En cuanto a la región selva alta, desde el 2017 hasta el 2021 las mujeres mostraron una inclinación por estudiar en universidades públicas, aunque en 2022 se observa un aumento en el interés por universidades privadas. Por último, la cantidad de mujeres en la región selva baja estudió en universidades públicas.

**Figura 4**

*Cantidad anual de mujeres en carreras STEM en universidades públicas y privadas, agrupadas por regiones geográficas en el Perú*




*Nota*. Elaborado con datos del Sistema de Recolección de Información para Educación Superior (SIRIES), por Minedu – Digesu, 2023. Datos al 6 de marzo del 2023.

## 4. DISCUSIÓN DE LOS RESULTADOS

De acuerdo con el análisis realizado de los datos extraídos de Minedu-Digesu (2023) durante los años 2017 al 2022, las mujeres y hombres han incrementado su participación en carreras STEM en el contexto peruano.

En el Perú, hay 47 universidades públicas y 46 universidades privadas con licencia. Asimismo, todas las regiones del país cuentan con al menos una universidad pública. Sin embargo, el acceso a la educación universitaria es limitado para los jóvenes de escasos recursos. De cada diez jóvenes que se postulan a universidades públicas, solo dos logran ingresar. En contraste, de cada diez postulantes a universidades privadas, siete consiguen ingresar. Además, estos últimos estudiantes deben contar con recursos económicos para poder hacer frente a la mensualidad universitaria, que depende de las políticas establecidas por cada institución (Ministerio de Educación, 2023). Esta situación se ve reflejada en los resultados cuantitativos de la participación de jóvenes en carreras STEM en las universidades públicas y privadas de la Figura 4

Durante el 2020, la pandemia representó un desafío para los jóvenes estudiantes, en varios aspectos, incluyendo el acceso, la conectividad, la disponibilidad de dispositivos electrónicos y la adaptación al entorno virtual. En la Figura 2 se muestra cómo varía la demanda —de mujeres y hombres— de carreras STEM, durante los años previos y posteriores a la pandemia.

Los estudiantes de las regiones de la sierra y selva presentan problemas en adaptarse a la vida académica universitaria y prefieren buscar oportunidades de trabajo y estudio en la región costa. Como se evidencia en la Figura 4, la costa es la región con la mayor cantidad de mujeres estudiantes, debido a la infraestructura y al nivel educativo de las universidades privadas (Moncada, 2017).

## 5. CONCLUSIONES

Según el análisis de los datos extraídos de Minedu-Digesu (2023), la participación de las mujeres en carreras STEM se ha incrementado en el Perú, a pesar de los desafíos y brechas existentes en la sociedad. Se registró una disminución de su participación en el 2020, probablemente debido a la pandemia del COVID-19. Esta crisis afectó a todos los peruanos y las universidades tuvieron que adaptarse rápidamente a este cambio hacia la educación virtual, modelo que se mantiene con las clases híbridas. De esta manera, a partir del 2021 existe un aumento gradual en la participación de las mujeres y hombres en las carreras STEM, aunque se evidencia una escasa participación en la sierra y la selva.

El incremento de mujeres en carrera STEM, que probablemente se debe a la motivación que reciben en el colegio, representa una evolución en la sociedad. Además, estas mujeres están rompiendo estereotipos sociales y sienten confianza para estudiar estas carreras. Finalmente, la generación actual de mujeres sirve de ejemplo para las generaciones futuras y alienta modelos femeninos a seguir en la innovación de diversas áreas de la ciencia y la tecnología.

## 6. AGRADECIMIENTOS

Se agradece el apoyo recibido del Proyecto ELLAS para la realización del póster presentado en el *workshop* Equality in Leadership for Latin American STEM, realizado el 2023 en la Universidad de Lima, y que constituye el origen del presente manuscrito.

## REFERENCIAS

Bello, A. (2020). *Las mujeres en ciencias, tecnología, ingeniería y matemáticas en América Latina y el Caribe*. ONU Mujeres. https://lac.unwomen.org/es/digiteca/publicaciones/2020/09/mujeres-en-ciencia-tecnologia-ingenieria-y-matematicas-en-america-latina-y-el-caribe

Dávila, G., Guzmán, I., Quintanilla, C., & Maciel, C. (2022). Venciendo los desafíos para la inclusión de mujeres en STEM. *Actas del Congreso Internacional de Ingeniería de Sistemas*, 44-47. https://doi.org/10.26439/ciis2022.6067

Duke Escobar, V. G., Torres Sigüenza, J. O., García Perdido, M. U., & Toledo Martínez, C. S. (2021). Factores que inciden en la elección de carreras STEM en la educación

universitaria de El Salvador. *Anuario de Investigación: Universidad Católica de El Salvador*, *10*, 23-38. https://doi.org/10.5377/aiunicaes.v10i1.12487

Garduño, E., & Reyes, A. (2022). *Mujeres y educación en STEM: una mirada con perspectiva de género. Apuntes para México*. Mujeres Unidas por la Educación - Movimiento STEM. https://www.movimientostem.org/wp-content/uploads/2022/02/Mujeres-y-educacion-en-STEM-una-mirada-con-perspectiva-de-genero.pdf

Kochhar, K., & Dabla-Norris, E. (2018, 20 de noviembre). Las mujeres, la tecnología y el futuro del trabajo. *IMF Blog*. https://www.imf.org/es/Blogs/Articles/2018/11/16/blog-Women-Technology-the-Future-of-Work

Hernández Herrera, C. A. (2021). Las mujeres STEM y sus apreciaciones sobre su transitar por la carrera universitaria. *Nova Scientia*, *13*(27), 26. https://doi.org/10.21640/ns.v13i27.2753

López Simó, V., Couso Lagarón, D., & Simarro Rodríguez, C. (2020). Educación STEM en y para el mundo digital: el papel de las herramientas digitales en el desempeño de prácticas científicas, ingenieriles y matemáticas. *Revista de Educación a Distancia (RED)*, *20*(62). https://doi.org/10.6018/red.410011

Ministerio de Educación - Dirección General de Educación Superior Universitaria. (2023). Sistema de Recolección de Información para Educación Superior (SIRIES).

Ministerio de Educación. (2023). *La universidad en cifras*. https://repositorio.minedu.gob.pe/bitstream/handle/20.500.12799/9077/La%20Universidad%20en%20Cifras.pdf?sequence=1&isAllowed=y

Moncada, F. C. (2017). Perspectivas de jóvenes universitarios sobre la educación superior: una visión intercultural. *Revista Psicológica Herediana*, *9*(1-2), 34. https://doi.org/10.20453/rph.v9i1-2.3004

Razo Godínez, M. L. (2008). La inserción de las mujeres en las carreras de ingeniería y tecnología. *Perfiles Educativos*, *30*(121), 63-96. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-26982008000300004&lng=es&tlng=es

STEM para todas. (2023). https://www.stemparatodas2023.org

Valero-Matas, J. A., & Coca Jiménez, P. (2021). La percepción de las materias STEM en estudiantes de primaria y secundaria. *Sociología y Tecnociencia, 11*(Extra 1), 116-138. https://revistas.uva.es/index.php/sociotecno/article/view/5144/3799

Vivar, F. (2023). La ciencia no tiene género. En Dirección de Comunicaciones e Imagen Institucional de la Universidad del Pacífico (Ed.), *Desde la academia. Retos para la mujer peruana* (p. 6). https://www.up.edu.pe/prensa/noticias/up-publica-desde-la-academia-8m

# Aplicación *cloud native* en el contexto de una ingeniería de *software* continua

Zoraida Mamani Rodriguez
zmamanir@unmsm.edu.pe
https://orcid.org/0000-0002-2590-8387
Universidad Nacional Mayor de San Marcos, Perú

Recibido: 28 de marzo del 2024 / Aceptado: 23 de mayo del 2024
doi: https://doi.org/10.26439/interfases2024.n19.7038

RESUMEN. Una aplicación *cloud native* es un tipo de *software* que ha sido diseñado específicamente para ejecutarse en la nube, con enfoque distribuido, elástico, escalado horizontal y compuesto de microservicios con implementación autónoma. Asimismo, se diseñan con arquitecturas web *cloud native*, operan en una plataforma elástica de autoservicio y se caracterizan por su resiliencia y elasticidad. La ingeniería de *software* continua es un proceso que busca articular la ingeniería de requisitos, el desarrollo y las operaciones en un bucle continuo, con una retroalimentación recíproca, con la finalidad de producir un *software* de calidad. En ese contexto, el presente trabajo propone el diseño e implementación de una aplicación *cloud native* en una perspectiva de ingeniería de *software* continua, aplicada al caso de estudio SIGCON. Usa el modelo de servicio *cloud* CaaS, aplica el patrón BFF en la construcción del *software*, realiza contenedorización del *frontend*, *backend* y almacenamiento, y expone los resultados.

PALABRAS CLAVE: *cloud native application* / modelos de servicios *cloud* / patrón BFF / ingeniería de *software* continua

## CONTINUOUS SOFTWARE ENGINEERING OF A CLOUD-NATIVE APPLICATION

ABSTRACT. A cloud-native application is a software specifically designed to run in the cloud, focusing on distributed, elastic, horizontally scaled, and microservice-based architecture with autonomous deployment. These applications are designed with cloud-native web architectures, operate on an elastic self-service platform, and stand out because of their resilience and elasticity. Continuous software engineering integrates requirements engineering, development, and operations in a continuous loop with reciprocal feedback to produce quality software. The present work proposes to design and implement a cloud-native application applied to the SIGCON case study from a continuous software engineering perspective. It uses the CaaS cloud service

model, applies the BFF pattern in software construction, containerizes the frontend, backend, and storage, and presents the results.

KEYWORDS: Cloud Native Application / Cloud Service Models / BFF Pattern / Continuous Software Engineering

## 1. INTRODUCCIÓN

La industria del *software* evoluciona aceleradamente. Nuevos enfoques se orientan a la construcción de *software* nativo de la nube, ejecutándose bajo modelos de computación "sin servidor", con microservicios autónomos desplegados en contenedores y diseñados con arquitecturas web componibles. Estas tendencias tecnológicas resilientes se exponen en la literatura. Por otro lado, existe una alta demanda laboral a nivel global de profesionales idóneos en la construcción de aplicaciones *cloud native*. En consecuencia, corresponde a la academia la formación de recursos humanos con competencias y capacidades en la construcción de *software* bajo estos nuevos enfoques, que les permita asumir los desafíos de la industria. El principal aporte de la investigación se centra en diseñar e implementar una aplicación *cloud native* en el contexto de una ingeniería de *software* continua. Utiliza arquitecturas *cloud* componibles, patrón *cloud* BFF, metodología ágil Scrum, cultura DevOps en el desarrollo colaborativo de aplicaciones escalables, con recursos humanos en proceso formativo, ágiles, rotativos, altamente motivados y enfocados en la automatización moderna de procesos de negocios. Este ecosistema de componentes tecnológicos, humanos, arquitecturas, data, servidores virtualizados y modelos de servicios *cloud,* lo exhibe como innovador en la industria peruana de *software*. Considerando lo expuesto, la presente investigación establece como objetivos: (1) diseñar una arquitectura CNA con una perspectiva de ingeniería de *software* continua; (2) implementar la propuesta en el caso de estudio SIGCON; (3) realizar el despliegue del *software* bajo el modelo de servicio *cloud* CaaS; y (4) evaluar los resultados.

## 2. MARCO TEÓRICO

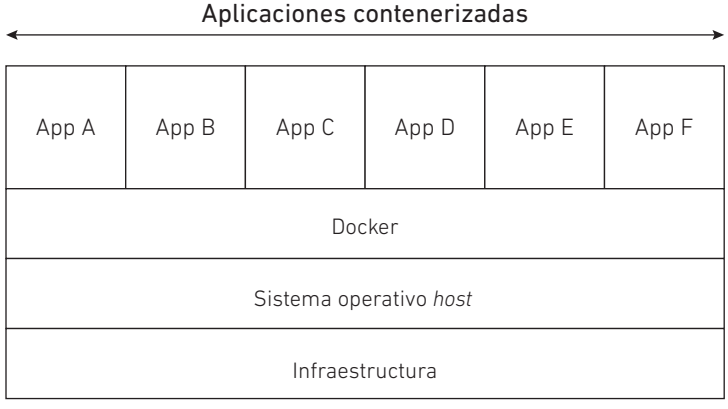
### Modelos de servicio *cloud*

El término *cloud* se populariza en el año 2006 con el lanzamiento de un servicio de nube pública de propósito general de parte de Amazon Web Services (AWS). Este tipo de modelo de servicio corresponde a infraestructura como servicio (IaaS), que consiste en proporcionar al consumidor procesamiento, almacenamiento, redes y otros recursos informáticos fundamentales, de manera que dicho consumidor pueda implementar y ejecutar *software* arbitrario, sistemas operativos y aplicaciones. El consumidor no gestiona ni controla la IaaS, pero tiene control sobre el sistema operativo (SO), tiempos de ejecución, escalamiento, código fuente de la aplicación, así como los aspectos relativos a los datos y configuración que residen en ella y al control limitado de componentes de red, como *firewalls*. Posteriormente, en el año 2009 destacan —en el ámbito tecnológico— las plataformas como servicios (PaaS), que ofrecen al consumidor el uso de la IaaS para desplegar sus aplicaciones. Aquí el consumidor no administra ni controla la IaaS, solo tiene control sobre las aplicaciones implementadas y los ajustes de configuración para su funcionamiento (Mell & Grance, 2012). En estos años, en el ámbito de la

comunidad de desarrolladores de *software* se popularizó la plataforma Heroku, debido a que simplificó el proceso de desarrollo y despliegue de las aplicaciones, haciendo más simple, eficiente y competitivo el proceso de despliegue. Como consecuencia de ello, se incrementó la productividad del desarrollo de *software* y se redujeron costos, pues el consumo por el uso de una PaaS es bajo demanda y se disponen de herramientas para monitorear y escalar las aplicaciones. El modelo de *software* como servicio ofrece al consumidor el uso de las aplicaciones del proveedor que se ejecutan en una IaaS. Este modelo es utilizado para distribuir aplicaciones en la nube a los usuarios a través de Internet y se pone a disposición bajo un modelo de pago que puede ser una suscripción o una compra. Hussein et al. (2019) describen un contenedor como una tecnología de virtualización liviana emergente que opera a nivel del SO para encapsular una tarea y sus dependencias de biblioteca para su ejecución. Es posible que diferentes contenedores se ejecuten en un SO, como se aprecia en la Figura 1. El modelo contenedor como servicio (CaaS) surge con la finalidad de resolver problemas de las aplicaciones desarrolladas en un determinado entorno PaaS, limitadas por las especificaciones de ese entorno. CaaS permite el despliegue de la aplicación independiente del entorno PaaS, eliminando los posibles conflictos, barreras o limitaciones que podrían originarse por la convivencia de varios servicios como base de datos, lenguajes de programación, servidores de aplicaciones, entre otros, en un mismo entorno PaaS. Asimismo, este tipo de aplicaciones que se ejecutan en un CaaS son portátiles, toda vez que pueden trasladarse a cualquier otro entorno de ejecución. Una plataforma abierta para ejecutar contenedores de aplicaciones es Docker (Docker, 2024), cuya arquitectura se muestra en la Figura 1.

**Figura 1**

*Modelo de servicio CaaS*

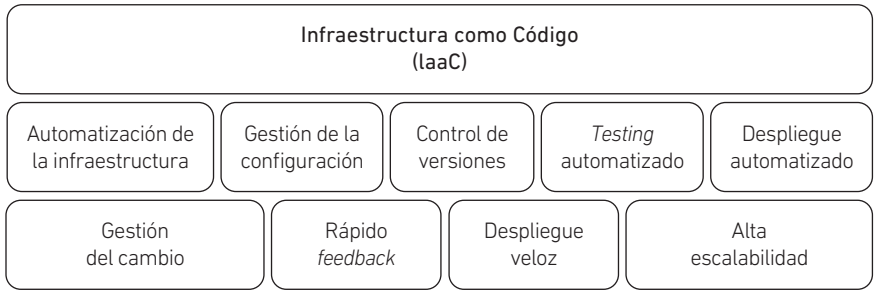



***Nota.*** Tomado de Docker, 2024, *Use containers to build, share and run your applications. https://www.docker.com/resources/what-container/*

El modelo función como servicio (FaaS) utiliza la computación *serverless* o sin servidor. Este tipo de modelo permite que las arquitecturas de microservicios funcionen considerando cada microservicio como una función, la cual es tomada por el proveedor FaaS, que ejecuta y gestiona la función desplegada. Este tipo de modelo es muy utilizado por los desarrolladores de *software*, dado que no tienen que preocuparse por la infraestructura ni por el despliegue, lo que permite la reducción de tiempo y de costos en la construcción de *software*. Entre las ventajas del modelo FaaS está que las tareas de escalamiento, mantenimiento, recuperación ante desastres, así como aspectos de seguridad, son realizadas por el proveedor del servicio FaaS. Como desventajas tenemos la pérdida de control del sistema, el proceso de pruebas (que es más complejo), una dependencia a largo plazo con el proveedor FaaS, además de tener que ceñirse estrictamente a sus requisitos para que la solución *software* funcione correctamente. AWS Lambda es una plataforma FaaS muy popular, seguida por Azure Functions y Google Cloud Functions (Habala et al., 2023). Infraestructura como código (IaaC) es un proceso de gestión de infraestructura de TI que aplica las mejores prácticas: desde el desarrollo de *software* DevOps hasta la gestión de recursos de infraestructura en la nube mediante código o *scripts* como máquinas virtuales (MV), redes, balanceadores de carga, bases de datos y otras aplicaciones en red. A estos *scripts* utilizados en la IaaC se les conoce como *scripts* de configuración como código, los cuales reducen el aprovisionamiento de recursos en la nube (Almuairfi & Alenezi, 2020; Buchanan, 2024). En la Figura 2 se aprecian las funcionalidades que ofrece este proceso de automatización de la infraestructura, gestión de la configuración, control de versiones, automatización de las pruebas, automatización del despliegue, gestión del cambio, rápido *feedback*, despliegue veloz, así como alta escalabilidad.

**Figura 2**

*Funcionalidades de la IaaC*




*Nota.* Tomado de I. Buchanan, 2024, *Infrastructure as code. How infrastructure as code (IaC) manages complex infrastructures*. Atlassian. https://www.atlassian.com/microservices/cloud-computing/infrastructure-as-code

En el trabajo de Duan et al. (2015) se ha realizado una amplia revisión de la literatura sobre el término XaaS y los comprendidos como *as a service*. *Anything* como servicio o cualquier cosa como servicio (XaaS) comprende cualquier herramienta, aplicación o recurso que se suministre a través de la nube mediante suscripción. Este tipo de servicio generalmente se rige por acuerdos del nivel de servicio (SLA) que proveedores y clientes especifican en el contrato. Este mecanismo permite ahorrar costos, ofrece una rápida comercialización, flexibilidad para centrarse en el negocio principal, escalabilidad y confiabilidad. Los modelos de servicios en la nube continúan evolucionando, como se puede apreciar en la Figura 3. Hemos identificado a la fecha hasta veinticuatro categorías.

**Figura 3**

*Modelos de servicio* cloud


| Acrónimo | Descripción |
|---|---|
| IaaS | Infraestructura como servicio |
| PaaS | Plataforma como servicio |
| SaaS | *Software* como servicio |
| FaaS | Función como servicio |
| CaaS | Contenedores como servicio |
| IaaC | Infraestructura como código |
| CaC | Configuración como código |
| XaaS | Cualquier cosa como servicio |
| AaaS | *Analytics* como servicio |
| BPaaS | Plataforma *backend* como servicio |
| MBaaS | *Backend* móvil como servicio |
| BaaS | *Backup* como servicio |
| RaaS | *Recovery* como servicio |
| DRaaS | *Disaster recovery* como servicio |
| HaaS | *Hardware* como servicio |
| MaaS | *Monitoring* como servicio |
| SECaaS | Seguridad como servicio |
| FWaaS | *Firewall* como servicio |
| DaaS | *Database* como servicio |
| NaaS | *Network* como servicio |
| STaaS | *Storage* como servicio |
| AIaaS | Inteligencia artificial como servicio |
| EssS | *Encryption* como servicio |
| RAaaS | Asignación de recursos como servicio en múltiples proveedores de nube |

## Aplicación *cloud native*

Una *cloud native application* o aplicación nativa de la nube (CNA) es un sistema distribuido, elástico, con escalado horizontal, compuesto de microservicios que aísla el estado en un mínimo de componentes con estado. La aplicación y cada una de sus unidades de implementación autónoma se diseñan de acuerdo con patrones de diseño centrados en la nube y operan en una plataforma elástica de autoservicio. Así lo definen Kratzke y Quint (2017). Otra definición similar describe a una CNA como un tipo de *software* que ha sido diseñado específicamente para ejecutarse en un entorno de nube. Para ser considerada una CNA, una aplicación *software* debe cumplir con ciertas características, como resiliencia y elasticidad. La resiliencia implica que una CNA debe anticiparse a los fallos y fluctuaciones en el contexto de los recursos de la nube, así como de los servicios de terceros necesarios

para su operatividad. Por su parte, la elasticidad comprende la capacidad de las CNA para ampliar el uso de recursos requerido, evitando el aprovisionamiento excesivo o insuficiente y considerando que la nube es un servicio medido que ofrece autoservicio bajo demanda y que, por tanto, requiere rápida elasticidad (Toffetti et al., 2017). La naturaleza de las CNA es que frecuentemente dependen de servicios de terceros, por lo que se incrementa el riesgo de que estos servicios puedan fallar o presentar insuficiencia en la calidad de su servicio. Las CNA utilizan una pila (*stack* en inglés) de *software* de código abierto para segmentar aplicaciones en microservicios, empaquetar cada microservicio en su propio contenedor y orquestar dinámicamente estos contenedores para optimizar el uso de recursos de la nube.

### Ingeniería de *software* continua

La ingeniería de *software* continua (CSE) es un proceso en auge, que busca articular la ingeniería de requisitos, el desarrollo y las operaciones en un bucle continuo, con retroalimentación recíproca, para producir *software* de calidad. La CSE es uno de los principios DevOps que enfatiza que la integración entre el desarrollo de *software* y su distribución operativa debe ser continua. DevOps mejora la colaboración entre las partes interesadas, los equipos de desarrollo y las operaciones. La práctica de la CSE es la entrega rápida, la minimización del tiempo de lanzamiento de nuevas funcionalidades, la mitigación de riesgos, y el impulso de mejoras o refactorizaciones bajo un proceso de entrega continua (Eramo et al., 2024).

## 3. METODOLOGÍA

En la sección anterior se han revisado los fundamentos de los modelos de servicios de la nube, las CNA y la CSE. En esta sección desarrollaremos la metodología que se aplicará en un caso de estudio. Esta se organiza de la siguiente manera: (A) diseño de la arquitectura CNA en el contexto de una CSE, y (B) implementación de una CNA mediante un caso de estudio.

El caso de estudio de la investigación está orientado a la administración de edificaciones modernas, como son las viviendas multifamiliares. Es una tendencia mundial el crecimiento vertical de las ciudades y la construcción de grandes edificios multifamiliares (Espinoza, 2020; Vega, 2021). Este tipo de predios mantienen áreas comunes, compartidas entre todos los propietarios de las unidades inmobiliarias o departamentos; estos pueden ser parques internos, gimnasios, piscinas, ascensores, salas de niños, salas de cine, zonas de parrillas, entre otros. Dado que estos espacios requieren mantenimiento permanente, se hace necesario contar con un área administrativa que se encargue de todo el proceso de gestión de áreas comunes, emisión y cobranza de recibos de mantenimiento, balance de ingresos y egresos, gestión de visitas, gestión de compras, gestión de recursos humanos, entre otras tareas que implique mantener el predio operativo, de

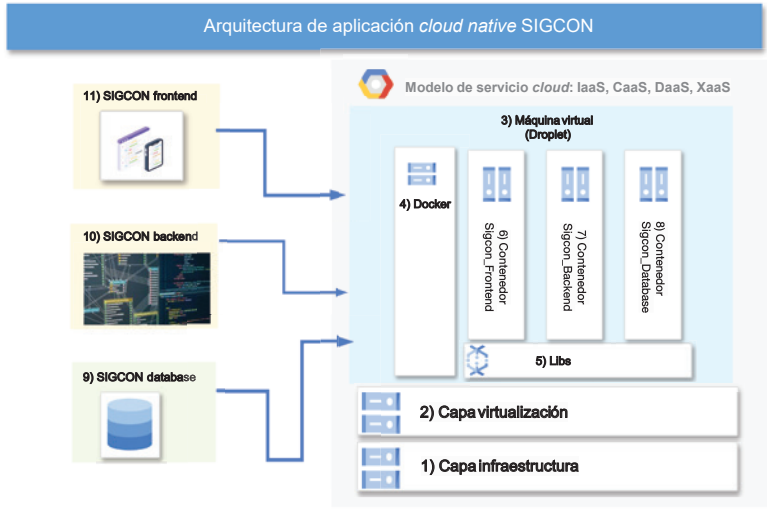
manera que sus propietarios puedan usarlo sin mayores preocupaciones (Casa Grande, 2023). Según lo descrito, existe una alta demanda por contar con *software* que automatice los procesos de gestión de mantenimiento del predio o condominio, y es en ese orden de ideas que proponemos el sistema de gestión de condominios (SIGCON) desde una perspectiva de arquitectura de aplicación *cloud native* en el contexto de una CSE.

### A. Diseño de la arquitectura CNA en el contexto de una CSE

En la Figura 4 se presenta la arquitectura del sistema SIGCON. La primera capa (1) corresponde a una IaaS con el proveedor *cloud* Digital Ocean. Se adquirió una MV con 1 GB memoria / 25 GB disco duro con SO Linux Ubuntu 22.10 x64. Esta MV se constituye en la IaaS del proyecto. Sobre esta capa se aplicó el modelo de servicio CaaS, específicamente desplegando tres contenedores (6, 7 y 8 en la Figura 4) gestionados por la herramienta Docker (4) y un conjunto de librerías comunes (5). Se utilizó CaaS debido al uso de *frameworks*; Flask por el lado del *backend*, uso de lenguaje python, angular por el lado del *frontend*, uso del lenguaje typescript, html, css, arquitecturas e interfaces en el desarrollo del sistema SIGCON. Con la finalidad de facilitar su despliegue y evitar posibles conflictos en el uso de puertos y librerías de terceros (también llamadas dependencias), se hizo necesario el aislamiento de la aplicación de su entorno. Por otro lado, la naturaleza de una CNA es su construcción a razón de microservicios y su despliegue en contenedores con fines de resiliencia y elasticidad. En ese orden de ideas, SIGCON fue diseñado bajo un contexto de servicios desacoplados: la clara separación del *backend* de la aplicación (10) con respecto al *frontend* (11), así como el despliegue de la base de datos mediante un modelo DaaS (9), así lo demuestra.

**Figura 4**

*Arquitectura física de la aplicación* cloud native *SIGCON*



### B. Implementación de una CNA mediante un caso de estudio

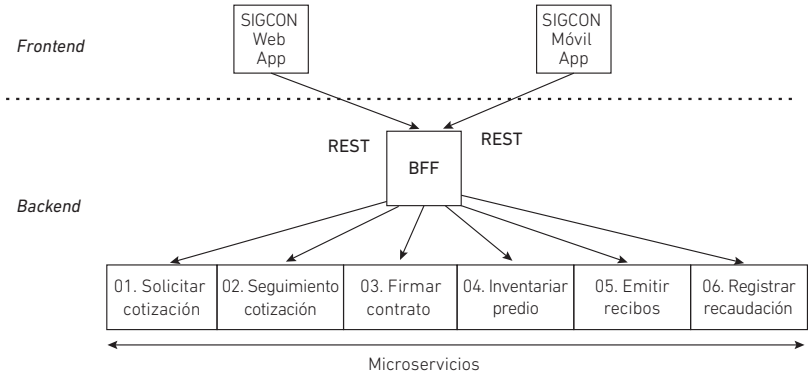
La implementación de la CNA enfocada al caso de estudio SIGCON inicia con el modelado del negocio, identificando cinco procesos clave como casos de uso de negocio (CUN): gestión de servicios (CUN 1.0), gestión de predios (CUN 2.0), gestión del mantenimiento (CUN 3.0), gestión de cobranza (CUN 4.0) y gestión de adquisiciones (CUN 5.0). La fase de requisitos permitió identificar los casos de uso del sistema, a partir de los cuales, aplicando la metodología ágil Scrum, se procedió a formular el *product backlog*, con seis requisitos funcionales fundamentales: 01. solicitar cotización, 02. seguimiento cotización, 03. firmar contrato, 04. inventariar predio, 05. emitir recibos y 06. registrar recaudación.

#### *Arquitectura del* software

La arquitectura del *software* del proyecto adopta inicialmente una arquitectura basada en tres capas (capa de presentación, capa de la lógica de aplicaciones y almacenamiento). Luego de aplicar los patrones generales de *software* para la asignación de responsabilidades, se descompone la capa de la lógica de aplicaciones, constituyéndose en una arquitectura multicapas compuesta por estratos verticales o subcapas (capa del dominio, capa de servicios y particiones horizontales), una por cada requisito identificado en la etapa de análisis del proyecto, de acuerdo con Larman (1999). El uso de una arquitectura web moderna conlleva a implementar arquitecturas multicapas utilizando el patrón *backend for frontend* (BFF), proyectándose a la resiliencia, escalamiento y elasticidad que caracteriza a las CNA. Este desacoplamiento corresponde a una división entre la V que reside en el lado del cliente y la M en el lado del servidor, considerando el estilo arquitectónico modelo-vista-controlador. BFF permite utilizar arquitecturas de *software* componibles de manera que el *frontend* y el *backend* evolucionen a diferentes ritmos y escalas. Esto hace posible adaptarse entre las necesidades de aplicaciones nativas, móviles, SPA (*single-page application*) y microservicios (Brown & Woolf, 2016). En la Figura 5 se expone la arquitectura lógica componible del proyecto SIGCON.

**Figura 5**

*Arquitectura lógica del proyecto SIGCON*



*Desarrollo del* frontend

El desarrollo del *frontend* ha sido realizado con el *framework* Angular 16.2.9, el cual permite crear aplicaciones de una sola página (SPA). Las SPA se implementan de forma nativa en los navegadores modernos, utilizan HTML5, CSS3 y JavaScript; almacenan el estado de la página en el cliente y se conectan al *backend* a través de servicios REST. Se conoce así a este enfoque porque todo el código (HTML, CSS y JavaScript) necesario para un conjunto de funcionalidades que pueden corresponder a múltiples pantallas o páginas lógicas, se recuperan como una solicitud de una sola página. JavaScript se encarga de toda la manipulación del modelo de objetos de documento HTML, la navegación de la página y el acceso a los datos del *backend*. Las SPA son utilizadas para aprovechar los principios del diseño responsivo, permitiendo optimizar la experiencia del usuario en cuanto al diseño y tamaño de la pantalla. Los scripts Media CSS se utilizan a menudo para incluir bloques específicos que solo se aplican a determinados tipos de pantalla. Esta técnica permite especificar un conjunto diferente de reglas CSS para tabletas, teléfonos móviles o portátiles, lo que da como resultado pantallas diseñadas y configuradas específicamente para esos dispositivos (Brown & Woolf, 2016). En ese orden de ideas, Angular es un *framework* basado en componentes para crear aplicaciones web escalables; está escrito en TypeScript. Implementa la funcionalidad principal y opcional como un conjunto de bibliotecas de TypeScript que importa a sus aplicaciones. En la Tabla 1 se presentan los principales *frameworks* utilizados en el proyecto.

**Tabla 1**

Frameworks *del proyecto SIGCON*

| Servicios | *Frameworks* |
|---|---|
| SIGCON_frontend | Angular 16.2.9, bootstrap, nginx, node,js |
| SIGCON_backend | Python, Flask, marshmallow, psycopg2, SQLAlchemy |
| SIGCON_database | Postgresql 15 |

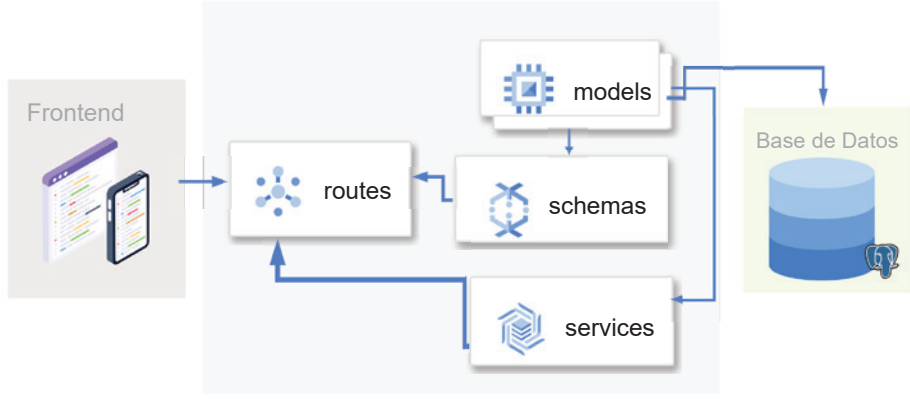
*Desarrollo del* backend

El *backend* del proyecto, como se aprecia en la Figura 6, se desarrolló utilizando el micro *framework* Flask debido a que permite un desarrollo fácil y rápido, pero con capacidad de escalar a aplicaciones más complejas. Este *framework* utiliza lenguaje de programación Python, de código abierto, no depende de ninguna plataforma, es un lenguaje dinámico, admite programación orientada a objetos y es funcional. En ese contexto, Flask proporciona el código necesario para poner en marcha de forma rápida y segura los servicios web (Flask, s. f.). Asimismo, se utilizaron librerías de terceros para agilizar el desarrollo, como por ejemplo el Flask-Marshmallow, una biblioteca de serialización y

deserialización de objetos, que se integra con Flask-SQLAlchemy —una extensión para Flask que agrega soporte para el mapeador relacional de objetos (ORM) SQLAlchemy—, y permite configurar objetos y patrones comunes para usar esos objetos, como una sesión vinculada a cada solicitud web, modelos y motores (FlaskSQLalchemy, s. f.). Por su parte, el ORM referido brinda el poder y flexibilidad de SQL y proporciona un conjunto completo de patrones de persistencia conocidos a nivel empresarial, diseñados para un acceso eficiente y de alto rendimiento a bases de datos, adaptados a un lenguaje de dominio simple (SQLAlchemy, s. f.). Otra librería de terceros utilizada es psycopg2, que es el adaptador para bases de datos PostgreSQL.

**Figura 6**

*Arquitectura del* backend *SIGCON*




*Despliegue del proyecto*

El proceso de despliegue siguió el diseño de la arquitectura CNA en el contexto de una CSE indicado en la Figura 4. Considerando el enfoque CaaS, se procedió a seleccionar el Docker como herramienta de contenedorización en este proyecto, debido a su simplicidad de uso, además de la documentación y el soporte de una amplia comunidad de desarrolladores que la respaldan. Un contenedor Docker es un paquete de *software* liviano, independiente y ejecutable que encapsula una aplicación y sus dependencias, bibliotecas y archivos de configuración (Merelli et al., 2019). Un contenedor dentro de un entorno virtualizado proporciona un entorno de ejecución consistente y aislado para que las aplicaciones se ejecuten sin problemas en diferentes entornos informáticos. El archivo de configuración Dockerfile para el despliegue del *backend* se indica a continuación: como se aprecia en la sección del código, línea 1, se indica la imagen Docker a utilizar. Una imagen de Docker es un repositorio alojado en Docker Hub, sirve como punto de partida para la mayoría de los desarrolladores de *software*, incluye sistemas operativos

y lenguajes de programación y es confiable dado que tiene pocas o ninguna vulnerabilidad (Docker, 2024). En el presente proyecto se utilizó la imagen Docker Python:3.8, que tiene instalado el SO Linux ubuntu, python:3.8. En la línea 2 se copia el código fuente del *backend* en el directorio predeterminado en ubuntu, carpeta app, la cual se establece como directorio de trabajo en la línea 3. En la línea 4 se instalan las dependencias requeridas para el funcionamiento del *backend*, las cuales se encuentran consignadas en el archivo requirements.txt del proyecto. En la línea 5 se establece el puerto a considerar en el despliegue y, finalmente, en la línea 6, se precisan los comandos para lanzar la ejecución del *backend*.

1. FROM python:3.8
2. COPY ./ /app
3. WORKDIR /app
4. RUN pip install -r requirements.txt
5. EXPOSE 5000
6. CMD [ “python3”, “-m” , “flask”, “run”, “--host=0.0.0.0”]

Luego de definir el archivo de configuración Dockerfile, ya se puede crear el contenedor. Para ello se ejecuta el comando: sudo docker build -t sigcon-backend-1.0.0. Esta instrucción dará lectura al Dockerfile y procesará las líneas contenidas allí. Finalmente, para ejecutar el contenedor se lanza la instrucción: sudo docker run -p 5000:5000 sigcon-backend-1.0.0.

Para la contenedorización del *frontend* igualmente se configuró el Dockerfile. Aquí se utilizó la distribución Linux Alpine, como se muestra en la línea 1, se estableció el directorio de trabajo, la carpeta app y en ella se realizó una copia de la carpeta dist que contiene el compilado del *frontend*. En la línea 4 se instala nginx, que es un es un servidor web ligero de alto rendimiento; luego, se requiere copiar el archivo de configuración nginx.conf que contiene los parámetros de configuración. Finalmente, en la línea 6 se copia la carpeta sigcon_frontend que se encuentra en la carpeta dist a la carpeta html.

1. FROM node:16-alpine3.11 AS build
2. WORKDIR /usr/src/app
3. COPY . .
4. FROM nginx:1.17.1-alpine
5. COPY nginx.conf /etc/nginx/nginx.conf
6. COPY --from=build /usr/src/app/dist/sigcon_frontend /usr/share/nginx/html

El procedimiento seguido para crear el contenedor del *backend* es el mismo para el *frontend*: sudo docker build -t sigcon-frontend-1.0.0. Y para ejecutar el contenedor: sudo docker run -p 4200:80 sigcon-frontend-1.0.0..

## 4. RESULTADOS

El paradigma CNA continúa evolucionando. Hoy ya se habla de ingeniería CNA, cuya implementación aún es compleja, puesto que implica contar con recursos humanos altamente especializados en *cloud native*, modelos de servicios *cloud*, diseño y construcción de arquitecturas *cloud native*, DevOps, integración y entrega continua de *software*, patrones de diseño para enfoques *cloud native*, todo ello complementado al proceso de CSE que deben utilizar los equipos de desarrollo.

El despliegue de los servicios de acuerdo con la arquitectura propuesta se puede apreciar en la Tabla 2. Se trata de contenedores desacoplados, con ejecución independiente. Los *scripts* definidos en la sección Despliegue del proyecto permiten realizar el proceso de integración y entrega continua de *software*.

**Tabla 2**

*Contenedores del proyecto SIGCON*

| Imagen | Comando | Puertos | Nombre Contenedor | URL |
|---|---|---|---|---|
| sigcon-back-end-1.0.0 | "python3 -m flask ru..." | 0.0.0.0:5000->5000/tcp | unruffled_ poitras | http://137.184.120.127:4200/principal |
| sigcon-fron-tend-1.0.0 | "nginx -g 'dae-mon of..." | 0.0.0.0:4200->80/tcp | upbeat_tesla | |
| sigcon-data-base-1.0.0 | "postgres 'dae-mon of..." | 0.0.0.0:5432->5432/tcp | epic_ brahmagupta | |

El utilizar el modelo de servicios CaaS en el proyecto, nos ha permitido disponer del prototipo de la aplicación en la nube a muy bajo costo. Este modelo adopta una infraestructura inmutable, pues aquí los contenedores no se reparan ni modifican. Si uno falla o requiere una actualización, se destruye y se aprovisiona uno nuevo, todo a través de los *scripts* consignados en el archivo Dockerfile.

El prototipo del producto *software* resultante de la investigación se encuentra publicado en la URL http://137.184.120.127:4200/principal.

## 5. DISCUSIÓN DE LOS RESULTADOS

Concordamos con lo señalado por Stine (2015) y Balalaie et al. (2016) sobre las arquitecturas CNA, las cuales buscan entregar *software* más rápido, con mayor aislamiento, con

tolerancia a fallos y recuperación automática, de manera que permiten un escalamiento horizontal de aplicaciones, e incorporar a la CNA otros entornos (móvil, sistemas heredados), todo ello bajo el enfoque de computación sin servidor.

La parte experimental de nuestra investigación nos permitió coincidir con Kratzke y Quint (2017) en el sentido de que los microservicios no son otra cosa que la descomposición de las funcionalidades de los sistemas tradicionales, conocidos como monolitos. La interacción entre estos es a través de APIs, las cuales siguen el estilo REST con serialización JSON u otro formato. Su despliegue se realiza en plataformas de infraestructuras ágiles de autoservicio, operando de manera autónoma, con escalamiento automatizado, bajo demanda de aplicaciones, gestión del estado, enrutamiento dinámico, equilibrio de carga y monitoreo de métricas.

## 6. CONCLUSIONES

- La presente investigación propuso el diseño e implementación de una aplicación *cloud native* en el contexto de una CSE. La propuesta usa el modelo de servicio cloud CaaS y aplica el patrón BFF en la construcción del proyecto SIGCON, con la finalidad de contenedorizar el despliegue del *frontend*, *backend* y almacenamiento en entornos virtualizados y desacoplados, con proyección a lograr resiliencia, escalamiento y elasticidad en el tiempo.
- La propuesta puede implementarse en otros dominios de negocios que requieran aplicar arquitecturas web modernas y componibles, utilizando el patrón BFF, modelo de servicio *cloud* CaaS. Estos enfoques permiten afrontar integración y entregas continuas de *software* en contextos de desarrollo acelerados.
- La implementación del proyecto se realizó con la participación de los estudiantes del curso Desarrollo de sistemas web de los semestres académicos 2023-I y 2023-II, de la escuela profesional de Ingeniería de Sistemas de la Universidad Nacional Mayor de San Marcos, y la docente y coordinadora del grupo de investigación Ingeniería web. Se aplicó la metodología ágil Scrum en la gestión del proyecto, la cultura DevOps en la formación de los equipos, con fines de lograr las competencias establecidas en el curso y el perfil de egreso del plan de estudios correspondiente.
- La investigación ha permitido formar recursos humanos según las tendencias actuales en ingeniería de *software*, área en constante evolución, con alta demanda laboral nacional e internacional. Estos profesionales requieren una formación actualizada, idónea y rigurosa, con un producto de valor como resultado.

## 7. TRABAJOS FUTUROS

De la presente investigación se desprenden algunas sublíneas no cubiertas, como la implementación de los patrones *service discovery* y *circuit breaker*, los cuales pueden ser abordados como trabajos futuros.

## REFERENCIAS

Almuairfi, S., & Alenezi, M. (2020). Security controls in infrastructure as code. *Computer fraud & security*, *2020*(10), 13-19. https://doi.org/10.1016/S1361-3723(20)30109-3

Balalaie, A., Heydarnoori, A., & Jamshidi, P. (2016). Microservices architecture enables DevOps: migration to a Cloud-Native architecture. *IEEE Software*, *33*(3), 42-52. https://doi.org/10.1109/MS.2016.64

Brown, K., & Woolf, B. (2016, octubre). *Implementation patterns for microservices architectures*. [presentación de paper]. PLoP'16: Proceedings of the 23rd Conference on pattern languages of programs, 1-35. https://www.hillside.net/plop/2016/papers/proceedings/papers/brown.pdf

Buchanan, I. (2024). *Infrastructure as code. How infrastructure as code (IaC) manages complex infrastructures*. Atlassian. https://www.atlassian.com/microservices/cloud-computing/infrastructure-as-code

Casa Grande. (2023, 12 de mayo). *Reglamento de convivencia de edificios y condominios*. https://www.administracionedificiosperu.com/2020/09/reglamento-de-convivencia-de-edificios.html

Docker (2024). *Use containers to build, share and run your applications.* https://www.docker.com/resources/what-container/

Duan, Y., Fu, G., Zhou, N., Sun, X., Narendra, N. & Hu, B. (2015). Everything as a service (Xaas) on the cloud: origins, current and future trends. *CLOUD'15: proceedings of the 2045 IEEE 8th International conference on cloud computing*, 621-628. https://doi.org/10.1109/CLOUD.2015.88

Eramo, R., Tucci, M., Di Pompeo, D., Cortellessa, V., Di Marco, A., & Taibi, D. (2024). Architectural support for software performance in continuous software engineering: a systematic mapping study. *Journal of systems and software*, *207*, 111833. https://doi.org/10.1016/j.jss.2023.111833

Espinoza, C. (2020, 22 de febrero). Crecimiento urbano y ciudades del futuro. *El Peruano.* https://elperuano.pe/noticia/90177-crecimiento-urbano-y-ciudades-del-futuro

Flask. (s. f.). *Flask. User's guide.* https://flask.palletsprojects.com/en/3.0.x/

FlaskSQLalchemy. (s. f.). *Flask SQLalchemy. User guide*. https://flask-sqlalchemy.palletsprojects.com/en/3.1.x/

Habala, O., Bobák, M., Šeleng, M., Tran, V., & Hluchý, L. (2023). Architecture of a function-as-a-service application. *Computing and informatics*, *42*(4), 878-895. https://doi.org/10.31577/cai_2023_4_878

Hussein, M. K., Mousa, M. H., & Alqarni, M. A. (2019). A placement architecture for a container as a service (CaaS) in a cloud environment. *Journal of cloud computing, 8*, 7. https://doi.org/10.1186/s13677-019-0131-1

Kratzke, N., & Quint, P-C. (2017). Understanding cloud-native applications after 10 years of cloud computing. A systematic mapping study. *Journal of systems and software*, *126*, 1-16. https://doi.org/10.1016/j.jss.2017.01.001

Larman, C. (1999). *UML y patrones: una introducción al análisis y diseño orientado a objetos*. Prentice Hall.

Mell, P., & Grance, T. (2012). *The NIST definition of cloud computing.* National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-145

Merelli, I., Fornari, F., Tordini, F., D'Agostino, D., Aldinucci, M., & Cesini, D. (2019). Exploiting Docker containers over grid computing for a comprehensive study of chromatin conformation in different cell types. *Journal of parallel and distributed computing*, *134,* 116-127. https://doi.org/10.1016/j.jpdc.2019.08.002

SQLAlchemy. (s. f.). *The Python SQL toolkit and object relational mapper*. https://www.sqlalchemy.org/

Stine, M. (2015). *Migrating to cloud-native application architectures*. O'Reilly Media.

Toffetti, G., Brunner, S., Blöchlinger, M., Spillner, J., & Bohnert, T. (2017). Self-managing cloud-native applications: design, implementation, and experience. *Future generation computer systems*, *72*, 165-179. https://doi.org/10.1016/j.future.2016.09.002

Vega, É. (2021, 5 de mayo). Crecimiento inmobiliario vertical de Lima muestra comportamientos diferenciados. *El Comercio*. https://elcomercio.pe/economia/negocios/crecimiento-inmobiliario-vertical-de-lima-muestra-comportamientos-diferenciados-mercado-inmobiliario-capeco-tinsa-nczenoticia/?ref=ecr

# Testing Asymmetric Encryption in a Sustainable Hacking Lab

Michael Dorin
mike.dorin@stthomas.edu
https://orcid.org/0000-0002-5798-0623
University of St. Thomas, St. Paul, MN. USA

Sergio Montenegro
sergio.montenegro@uni-wuerzburg.de
https://orcid.org/0000-0002-3958-3018
Julius-Maximilians-Universität Würzburg, Würzburg Germany

Received: April 15th, 2024 / Accepted: May 18th, 2024
doi: https://doi.org/10.26439/interfases2024.n19.7058

ABSTRACT. The Aerospace Information Technology Department (Computer Science VIII) at University of Würzburg explores many facets of aerospace systems, including secure telemetry and telecommand systems. Because satellites are expensive and indispensable, thorough protection and security research is necessary. Security algorithms are often processor-intensive, which can deprive payload applications of valuable execution cycles and even system power, thus making proper algorithm selection essential. A mechanism for execution and analysis on devices of similar capability to hardware systems used in space applications is essential for proper algorithm selection. This paper shows that it is possible to create an inexpensive and sustainable lab to efficiently and correctly test encryption algorithms and protocols using discarded tablet computers and inexpensive single-board computers. The lab constructed began by evaluating three public encryption key algorithms to assess computational space and time requirements. The three algorithms include an implementation of prime number-based Rivest-Shamir-Adleman (RSA) and two elliptic-curve cryptography-based key-exchange implementations. The initial results for the three algorithms show RSA memory requirements are not substantially different from the elliptic curve algorithms, but running times are comparatively slower. The first elliptic curve cryptography algorithm has moderate run time and space requirements, while the second one shows an improved run time but requires more space. This study reveals that testing algorithms using affordable lab devices can provide useful performance related data.

KEYWORDS: asymmetric encryption / sustainable hacking lab / satellite communications / RODOS

## PRUEBAS DE CIFRADO ASIMÉTRICO EN UN LABORATORIO DE *HACKING* SOSTENIBLE

RESUMEN. El Departamento de Tecnología de la Información Aeroespacial (Ciencias de la Computación VIII) de la Universidad de Würzburg explora muchos aspectos de los sistemas aeroespaciales, incluidos los sistemas seguros de telemetría y telemando. Debido a que los satélites son costosos e indispensables, es necesaria una investigación exhaustiva sobre su protección y seguridad. Los algoritmos de seguridad suelen requerir un uso intensivo de los procesadores, lo que puede privar a las aplicaciones de carga útil de ciclos de ejecución valiosos e incluso de energía del sistema. Por ello, es esencial una selección adecuada de los algoritmos que se utilizarán. Disponer de un mecanismo para la ejecución y el análisis, en dispositivos de capacidades similares a los sistemas y el hardware que se utilizan en las aplicaciones espaciales, es fundamental para una correcta selección de algoritmos. Este artículo muestra que es posible crear un laboratorio económico y sostenible para probar de manera eficiente y precisa los algoritmos de encriptación y protocolos utilizando tabletas descartadas y computadoras de placa única económicas. El laboratorio que se construyó con ellas se utilizó para evaluar tres algoritmos públicos de clave de encriptación para determinar los requisitos computacionales de espacio y tiempo. Los tres algoritmos incluyen una implementación del algoritmo de clave pública basado en números primos Rivest-Shamir-Adleman (RSA) y dos implementaciones de intercambio de clave basadas en criptografía de curva elíptica. Los resultados iniciales muestran que los requisitos de memoria de estos algoritmos no son sustancialmente diferentes, pero los tiempos de ejecución del algoritmo RSA son comparativamente más lentos. El primer algoritmo de criptografía de curva elíptica tiene requisitos moderados de tiempo de ejecución y espacio, mientras que el segundo muestra un tiempo de ejecución mejorado, pero requiere más espacio. Este estudio revela que probar algoritmos utilizando dispositivos de laboratorio asequibles puede proporcionar datos útiles acerca de su rendimiento.

PALABRAS CLAVE: encriptación asimétrica / laboratorio de hackeo sostenible / comunicaciones satelitales / RODOS

## 1. INTRODUCTION

The negative perception of hacking involves illicitly accessing computer systems and data (Sciglimpaglia Jr, 1991). Cybercriminals can use hacking to sabotage scientific and aerospace endeavors, hamper studies, breach security and endanger missions. Although researchers from the Aerospace Information Technology Department (Computer Science VIII) at the University of Würzburg have yet to encounter any security issues related to commanding spacecraft, as the cost of powerful hardware systems decreases, so does the cost of cyberattacks. This makes it necessary to build a hacking lab to start evaluating encryption algorithms.

Many researchers utilize virtual machines (VM) to create security labs. For example, Guarda et al. (2016) describe penetration testing using virtual environments, while Lee et al. (2023) describe simulation-based cybersecurity testing. As early as 2003, Garfinkel and Rosenblum (2003) were able to demonstrate a successful use of virtual machines for intrusion detection. VMs can work satisfactorily in many environments; however, since resources are shared overall, multiple VMs can create confusing time measurements when running different tests.

In regards to encryption algorithms, several papers in the current literature address encryption and telemetry/telecommand system (TM/TC) security. For example, López and Fraga (2016) describe a TM/TC encryption system using the Advanced Encryption Standard (AES). This paper suggests that AES is emerging as the new de-facto standard for satellite telemetry and telecommand systems. Another interesting paper on TM/TC security is from Hoang et al. This compelling paper discusses physical layer security (Hoang et al., 2021). Meanwhile, Herpel et al. (2016) address security by showing a software architecture for satellites. While these papers are helpful, they do not specifically address algorithms that are a good fit for the Realtime Onboard Dependable Operating System (RODOS). More information on RODOS can be found in Appendix A.

Regarding algorithms appropriate for RODOS, further examination of the literature covers more specific algorithm implementations. Many papers discuss the efficient implementation of Rivest-Shamir Adleman (RSA). Two examples from Nozaki et al. (2001) and Zhou and Tang. (2011) offer detailed analyses. There are also several papers on the implementation of elliptic curve cryptography (ECC). Point operations, for instance, can also be partially reduced by using the Hamming weight of the private key (commonly known as the multiplicand 'k'). In their work *Fast Algorithms for Common-Multiplicand Multiplication and Exponentiation by Performing Complements*, Chang et al. (2003) explain how to reduce the Hamming weight of the multiplicand which, in this case, can reduce point operations. Hamming weights have been discussed for more than fifty years and more information can be found in the work of Hamming (1970). Further research on using Hamming weights to save processing time has been conducted by various authors, including Kodali and Budwal (2013) and Huang et al (2010). Although these papers are interesting, they do not cover implementation details required by small devices.

Because previous research has not thoroughly addressed the security and encryption testing needs required for embedded applications in space using RODOS, this project explores the construction of a practical hacking lab to test the time it takes each of three asymmetric encryption algorithms to perform key exchange. This project selected specific RSA and ECC implementations having equivalent security levels when measured against symmetric encryption algorithms. In symmetric algorithms, if an n-bit key does not allow a faster attack than a brute force attack, then the algorithm is defined to have security level n (Lenstra, 2006). For this research, 80 bits of asymmetric encryption was determined to be the minimum acceptable level. RSA and ECC were selected to be compared at symmetric encryption levels of 80, 112, 128, 192 and 256 bits to evaluate algorithm performance. As such, corresponding RSA implementations of 1024, 2048, 3072, 7680 and 15360 bits were evaluated, while the corresponding curves B163, B233, B283, B409 and B571 were selected for ECC testing (Brown et al., 2001). Algorithms ECC were implemented in C using an embedded ECC library (Kokke, 2017). Algorithms for RSA were also implemented in C and a large number library (BigDigits multiple-precision arithmetic source code, s/f). This paper demonstrates that creating a hacking lab and testing algorithms for ground, air and space communications is feasible and practical.

## 2. PROJECT FOUNDATION AND ALGORITHMS

### 2.1 TM/TC and RODOS

Every satellite and spacecraft requires telemetry and command interfaces at one or more ground stations to receive payload data and to monitor and control the spacecraft. Telemetry (TM) refers to the process of collecting and transmitting data from the spacecraft to the ground station, where this data is visualized for operators and engineers. Telecommand (TC) interfaces allow operators to send commands and instructions from the ground station to the spacecraft (Maral et al., 2020).

Typically, TM and TC are developed separately, with one team developing the TM software for the spacecraft side and another team developing the TC counterpart for the ground station. However, RODOS employs an application generator called Corfu to specify telemetry and telecommands. Corfu generates the code for the spacecraft and for the ground station. For the spacecraft, it generates the distribution of telecommands and a frame to interpret and execute each of them, while for TM it generates the frame to collect, pack and send data from the spacecraft to ground station. For the ground station side, Corfu generates decoders and graphical representations for all telemetry data. It generates buttons to send commands and fields to enter their parameters. These commands and their parameters are then encoded and sent to the spacecraft.

### 2.2. Encryption Algorithms Tested

Saltzer and Schroeder (1975) suggested that security mechanisms should be as small as possible to be of aid in their verification, which is undoubtedly true in resource-starved environments. To allow for the creation of small mechanisms and achieve secure communication, a practical hacking lab was built, and RSA plus two variations of ECC were tested.

#### *2.2.1 Rivest-Shamir-Adleman*

The first algorithm tested was an RSA public key authentication. RSA was invented in 1977 and is based on the difficulty of factoring large composite numbers into their prime factors. The algorithm relies on a public-private key pair, where the public key is used for encryption and the private key is used for decryption (Salami et al., 2023). More basic information on RSA can be found in Appendix B. For this test, a complete RSA key exchange was simulated. RSA was selected for this test because of its perceived compactness and simplicity of implementation.

#### *2.2.2 Elliptic Curve Cryptography*

This paper does not describe the details of ECC, but information is provided in Appendix C. Two variants of ECC multiply were evaluated simulating a key exchange using Elliptic Curve Diffie-Helman (ECDH). More information on ECDH can be found in Appendix D. The first variant employs the hamming weight of the multiplicand to potentially reduce the overall number of operations required for the multiply. The second approach also employs Hamming weights, but in this variant an entry in the lookup table is created for each bit of the curve. For example, in the case of the 163-bit curve, 163 entries are created. As with standard multiplication, the private key 'k' is shifted to the right. However, instead of doing a 'double and add' when odd, the point value for the particular bit is found in the lookup table and is then added to a running sum, creating the product. This eliminates the need for the point to double each pass through the loop. As before, this algorithm can potentially be made even faster using Hamming weights. To take advantage of Hamming weights, an inverse value of k, inverseK, is subtracted from the next power of two above k, np. The resulting calculation is

$$\mathrm{np} = 2^{\,\mathrm{int}(\log 2(\mathrm{k}))+1}$$

The value of inverseK is calculated by subtracting k from the next power, as

$$\mathrm{inverseK} = \mathrm{np} - \mathrm{k}$$

Next, choose k or inverseK based on the Hamming weight. If inverseK is used, an extra subtraction is done at the end to transform the result to the original k, but the algorithm guarantees that the number of adds required is never greater than the number

of bits in k, though there is the Hamming math at the beginning and a point subtract required at the end when inverseK is used. Source code for this algorithm is provided in Appendix E. The algorithm demonstrated here is a refinement of the implementation made by Dorin (2009) and was inspired by work conducted in the 1960s when speeding up normal multiplication was required. For instance, Mitchell (1962) described computer multiplication and division using binary logarithms. More recently, Koshelev (2024) describes using Hamming weight for elliptic curve hashing algorithms. Nascimento et al. (2015) describe the use of lookup tables in their 8-bit implementation of ECC.

### 2.3 Simulation Equipment

Establishing a sustainable hacking lab requires completing common steps regardless of the algorithms being tested. The first action was acquiring suitable devices. For this study, both Linux-based devices and single-board computers were desired. Regarding Linux-based devices, systems were recommended to have at least 4 GB of RAM and at least 20 GB of persistent storage (The Linux Mint Team, 2024a). It was not difficult to find suitable devices as many surplus devices are sold on eBay for less than USD20 each. The final tablets arrived in the form of a collection of electronics generously donated to the University of St. Thomas. Those planning to set up an analogous lab can likely find equipment by contacting local electronics recyclers. Discarded electronics are often available at a low price or even at no cost. Recycling electronics can save money and increase the sustainability of the lab's construction. Specifications for the selected are available in Table 1. An image of a selected tablet is shown in Figure 1.

**Table 1**

*Tablet specifications*

| Component | Details |
|---|---|
| CPU Frequency | 1.4 Gigahertz |
| Processor Type | Intel Atom (Quad-Core) |
| Available Networking | Wi-Fi and Bluetooth |
| USB Ports | 1 External (3 total) |
| Available RAM | 4 Gigabytes |
| Persistent Storage | 30 Gigabytes |
| Operating System | Linux (Mint) |

The STM32F407-DISC1 was selected given that an actual embedded target was also desired for this project. According to the manufacturer, the STM32F407VGT6 microcontroller features a 32-bit Arm Cortex-M4 with an FPU core, 1-MB of Flash memory, and 192 KB of RAM (STMicroelectronics, 2024a). See Figure 2.

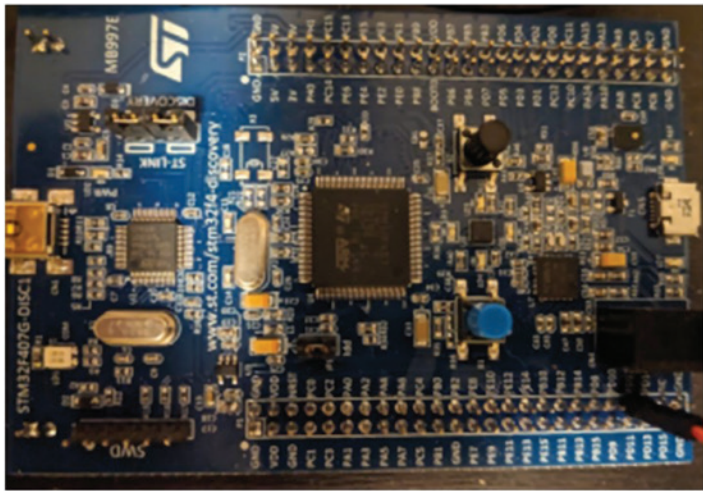
The tablets were re-imaged with the Mint distribution of the Linux operating system (The Linux Mint Team, 2024b). Mint was selected because it is stable, well-supported and easy to install. GNU compiler collection (GCC) tools were used for development on the Linux devices (Fenlason & Stallman, 1998). The STM32 Discovery board was used as-is out of the box. The STM32CubeIDE was used for development and project generation (STMicroelectronics, 2024b). Standard settings given by the STM32CubeIDE were not modified.

**Figure 1**

*The tablet device used for testing.*

**Figure 2**

*The STM32 discovery board used in testing*



## 3. RESULTS

On each hardware platform, key-exchange mathematics were performed on the algorithm being tested. This test code was executed continuously on the STM32 and an oscilloscope was used to measure timing. The key value used in the exchange remained unchanged for all tests. Table 2 displays a summary of results for the minimally acceptable security level, equivalent to 80 bits symmetric encryption, on both platforms.

**Table 2**

*Program response time for 80 bit security strength*

| Device | Build | Test Program Size (bytes) | Response Time (msec.) |
|---|---|---|---|
| STM32 | RSA | 37,980 | 3,810 |
| STM32 | ECDH Hamming Only | 28,072 | 2,730 |
| STM32 | ECDH Hamming and Lookup Table | 37,296 | 2,095 |
| Linux | RSA | 77,904 | 184.45 |
| Linux | ECDH Hamming Only | 36,776 | 46.13 |
| Linux | ECC Hamming and Lookup Table | 37,664 | 31.49 |

### 3.1 RSA

The first algorithm tested was RSA, using the STM device. Table 3 shows the required time for completion as well as the size of the executable. Note that the terms text, data, and bss refer to different sections of program memory. The text section contains executable code. The bss section contains uninitialized variables. The data section contains global and static variables (Pesch et al., 1993). Table 3 reports the STM32 results, Table 4 reports the results from the Linux system.

**Table 3**

*RSA key exchange simulation (STM32)*

| Security Strength | RSA bits | Program Size Bytes (text, data, bss) | Response Time (msec) |
|---|---|---|---|
| 80 | 1024 | 35,228 bytes | 3,810 |
| 112 | 2048 | 38,848 bytes | 27,970 |
| 128 | 3072 | 39,360 bytes | 90,400 |
| 192 | 7680 | 41,664 bytes | 844,926* |
| 256 | 15360 | 45,504 bytes | 3,824,620* |

***Note.*** * indicates that operation time is estimated.

**Table 4**

*RSA key exchange simulation (Linux)*

| Security Strength | RSA bits | Program Size Bytes (a.out) | Response Time (msec) |
|---|---|---|---|
| 80 | 1024 | 77,904 bytes | 184.45 |
| 112 | 2048 | 77,904 bytes | 1,350.69 |
| 128 | 3072 | 82,000 bytes | 4,399.72 |
| 192 | 7680 | 82,000 bytes | 65,894.26 |
| 256 | 15360 | 86,096 bytes | 516,835.02 |

### 3.2 ECDH Using Basic Multiply with Hamming Weight

The next algorithm tested was ECDH using ECC basic multiply with Hamming weight. This version of multiplication did not use a lookup table. STM32 results are shown in Table 5, the Linux results are shown in Table 6.

**Table 5**

*ECDH key exchange simulation (No lookup) (STM32)*

| Security Strength | ECC Algorithm | Program Size Bytes (text, data, bss) | Response Time (msec) |
|---|---|---|---|
| 80 | B163 | 28,156 | 312.94 |
| 112 | B233 | 28,176 | 545.0 |
| 128 | B283 | 28,192 | 720.0 |
| 192 | B409 | 28,228 | 1,475.0 |
| 256 | B571 | 28,244 | 2,570.0 |

**Table 6**

*ECDH key exchange simulation (No lookup) (Linux)*

| Security Strength | ECC Algorithm | Program Size Bytes (a.out) | Response Time (msec) |
|---|---|---|---|
| 80 | B163 | 36,776 | 46.13 |
| 112 | B233 | 36,800 | 84.5 |
| 128 | B283 | 36,800 | 113.11 |
| 192 | B409 | 36,816 | 232.17 |
| 256 | B571 | 36,840 | 444.39 |

### 3.3 ECDH Using Basic Multiply with Hamming Weight and Lookup Table

The final algorithm tested was ECDH using ECC basic multiply with Hamming weight and lookup table. Table 7 reports STM32 results, Table 8 shows results with the Linux system.

**Table 7**

*ECDH key exchange simulation (Lookup Table) (STM32)*

| Security Strength | ECC Algorithm | Program Size Bytes (text, data, bss) | Response Time (msec) |
|---|---|---|---|
| 80 | B163 | 37,012 | 214.78 |
| 112 | B233 | 44,492 | 372.3 |
| 128 | B283 | 48,872 | 495.38 |
| 192 | B409 | 71,464 | 1,005.0 |
| 256 | B571 | 111,224 | 1,755.0 |

**Table 8**

*ECDH key exchange simulation (Lookup Table) (Linux)*

| Security Strength | ECC Algorithm | Program Size Bytes (a.out) | Response Time (msec) |
|---|---|---|---|
| 80 | B163 | 37,664 | 31.49 |
| 112 | B233 | 44,832 | 58.25 |
| 128 | B283 | 49,184 | 78.44 |
| 192 | B409 | 75,808 | 156.57 |
| 256 | B571 | 119,584 | 306.92 |

### 3.4 Power Consumption

Power consumption was measured during all of the tests. On Linux, irrespective of algorithm, PowerTOP (Zeitouny & Akturan, 2013) estimated 1.53 Watts of power consumption and 100 % CPU utilization. On the STM-32 device, it was possible to measure power usage by putting a meter in-line with the power supply. Power consumption varied between 47.3 mA and 52.9 mA. STM32 Results are shown in Table 9.

**Table 9**

*Power consumption STM32*

| Security Strength | ECC – No lookup (mA.) | ECC – Lookup (mA.) | RSA (mA.) |
|---|---|---|---|
| 80 | 47.3 | 47.4 | 52.9 |
| 112 | 47.0 | 46.9 | 51.0 |

(*continues*)

(*continued*)

| Security Strength | ECC – No lookup (mA.) | ECC – Lookup (mA.) | RSA (mA.) |
|---|---|---|---|
| 128 | 47.5 | 47.3 | 49.4 |
| 192 | 47.5 | 47.5 | NA |
| 256 | 47.3 | 47.8 | NA |

Setting up the lab hardware was trouble-free, but future work may require performance tuning of the STM32 board. Setting up the development systems for the STM32 board and the Linux system went smoothly. While the STM32 and Linux device did run all the required applications, the STM32 required substantial time to process large numbers. It was necessary to use quadratic regression to estimate response times for RSA-7680 and RSA-15360.

When it came to space requirements, the RSA algorithm was expected to be the best choice given that the basic algorithm is comparatively uncomplicated. However, the ECC B163 algorithm without a lookup table turned out to be the best choice in terms of space. This was due to the space required by the large integer math routines needed for implementation. Hamming ECC with lookup table multiply demanded the most space, with ECC multiply without the lookup table falling in between.

Considering the required time per operation, ECC using Adaptive Hamming with lookup was the fastest of all three algorithms. Calculations were approximately 30 % faster than tests ran without the lookup table on both the STM32 and Linux boards. RSA was the slowest. ECDH with basic ECC multiply again fell in between. Both elliptic curve algorithms were faster than the equivalent RSA algorithm for all tests.

In terms of power, there does appear to be a small advantage for using ECC algorithms instead of traditional RSA. However, this advantage is marginal and warrants further exploration.

## 4. CONCLUSIONS

This research had multiple goals. The first goal was to determine whether it is possible to create a practical hacking lab to test security algorithms and provide quality data at an affordable price. While validating the first goal, a second goal was established to measure different asymmetric encryption algorithms using the aforementioned hacking lab. To this end, a hacking lab was created from STM32 boards and discarded tablet PCs. The selected devices ran the necessary software and performed tests on RSA and two variants of ECC multiple encryption algorithms.

Developing this hacking lab shows that a practical, inexpensive and sustainable solution for enhancing security in aerospace environments is possible. Several avenues for further research are now open. For instance, formally testing key exchanges and

synchronous algorithms could prove essential for future projects. In addition, testing block versus streaming ciphers for applications specific to RODOS TM/TC could also prove fruitful. Finally, further research could be directed toward refining the testing procedures and expanding the scope of the lab's capabilities. The results demonstrated in this project indicate a promising future for research on hacking labs.

## 5 APPENDICES

### Appendix A: RODOS

Realtime Onboard Dependable Operating System (RODOS) is a real-time operating system for embedded systems. It was designed for space applications but is suitable for any application demanding high dependability. An important feature of RODOS is its integrated real-time middleware. Developing the control and payload software on top of middleware provides maximum modularity for developers. Applications/modules can be developed independently and can be easily swapped without worrying about side effects since all modules are encapsulated as Building Blocks (BB) and can be accessed by other resources and access other resources through well-defined interfaces.

RODOS was implemented as a software framework in C++ with an object-oriented application interface. It is organized into layers: the lowest layer (1) is responsible for controlling the embedded system hardware (HAL: Hardware abstraction layer). The next layer (2), the kernel, administers local resources, threads and time. On top of the kernel is the middleware (layer 3), which enables communication between BBs using a publisher-subscribe multicast protocol. On top of the middleware sit user-implemented applications (layer 4), which are distributed software networks of simple BBs. The BBs API on the top of the middleware is a service-oriented interface. BBs interact by providing services to other BBs and using services from other BBs.

RODOS can be executed on different hardware platforms (arm, PPC, X86, Spark, Leon) and on top of Linux or Posix operating systems. It is possible to test and simulate target applications on various devices using RODOS on top of Linux.

### Appendix B: Rivest-Shamir-Adleman (RSA)

RSA encryption is used to secure data transmission, employing a public key for encryption and a private key for decryption. Using the public key, anyone can encrypt a message. However, the private key is required to decipher them.

*Key generation*

1. Choose two large prime numbers, $p$ and $q$
2. Compute their product, $n = pq$, which becomes the modulus for both keys.

3. Calculate the totient function $\phi(n) = (p - 1)(q - 1)$.
4. Choose an integer $e$ such that $1 < e < \phi(n)$ and $\gcd(e, \phi(n)) = 1$. This becomes the public exponent.
5. Compute the modular inverse of $e$ modulo $\phi(n)$ to get $d$, which is the private exponent.

– To encrypt a message $M$, compute $C \equiv M^e \bmod n$.

– To decrypt the ciphertext $C$, compute $M \equiv C^d \bmod n$.

RSA's security relies on the difficulty of factoring the modulus $n$ into its prime factors. As long as the prime factors $p$ and $q$ are sufficiently large, factoring $n$ and breaking the RSA encryption is computationally infeasible (Salami et al., 2023).

**Appendix C: Elliptic Curve Cryptography**

Elliptic Curve Cryptography (ECC) is a means of performing asymmetric encryption with faster processing times than other algorithms. ECC leverages the properties of elliptic curves over finite fields and can be used to provide encryption and create digital signatures. An advantage to ECC is the ability to offer equivalent security to traditional cryptographic algorithms like RSA albeit with smaller key sizes, making it more efficient for resource-constrained environments. Elliptic curves used in ECC are defined by equations of the form $y2 = x3 + ax + b.$, where a and b are coefficients chosen based on specific mathematical properties. These curves exhibit various properties, including having a group structure defined by point addition and scalar multiplication (Edoh, 2004).

The simplest form of point multiplication is commonly called "the double and add method" and is almost effortless to implement (Eisentraeger et al., 2002). Algorithm 1 below illustrates how to multiply the scalar value $k$ by a specific point on the curve, $P$.

Input: Scalar multiplier $k$, point $P$

Output: Resultant point $Q = kP$

Set $Q = \Phi$ ; // point-at-infinity

while ($k \neq 0$) do

if ($k$ *is odd*) then

$Q = Q + P$;

end

$P = 2P$;

$k = k/2$;

end

return $Q$;

The scalar *k* is shifted right for each iteration through the loop. Whenever bit zero of *k* is one, the current point value is added to the output. The point value *P* is doubled during each pass through the loop until *k* is zero, when the point *Q* is returned (Eisentraeger et al., 2002; Anoop, 2007).

**Appendix D: Elliptic Curve Diffie-Helman (ECDH) Algorithm**

1. Alice selects a (secret) random natural number *a*, calculates P = *a* * G. P is sent to Bob.

   *a* is Alice's private key.

   *P* is Alice's public key.

2. Bob selects a (secret) random natural number *b*, calculates Q = b * G. Q is sent to Alice.

   *b* is Bob's private key.

   Q is Bob's public key.

3. Alice calculates S = a * Q = a * (b * G).

4. Bob calculates T = b * P = b * (a * G).

   T = S = the new shared secret.

   More information can be found in RFC4492 (Blake-Wilson et al., 2006)

   and Kokke (2017).

**Appendix E: Elliptic Curve Example Source Code**

```
#include "ecdh.h"
void gf2point_hamming_and_lookup_mutliply(gf2elem_t x,
        gf2elem_t y, const scalar_t exp)
{
    scalar_t one = {1};  //scalar value one, used for shifting
    scalar_t nextPowerOfTwoAbove = {0};
    scalar_t result = {0};
    int numberofHighestOneBit;

    numberofHighestOneBit = bitvec_degree(exp);
    /* bitvec_degree holds number of the highest one-bit + 1 */
    bitvec_lshift(nextPowerOfTwoAbove, one, numberofHighestOneBit );
    int hammingWeight = hamming_weight(exp);
    bigint_subtract(result, nextPowerOfTwoAbove, exp);
```

```
    int hammingWight2 = hamming_weight(result);

if (hammingWeight < hammingWight2) {
   gf2point_multiply_with_lookup(x, y, exp);
} else {
    gf2point_multiply_with_lookup(x, y, result );
    gf2field_add(y,x,y );
    gf2point_add(x, y, Tx[numberofHighestOneBit+1],
    Ty[numberofHighestOneBit+1]);
    }
}
void gf2point_multiply_with_lookup(gf2elem_t x, gf2elem_t y,
     const scalar_t exp)
{
    int index = 0;
    int currentBit = 0x1;
    int currentByte = 0;

    for (index = 1; index < ECC_PRV_KEY_SIZE * 8; index++) {

        if (exp[currentByte] & currentBit) {
          gf2point_add(x, y,Tx[index], Ty[index]);
        }
        currentBit =currentBit << 1;
        if (currentBit == 0x00000000)
        {
          currentBit = 0x1;
          currentByte+=1;
        }
    }
}
```

*Note*. Requires tiny-ECDH-c from Kokke (2017).

## REFERENCES

Anoop, M. S. (2007). Elliptic curve cryptography. An implementation guide. https://informatika.stei.itb.ac.id/~rinaldi.munir/Kriptografi/2014-2015/ECC_Tut_v1_0.pdf

BigDigits multiple-precision arithmetic source code (s/f). *DI Management*. https://www.di-mgt.com.au/bigdigits.html

Blake-Wilson, S., Nystrom, M., Hopwood, D., Mikkelsen, J. & Wright, T. (2006). *Network working group*. https://www.rfc-editor.org/pdfrfc/rfc4366.txt.pdf

Brown, M., Hankerson, D., López, J. & Menezes, A. (2001). Software implementation of the NIST elliptic curves over prime fields. In D. Naccache (Ed.), *Topics in Cryptology — CT-RSA 2001.* (pp. 250-265). Springer. https://doi.org/10.1007/3-540-45353-9_19

Chang, C-C., Kuo, Y-T. & Lin, C-H. (2003). Fast algorithms for common-multiplicand multiplication and exponentiation by performing complements. *Proceedings 17th International Conference on Advanced Information Networking and Applications (AINA)* (pp. 807-811). IEEE Computer Society. https://doi.org/10.1109/AINA.2003.1193005

Dorin, M. (2009). *Implementation of standards based public key cryptography for small processor based systems* [Master's thesis] Metropolitan State University, St. Paul, Minnesota.

Edoh, K. D. (2004). Elliptic curve cryptography: Java implementation. *Proceedings of the 1st Annual Conference on Information Security Curriculum Development* (pp. 88-93). Association for Computing Machinery. https://doi.org/10.1145/1059524.1059542

Eisentraeger, K., Lauter, K. & Montgomery, P. L. (2002). An efficient procedure to double and add points on an elliptic curve. *Cryptology ePrint Archive, paper 2002/112*. https://eprint.iacr.org/2002/112.

Fenlason, J. & Stallman, R. (1998). *The GNU Profiler*. https://ftp.gnu.org/old-gnu/Manuals/gprof-2.9.1/html_mono/gprof.html

Garfinkel, T. & Rosenblum, M. (2003). A virtual machine introspection based architecture for intrusion detection. *Network and Distributed System Security Symposium*, *3*. https://suif.stanford.edu/papers/vmi-ndss03.pdf

Guarda, T., Orozco, W., Augusto, M. F., Morillo, G., Arévalo Navarrete, S. & Mota Pinto, F. (2016). Penetration testing on virtual environments. In: *Proceedings of the 4th International Conference on Information and Network Security (ICINS '16)* (pp. 9-12). https://doi.org/10.1145/3026724.3026728

Hamming, R. W. (1970). On the distribution of numbers. *Bell System Technical Journal*, *49*(8), 1609-1625. https://doi.org/10.1002/j.1538-7305.1970.tb04281.x

Herpel, H-J., Kerep, M., Montano, G., Eckstein, K., Schön, M. & Krutak, A. (2016). MILS compliant software architecture for satellites. *MILS@HiPEAC*. https://core.ac.uk/download/pdf/144785917.pdf

Hoang, T. M., Duong, T. Q., Tuan, H. D., Lambotharan, S. & Hanzo, L. (2021). Physical layer security: detection of active eavesdropping attacks by support vector machines. *IEEE Access*, *9*, 31595-31607. https://doi.org/10.1109/ACCESS.2021.3059648

Huang, X., Shah, P. G. & Sharma, D. (2010). Minimizing hamming weight based on 1's complement of binary numbers over GF ($2^m$)). *12th International Conference on Advanced Communication Technology (ICACT)*, 1226-1230. https://researchsystem.canberra.edu.au/ws/portalfiles/portal/28927012/full_text_published_15.pdf

Kodali, R. K. & Budwal, H. S. (2013). High performance scalar multiplication for ECC. *2013 International Conference on Computer Communication and Informatics* (pp. 1-4). https://doi.org/10.1109/ICCCI.2013.6466286

Kokke. (2017). *Small and portable implementation of ECDH in C*. https://github.com/kokke/tiny-ECDH-c

Koshelev, D. (2024), *Some remarks on how to hash faster onto elliptic curves*. *Journal of Computer Virology and Hacking Techniques*. (2024). https://doi.org/10.1007/s11416-024-00514-4

Lee, D. H., Kim, C. M., Song, H. S., Lee, Y. H. & Chung, W. S. (2023). Simulation-based cybersecurity testing and evaluation method for connected car V2X application using virtual machine. *Sensors*, *23*(3), 1421. https://doi.org/10.3390/s23031421

Lenstra, A. (2006). *Key lengths contribution to the handbook of information security*. https://blkcipher.pl/assets/pdfs/NPDF-32.pdf

López, D. & Fraga, E. (2016). Tm/tc encryption system. In: *14th International Conference on Space Operations*, Article 2330. American Institute of Aeronautics and Astronautics. https://arc.aiaa.org/doi/10.2514/6.2016-2330

Maral, G., Bousquet, M. & Sun, Z. (2020). *Satellite communications systems: systems, techniques and technology*. Wiley.

Mitchell, J. N. (1962). Computer multiplication and division using binary logarithms. *IRE Transactions on Electronic Computers*, *EC-11*(4), 512–517. https://doi.org/10.1109/TEC.1962.5219391

Nascimento, E., López, J. & Dahab, R. (2015). Efficient and secure elliptic curve cryptography for 8-bit AVR microcontrollers. In R. Chakraborty, P. Schwabe & J. Solworth (Eds.) *Security, privacy and applied cryptography engineering. Lecture notes in computer science, 9354*, pp. 289-309. Springer. https://doi.org/10.1007/978-3-319-24126-5_17

Nozaki, H., Motoyama, M., Shimbo, A. & Kawamura, S. (2001). Implementation of RSA algorithm based on RNS Montgomery multiplication. In C. K. Koc, D. Naccache & C. Paar (Eds.), *Cryptographic hardware and embedded systems—CHES 2001. Lecture Notes in Computer Science*, *2162*, 364-376. Springer. https://doi.org/10.1007/3-540-44709-1_30

Opus IVS. (2024). Opus IVS.About Us https://www.opusivs.com/about/

Pesch, R. H., Osier, J. M. & Support, C. (1993). *The Gnu binary utilities*. https://web.mit.edu/gnu/doc/html/binutils_toc.html

Salami, Y., Khajehvand, V. & Zeinali, E. (2023). Cryptographic algorithms: a review of the literature, weaknesses and open challenges. *Journal of Computer & Robotics*, *16*(2), 63-115. https://doi.org/10.22094/jcr.2023.1983496.1298

Saltzer, J. H. & Schroeder, M. D. (1975). The protection of information in computer systems. *Proceedings of the IEEE*, *63*(9), 1278–1308. https://doi.org/10.1109/PROC.1975.9939

Sciglimpaglia Jr., R. J. (1991). Computer hacking: a global offense. *Pace International Law Review*, *3*(1), 204-266. https://doi.org/10.58948/2331-3536.1020

STMicroelectronics. (2024a). STM32F4DISCOVERY - Discovery kit with STM32F407VG MCU. https://www.st.com/en/evaluation-tools/stm32f4discovery.html

STMicroelectronics. (2024b). STM32CubeIDE - Integrated development environment for STM32. https://www.st.com/en/development-tools/stm32cubeide.html

The Linux Mint Team. (2024a), Linux Mint - FAQ. Linux Mark Institute. https://linuxmint.com/faq.php

The Linux Mint Team. (2024b), Linux Mint - Download. Linux Mark Institute. https://www.linuxmint.com/download.php

Zeitouny, C. & Akturan, C. (2013). *Linux* power efficiency analysis methods. A look at power efficiency analysis methods under Linux environments*. Intel corporation. https://www.intel.com/content/dam/develop/external/us/en/documents/linux-power-efficiency-analysis-methods-2.pdf

Zhou, X. & Tang, X. (2011). Research and implementation of RSA algorithm for encryption and decryption. *Proceedings of 2011 6th international forum on strategic technology* (pp. 1118-1121). https://doi.org/10.1109/IFOST.2011.6021216

# Análisis de la brecha entre la universidad y la industria del *software* en la República Argentina: una perspectiva docente y posibles soluciones

Marcelo López-Nocera
mlopeznocera@gmail.com
https://orcid.org/0009-0006-8102-8639
Universidad Tecnológica Nacional, Argentina

María F. Pollo-Cattaneo
flo.pollo@gmail.com
https://orcid.org/0000-0003-4197-3880
Universidad Tecnológica Nacional, Argentina

Francisco Redelico
francisco.redelico@gmail.com
https://orcid.org/0000-0002-6945-2916
Instituto de Medicina Traslacional e Ingeniería Biomédica, Argentina

Recibido: 15 de abril del 2024 / Aceptado: 18 de mayo del 2024
doi: https://doi.org/10.26439/interfases2024.n19.7061

RESUMEN. Se realiza un estudio de campo para analizar la brecha existente entre universidad y empresa. El objetivo específico es la identificación de las variables relevantes en el nivel cognitivo según la teoría institucional (como motivaciones, creencias, conceptos y percepciones) para los agentes estudiados (en este caso, docentes universitarios) y su cuantificación. En ese sentido, se delimita el universo de estudio a la Universidad Tecnológica Nacional y a las carreras relacionadas con la industria del *software*, para así analizar la posibilidad de reducir la brecha y mejorar la relación entre ambas.

PALABRAS CLAVE: spin-off / UIC / colaboración entre industria y universidad / industria del *software*, espacios intersticiales, teoría institucional

## ANALYSIS OF THE GAP BETWEEN UNIVERSITY AND SOFTWARE INDUSTRY IN ARGENTINA: A TEACHING PERSPECTIVE AND POTENTIAL SOLUTIONS

ABSTRACT. This study examines the gap between universities and companies. It specifically aims to identify relevant cognitive variables according to institutional theory (such as motivations, beliefs, concepts, and perceptions) for the agents studied

(in this case, university professors) and quantify these variables. The research focuses on the National Technological University and the programs related to the software industry to explore ways to reduce the gap and improve the relationship between the two sectors.

KEYWORDS: spin-off / UIC / software industry / interstitial spaces / institutional theory

## 1. INTRODUCCIÓN

La interacción colaborativa entre universidad y empresa (UIC) ha aumentado considerablemente en la última década en todo el mundo (Bahdanava et al., 2024), incluyendo la República Argentina (Di Meglio, 2024), donde la industria del *software* ha crecido aproximadamente en un 15 % durante ese periodo (Krepki, 2024). Esto ha generado un aumento de la demanda estudiantil en los últimos veinte años por las carreras universitarias relacionadas con los sistemas de información, de modo tal que la inscripción en esas carreras representa aproximadamente el 5 % de la matrícula total universitaria (Lauric et al., 2024). Existe, además, en el entorno social, la conciencia sobre el fortalecimiento del vínculo entre universidad y empresa, dada la necesidad de la propia sociedad de beneficiarse de los resultados de la investigación académica (Jiménez & Castellanos, 2008).

Sin embargo, se puede observar que ese crecimiento de la matrícula no acompaña cuantitativamente la demanda de la industria y también que existe un desfase cualitativo, dado que no hay suficiente oferta de recursos humanos con la formación requerida para cubrir perfiles técnicos específicos (Grosso, 2019). En este contexto, se identifica una brecha entre lo que el mercado demanda y lo que la academia universitaria ofrece. Se considera un interesante objeto de estudio la identificación de las principales variables que inciden en la conformación estructural de dicha brecha. El objetivo del presente trabajo es, por consiguiente, analizar las características de la brecha existente entre lo que la industria del *software* demanda a la universidad y lo que ella ofrece, y viceversa. La investigación se sitúa en el contexto actual de la Universidad Tecnológica Nacional, Facultad Regional Buenos Aires (UTN FRBA), desde la perspectiva del docente universitario y de las dificultades que encuentra para su vinculación efectiva con dicha industria. Debido a la importancia del enfoque de los docentes para el tema abordado, para un primer análisis se llevó a cabo un estudio de campo que consistió en un conjunto de entrevistas con un cuestionario previamente elaborado.

El presente artículo se estructura de la siguiente manera: estado de situación de la cuestión abordada (sección 2), metodología utilizada para llevar adelante la investigación (sección 3), resultados obtenidos (sección 4), conclusiones elaboradas a partir de dichos resultados (sección 5), propuesta de líneas de investigación futuras (sección 6) y referencias bibliográficas consultadas (sección 7).

## 2. ESTADO DE SITUACIÓN

Furnari (2014) define el concepto de espacio intersticial (EI) como una oportunidad de interacción —incluso esporádica— y de fomento de nuevas prácticas entre dos mundos diversos, pero con intereses comunes. La teoría institucional (TI) clásica (Scott, 1987), en cambio, se orienta a la consolidación de prácticas preexistentes. Por otra parte, es aceptado que la educación universitaria juega un rol preponderante en el desarrollo de las *spin-off* (SO) (Hernández et al., 2015), así como en su papel de moderadora y

vinculadora del sistema conocido como la triple hélice (Etzkowitz & Leydesforff, 1997), compuesta por universidad, empresa y Estado. Estas SO funcionan como un ejemplo de aplicación práctica de los EI. A su vez, la universidad emprendedora —la cual se puede definir, siguiendo a Gibson y Foss (2017), como aquella que emprende acciones comerciales en favor de sus desarrollos científicos y tecnológicos, y que enseña y promueve el emprendedurismo— puede ser considerada un sistema complejo en el que un conjunto de agentes heterogéneos entre sí (docentes, investigadores, empresarios, funcionarios estatales y agentes de la sociedad civil) interactúan y conforman una cuádruple hélice (Alderete et al., 2020). Zachman y Redchuk (2016) han estudiado los vínculos que conforman dicho sistema en la República Argentina. Taucean et al. (2018), por otro lado, analizan las distintas aproximaciones de las *higher education institutions* al emprendedurismo. Bergenholtz y Bjerregaard (2014) estudian la importancia de los lazos (débiles o fuertes) y cómo estos se relacionan con las condiciones institucionales. En Arza y Vazquez (2010) se exploran, asimismo, los canales de interacción más efectivos para generar diferentes beneficios entre docentes, investigadores y empresas; sus resultados plantean diferentes cuestiones, una de las cuales es que se requiere un cambio en los incentivos para los investigadores, para que las interacciones generen nuevos beneficios. Los hallazgos de la investigación de Peksatici y Ergun (2019) revelan, adicionalmente, que las diferentes lógicas institucionales de la industria y de los programas educativos dan lugar a una brecha entre lo que los gerentes de la industria valoran y lo que ofrecen las instituciones universitarias. Massaro (2016), finalmente, hace una indirecta referencia a los EI y a las SO: señala que la vinculación tecnológica tampoco es ajena a la necesidad de federalizar políticas y que, por ello, son las universidades las que —por su propia naturaleza autárquica— presentan las mejores condiciones para adaptar sus políticas internas de vinculación a la realidad del progreso regional, aportando a tal efecto un espacio común (EI) para la promoción del conocimiento y el despliegue de la creatividad, para su puesta en valor. Esto puede hacerse ya sea a través de sistemas de vinculación tecnológica con el tejido empresarial e industrial regional como mediante la promoción de efectivos procesos de desarrollo y expansión de nuevas empresas tecnológicas (SO), a fin de nutrir ese tejido con mayor competitividad y productividad.

En este contexto, resulta de interés el análisis de la brecha identificada, sus características y las principales variables que inciden en su conformación estructural. El estudio se particulariza en la UTN FRBA actual, desde la perspectiva del docente universitario y de las dificultades que este encuentra para la vinculación efectiva de la academia con la industria.

## 3. METODOLOGÍA

En base a métodos de recolección de datos y de análisis estadístico utilizados previamente por diversos autores (Gibson & Foss, 2017; Taucean et al., 2018; Ortiz, 2019), se elaboraron encuestas y entrevistas como herramientas de investigación. Estas fueron

realizadas durante el año 2023, de acuerdo con el diseño curricular y al plan de estudios de la carrera de Ingeniería en Sistemas de Información —aprobados por el consejo superior de la UTN FRBA, con la Ordenanza 1877 del 15 de junio del 2022 (Universidad Tecnológica Nacional, 2022)— para los ingresantes a partir de 2023. Se aplicó un cuestionario de 88 preguntas (disponibles en el Apéndice A) a una población de 212 docentes del Departamento de Ingeniería en Sistemas de Información de la UTN FRBA, sobre un universo total de aproximadamente 300 personas. Al respecto, se recomienda disponer del listado de preguntas del cuestionario para su consulta durante la lectura del presente artículo. A los efectos del análisis cuantitativo, en la Tabla 1 se describen las preguntas de investigación planteadas para el presente estudio, con sus respectivas motivaciones:

**Tabla 1**

*Preguntas de investigación y motivación*

| Preguntas de investigación (PI) | Motivaciones investigativas (MI) |
|---|---|
| PI1) ¿Qué distribución en datos personales tienen los docentes universitarios estudiados y de qué modo influye eso en su relación con la industria? | MI1) Conocer la distribución estadística (rango de edad y género) de los docentes que actúan en el vínculo e indagar sobre su incidencia |
| PI2) ¿Qué tipo de experiencia o formación curricular tienen los docentes que interactúan con la industria? | MI2) Conocer cómo influye la formación curricular en el vínculo |
| PI3) ¿Qué vinculación/tipo de vínculo se utiliza con la industria? | MI3) Conocer los distintos tipos de vínculos que los docentes establecen con la industria y su incidencia en ellos |
| PI4) ¿Cuáles son las características y connotaciones de ese vínculo? | MI4) Conocer las características de cada tipo de vínculo y la incidencia de los docentes en ellos |
| PI5) ¿Qué conocimiento se tiene acerca de spin-off y en qué grado se las gestiona o utiliza? | MI5) Cuantificar el grado de conocimiento y gestión que los docentes tienen acerca de las spin-off |
| PI6) ¿Qué conocimiento se tiene acerca de la actividad científica argentina y en qué grado se interactúa con ella? | MI6) Cuantificar el grado de conocimiento de los docentes acerca de la actividad científica argentina y medir el nivel de interacción entre el mundo universitario y científico y entre el mundo científico y empresarial (desde la perspectiva docente). |

Las relaciones existentes con las preguntas de investigación planteadas permiten establecer agrupamientos para cada una de las preguntas efectuadas en las entrevistas, los que se representan en la Tabla 2.

**Tabla 2**

*Grupos de preguntas de las entrevistas relacionados con cada pregunta de investigación*

| Grupos de preguntas relacionadas en las entrevistas | Preguntas de investigación con las que se encuentra relacionado el grupo |
|---|---|
| Grupo 1: Desde la pregunta 1 a la pregunta 4 inclusive | PI1) |
| Grupo 2: Desde la pregunta 5 a la pregunta 10 inclusive | PI1) y PI2) |
| Grupo 3: Desde la pregunta 11 a la pregunta 20 inclusive | PI2) y PI3) |
| Grupo 4: Desde la pregunta 21 a la pregunta 41 inclusive | PI2), PI3) y PI4) |
| Grupo 5: Desde la pregunta 42 a la pregunta 61 inclusive | PI3) y PI4) |
| Grupo 6: Desde la pregunta 62 a la pregunta 68 inclusive | PI3), PI4) y PI5) |
| Grupo 7: Desde la pregunta 69 a la pregunta 78 inclusive | PI4), PI5) y PI6) |
| Grupo 8: Desde la pregunta 79 a la pregunta 88 inclusive | PI2), PI3), PI4), PI5) y PI6) |

Las respuestas dadas por cada docente pueden consultarse en el Apéndice B. Con el fin de consolidar la información con base en las respuestas recabadas, se presentan en la siguiente sección los resultados obtenidos para cada uno de los agrupamientos.

## 4. RESULTADOS OBTENIDOS

Para un mejor análisis de los resultados, se definen en esta sección nueve apartados, siendo los ocho primeros correspondientes a cada uno de los grupos definidos anteriormente (en la Tabla 2). A saber: 4.1 (grupo 1), 4.2 (grupo 2), 4.3 (grupo 3), 4.4 (grupo 4), 4.5 (grupo 5), 4.6 (grupo 6), 4.7 (grupo 7) y 4.8 (grupo 8). El noveno apartado (4.9) corresponde a la discusión misma de los resultados obtenidos, los que se desglosan a continuación.

### 4.1 Grupo 1

En la Figura 1 se muestran los resultados obtenidos para el grupo 1, al que pertenecen las preguntas 1 a 4 (P1-P4).

A partir de las respuestas se puede inferir, efectuando un análisis de Pareto (Sales, 2021), que la gran mayoría de la población entrevistada corresponde a personas de género masculino, en un rango de 50 a 65 años, con experiencia en enseñanza en carreras de grado en universidades públicas y sin experiencia en enseñanza en carreras de posgrado. Aquellos que sí tienen esa experiencia, en su mayoría, la adquirieron tanto en universidades privadas como en públicas.

### 4.2 Grupo 2

A continuación, en la Figura 2 se muestran los resultados para el rango de preguntas que va de la 5 a la 10 (P5-P10).

**Figura 1**

*Distribución de datos personales de los docentes entrevistados*

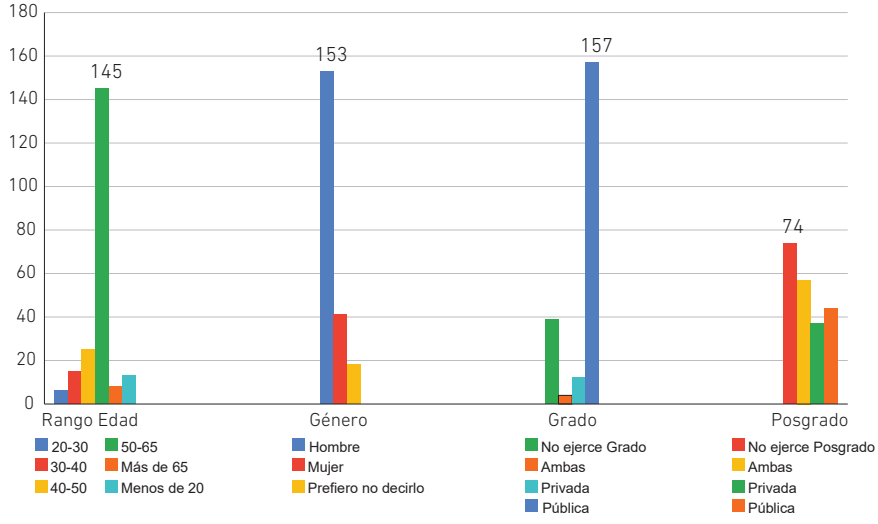
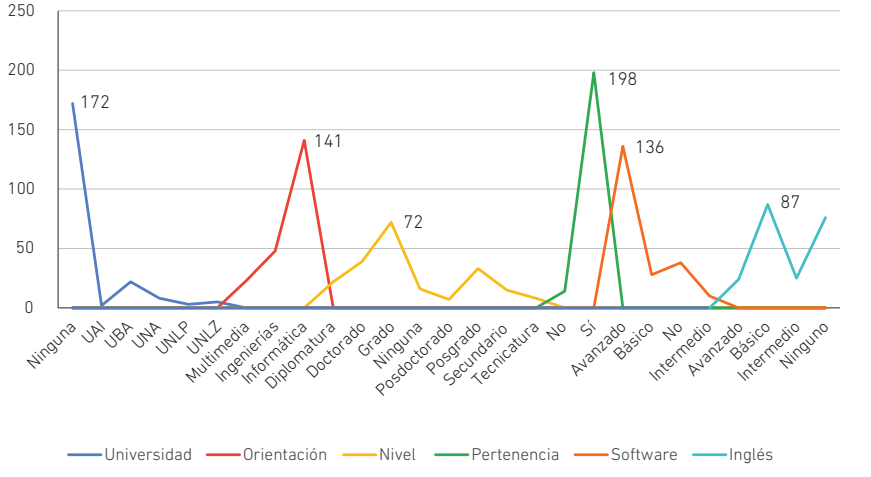



**Figura 2**

*Formación curricular de los docentes entrevistados*



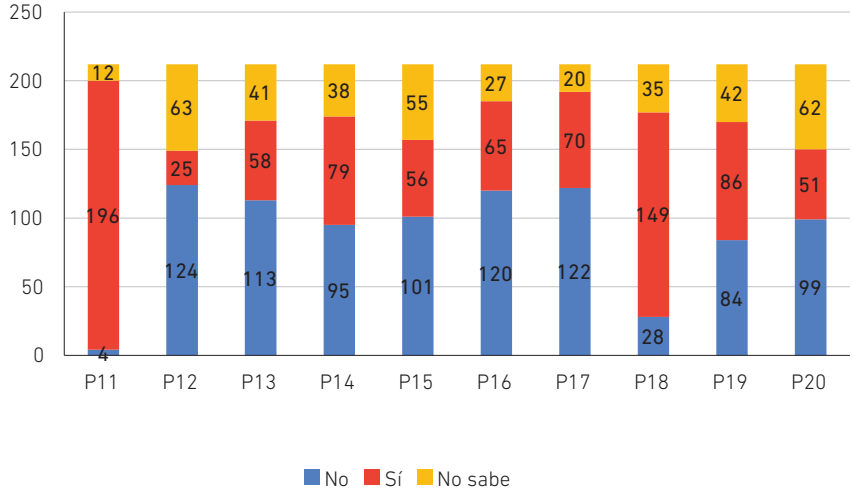

Siguiendo el análisis, se puede ver que la gran mayoría de la población entrevistada no dicta clases en ninguna otra universidad aparte de la UTN, lo hace fundamentalmente en el curso de Informática, sobre todo en carreras de grado y en la misma área en la cual se formó. Asimismo, se constata que la mayor parte tiene un manejo avanzado de *software*, pero las materias que dicta exigen solamente un manejo básico del idioma inglés.

### 4.3 Grupo 3

Seguidamente, se muestran en la Figura 3 los resultados obtenidos para las preguntas que van de la 11 a la 20 (P11-P20).

**Figura 3**

*Tipo de vínculo entre universidad e industria*



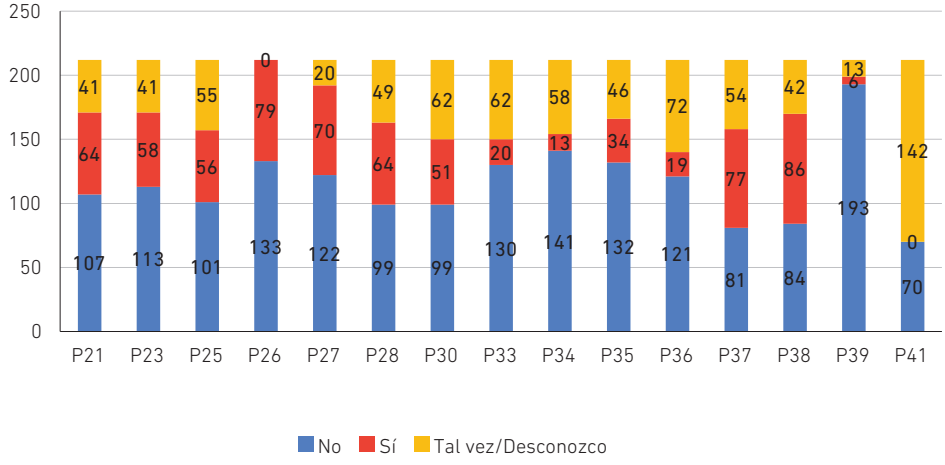

Se puede concluir que la gran mayoría de la población entrevistada conoce la existencia de un departamento de sistemas en su facultad, y piensa que se lleva adelante una política de colaboración abierta (Hernández et al., 2015) entre la universidad y la industria, aunque muchos entrevistados manifestaron no conocer el tema en detalle. Además, los docentes entrevistados opinan que no existen suficientes variantes en general (tecnicaturas, diplomaturas, etcétera) de las carreras de grado, que no existe suficiente oferta de carreras cortas con salida laboral inmediata, que la planificación curricular de las carreras no se diagrama en función de las necesidades de la industria ni considerando su crecimiento futuro. Por otra parte, señalaron también que los programas de estudio y los contenidos no se modifican en función de la actualización que la misma industria lleva adelante (Zachman & Redchuk, 2016). Pese a ello, también la mayoría constata que en las carreras se realizan prácticas integradoras finales para un más natural inicio laboral del graduado. Respecto de si se toman en cuenta las demandas de capacitación de la industria para confeccionar los programas de estudio, las respuestas estuvieron parejamente divididas. Por último, al preguntarse si se forma suficientemente a los estudiantes en habilidades blandas (Grosso, 2019) para una mejor inserción futura en el mundo laboral, la respuesta mayoritaria es negativa.

### 4.4 Grupo 4

Este grupo, que abarca las preguntas 21 a 41 (P21-P41) corresponde al tipo de vínculo entre universidad e industria y a sus connotaciones. A continuación, se muestra la Figura 4, en la que se detallan las respuestas a todas las preguntas tricotómicas (Sí, No, Tal vez/Desconozco) de este grupo de interrogantes.

**Figura 4**

*Respuestas a preguntas tricotómicas del grupo 4*

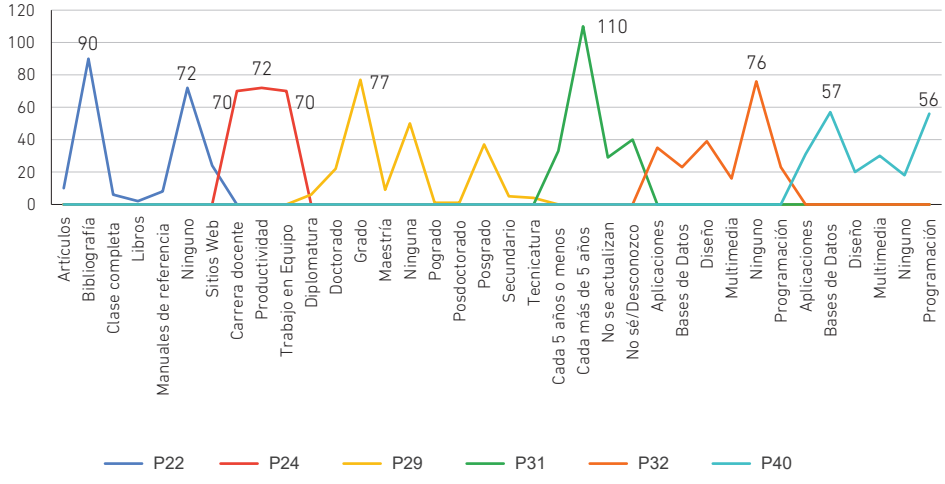



El análisis de estas preguntas muestra que, mayoritariamente, los docentes entrevistados afirman que la Universidad no cuenta con presupuesto propio para integrarse mejor con la industria; que no se imparten suficientes conocimientos sobre *software* en las clases ni se forma adecuadamente al estudiantado en habilidades blandas; que los docentes no trabajan con frecuencia fuera del ámbito académico y, por tanto, no tienen amplia experiencia laboral en otros ámbitos. Además, consideran que no se imparten los conocimientos prácticos necesarios para la futura inserción laboral de los estudiantes, ni se actualizan los planes de estudio universitarios en función de las demandas de la sociedad. De acuerdo con las respuestas, las empresas no ofrecen ayuda económica para que los estudiantes cursen formación de posgrado y tampoco existen convenios formales de colaboración para que los docentes realicen este tipo de estudios. Adicionalmente, las respuestas indican que no se realizan suficientes prácticas para la inserción de los estudiantes universitarios en el mercado laboral, que no se cuenta con personal docente certificado oficialmente (por los organismos profesionales correspondientes) para el dictado de las materias relacionadas con los conocimientos requeridos por dicha certificación y, finalmente, que las materias optativas que la universidad ofrece cada cuatrimestre no ayudan a reducir la brecha con la industria (Taucean et al., 2018).

Por otra parte, en la Figura 4 se observa la distribución de las respuestas obtenidas a las restantes preguntas del grupo. A saber: manejo de presupuesto propio (P21), conocimientos sobre *software* impartidos en las materias dictadas por los docentes (P23), suficiencia de la formación de los estudiantes en habilidades blandas (P25), situación laboral de cada docente fuera del ámbito académico (P26), experiencia laboral de los docentes fuera del ámbito académico (P27), impartición de conocimientos prácticos relacionados con la futura inserción laboral del estudiantado (P28), actualización de los planes de estudio de la universidad en función de satisfacer las demandas de la sociedad (P30), ayuda ofrecida por las empresas para cursar formación de posgrado (P33), existencia de convenios formales de colaboración para cursar estudios de posgrado (P34), suficiencia de las pasantías realizadas por el estudiantado en el mercado laboral (P35), existencia de personal docente certificado (P36), valor agregado de las carreras superiores (P37), renovación de la oferta de materias optativas para actualización profesional y su orientación a los requerimientos industriales (P38 y P39) y existencia de presupuesto externo o gubernamental para cumplimentar regulaciones industriales existentes (P41).

**Figura 5**

*Respuestas a preguntas de opción múltiple del grupo 4*




En la Figura 5, en la pregunta sobre cuál es el principal contenido en inglés impartido en las clases, mayormente las de grado, la mayoría de las respuestas hace referencia a la bibliografía. Sobre qué características son importantes a la hora de evaluar las acciones de un colaborador docente, la mayoría se inclina (muy ligeramente) por la productividad, aunque las otras opciones son votadas parejamente también. Se contesta mayoritariamente que los planes de estudio se actualizan con una frecuencia mayor a los cinco años. Además, la mayoría de entrevistados opina que no se imparten suficientes contenidos

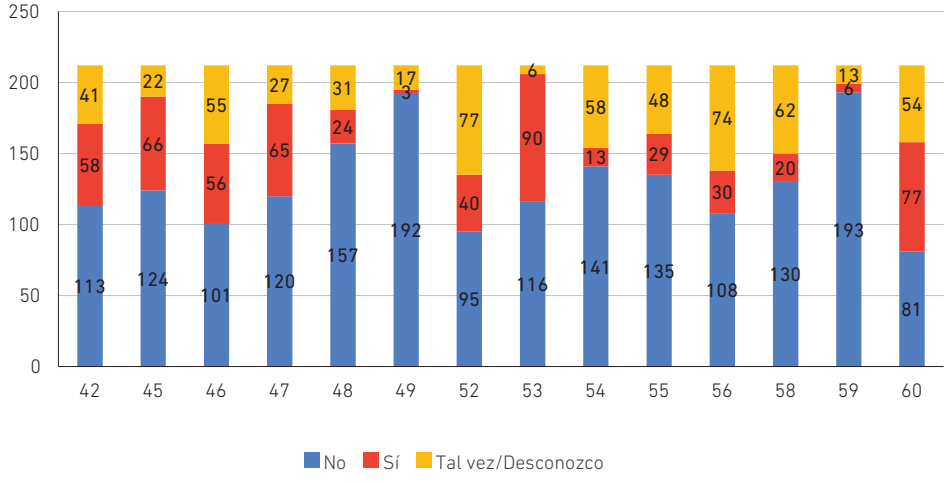
sobre *software* en las materias optativas y que la mayor oferta de estas se encuentra relacionada con bases de datos.

### 4.5 Grupo 5

El grupo 5 de preguntas (P42-P61) corresponde también al vínculo entre la universidad y la industria. El análisis puede observarse en la Figura 6, donde se representan todas las preguntas tricotómicas de este grupo.

**Figura 6**

*Respuestas a preguntas tricotómicas del grupo 5*



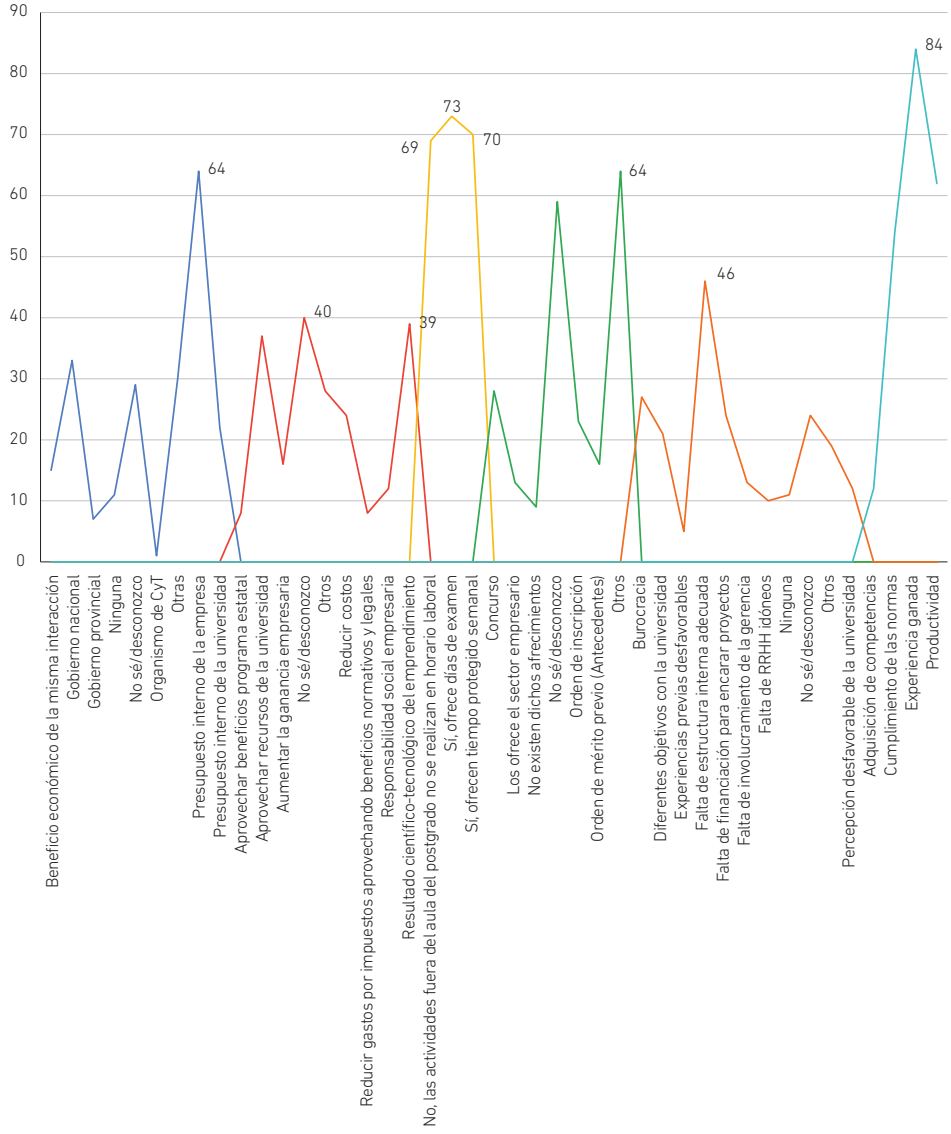

El análisis muestra que, en su mayoría, los docentes afirman que en la Universidad no se efectúan prácticas rentadas en integración con la industria, no se realizan suficientes pasantías, no existe un programa empresarial que ayude a los graduados universitarios a insertarse en el mercado laboral y que el mercado laboral no contrata graduados en puestos acordes a su formación. Además, la mayor parte de entrevistados considera que no se forma al estudiantado en gestión empresarial, no se ofrecen programas específicos de capacitación para obtener las certificaciones requeridas actualmente por la industria y que su labor docente no tiene ningún tipo de vínculo con el sector empresarial. También aparece en los resultados que en las empresas no se mide de manera alguna la brecha entre lo que la universidad ofrece y lo que la industria requiere; que no existe un marco regulador, ni interno ni externo, que sirva a la relación de la universidad con la industria; que no existe suficiente interrelación entre el espacio académico de la universidad y las instituciones gubernamentales; que no existen roles bien definidos en la estructura organizacional universitaria para su integración con la industria, y que no existe suficiente interrelación, en general, entre la universidad y la

sociedad en su conjunto (Arza & Vazquez, 2010). Finalmente, en cuanto a si se tienen presentes las normas industriales en la gestión diaria de la universidad, los porcentajes de respuestas afirmativas y negativas son parejos.

Por otra parte, en la Figura 7 se observa la distribución de las respuestas obtenidas de las preguntas de opción múltiple del grupo 5.

**Figura 7**

*Respuestas a preguntas de opción múltiple del grupo 5*



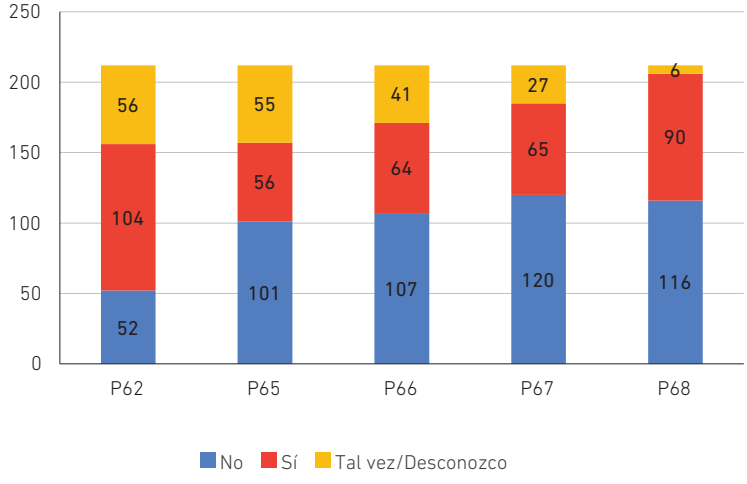
Se puede ver que, como principal fuente de financiación de los proyectos, se hace referencia en primer lugar al presupuesto interno de la empresa, seguido por otros ítems. Sobre cuál sería el principal objetivo de los proyectos de vinculación entre el sistema científico-tecnológico y la empresa, una gran parte se inclina (muy ligeramente, entre muchos otros ítems) por el aprovechamiento de los recursos de la universidad, como así también —en casi igual medida— por el resultado científico-tecnológico del emprendimiento. Sin embargo, la mayoría de los entrevistados expresa su ignorancia o desconocimiento al respecto. Se destaca, por otra parte, que las empresas ofrecen días libres para examen a los estudiantes (P50). La mayoría manifiesta no conocer cómo se cubren las necesidades de capacitación. La falta de estructura interna adecuada en la universidad es identificada como la principal dificultad para vincularse con las empresas. En lo que respecta a las características que deben considerarse para evaluar el desempeño de un alumno pasante en una empresa, la mayoría de los entrevistados se inclina por la adquisición de competencias, por sobre el cumplimiento de las normas industriales, la experiencia ganada y la productividad.

### 4.6 Grupo 6

El grupo 6 (P62-P68) concierne también al vínculo entre universidad e industria y la formación misma de SO. En la Figura 8 se presentan las respuestas a todas las preguntas tricotómicas de este grupo.

**Figura 8**

*Respuestas a preguntas tricotómicas del grupo 6*



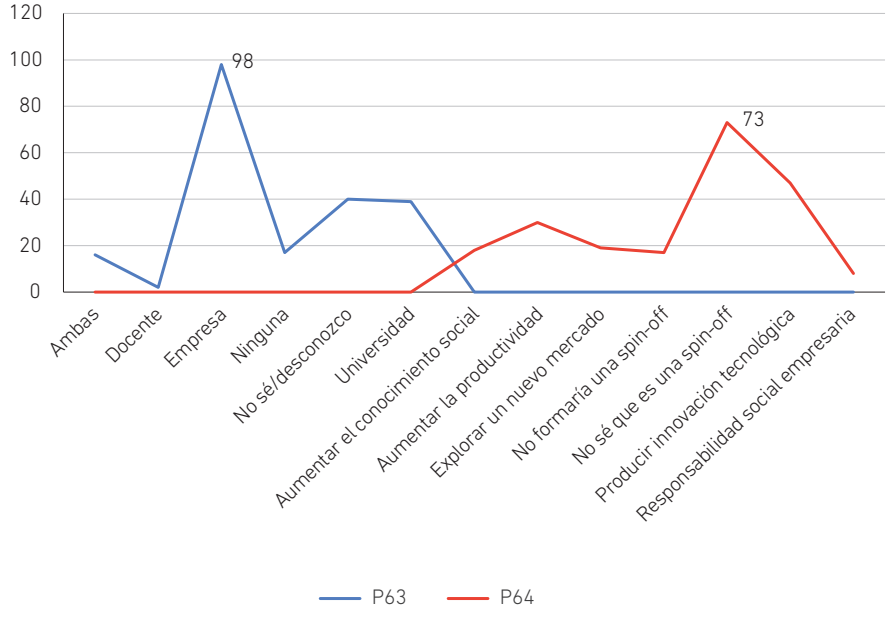

El análisis muestra que, en su mayoría, las personas consultadas afirman lo siguiente: en la universidad están claros los objetivos organizacionales respecto de la integración con la industria, pero el Departamento de Sistemas no tiene en cuenta (en su

quehacer diario) las demandas de ese sector. Por otro lado, no se apoya a los estudiantes para conseguir trabajo, no se ofrece a los postulantes una posible salida laboral al final de la carrera y, en general, no hay propuestas de inserción laboral para el estudiantado, salvo algunas, limitadas por los cupos (Massaro, 2016).

Por otra parte, en la Figura 9 se observa la distribución de las respuestas obtenidas a las preguntas de opción múltiple de este grupo; es decir, P63 y P64.

**Figura 9**

*Respuestas a preguntas de opción múltiple del grupo 6*



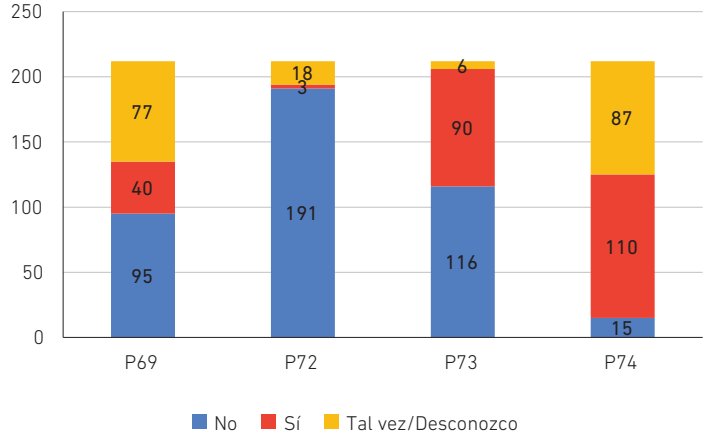

A la pregunta acerca de cómo se originan los vínculos entre empresas y docentes, se observa que la mayoría refiere principalmente a las mismas empresas. Sobre cuál sería su principal motivación para formar una SO, una gran parte se inclina (entre muchos otros ítems) por la producción de innovación tecnológica. Sin embargo, la mayoría de entrevistados indica no conocer lo que es una *spin-off*.

### 4.7 Grupo 7

Este grupo alude también a las características del vínculo entre universidad e industria, pero también al desarrollo de SO y de la actividad científica. El análisis puede sintetizarse primero en la Figura 10, que representa las respuestas de todas las preguntas tricotómicas (Sí, No, Tal vez/Desconozco) del grupo.

**Figura 10**

*Respuestas a preguntas tricotómicas del grupo 7*

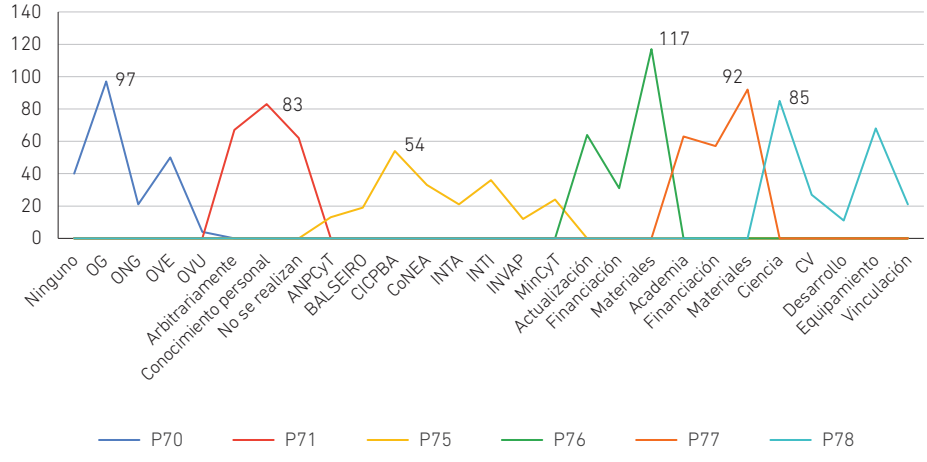



La mayoría de docentes que respondieron el cuestionario afirma que en la universidad no existe una estructura gerencial para trabajar la vinculación con la industria, no se consideran las normas regulatorias en la gestión diaria y no se cuenta con un plantel científico abocado a la integración con la industria (Jiménez & Castellanos, 2008). Por otra parte, se tiene amplio conocimiento de las actividades que realiza el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

La Figura 11 representa la distribución de las respuestas obtenidas a las restantes preguntas del grupo 7; es decir, P70, P71, P75 y P78.

**Figura 11**

*Respuestas a preguntas de opción múltiple del grupo 7*



La figura muestra que la mayoría hizo referencia a las organizaciones gubernamentales como agentes para realizar convenios con la industria. Sobre cuál sería el procedimiento básico para realizar estos convenios, una gran parte se inclina por los contactos personales. Respecto del conocimiento de organizaciones científicas (más allá del CONICET), la más mencionada es la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CICPBA). En relación a las motivaciones de un investigador para relacionarse con las empresas, se señala mayoritariamente el acceso a materiales, al igual que pasa con las motivaciones de un empresario para relacionarse con los organismos de investigación. Por último, acerca de cómo debería evaluarse la actividad de un investigador, se destaca su eventual contribución a la ciencia.

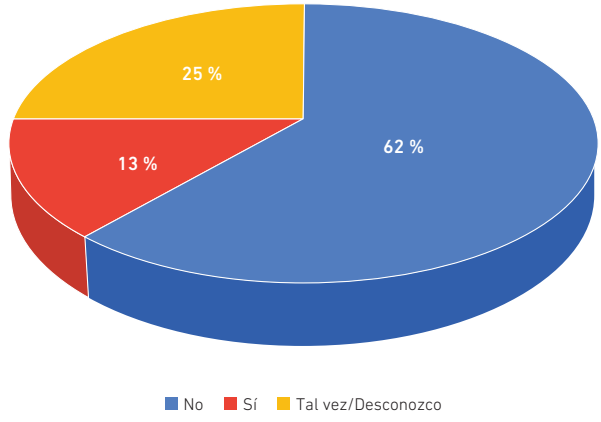
### 4.8 Grupo 8

En relación con el grupo 8, se presenta un gráfico consolidado, que se observa en la Figura 12. Allí se manifiesta la opinión general acerca del conocimiento (o no) del vínculo entre universidad y empresa, y su eventual conformidad con tal vinculación.

Los datos obtenidos en las entrevistas que dan lugar a la Figura 12 sugieren que, en promedio consolidado para todo el grupo, el 62 % de los entrevistados desconoce las iniciativas de integración con la industria o gobierno (y además no está conforme con ellas), mientras que solo el 13 % afirma conocerlas en detalle y, además, estar de acuerdo con ellas. A su vez, la cuarta parte de la población entrevistada manifiesta tener un conocimiento parcial o nulo y, por ende, carece de opinión formada al respecto.

**Figura 12**

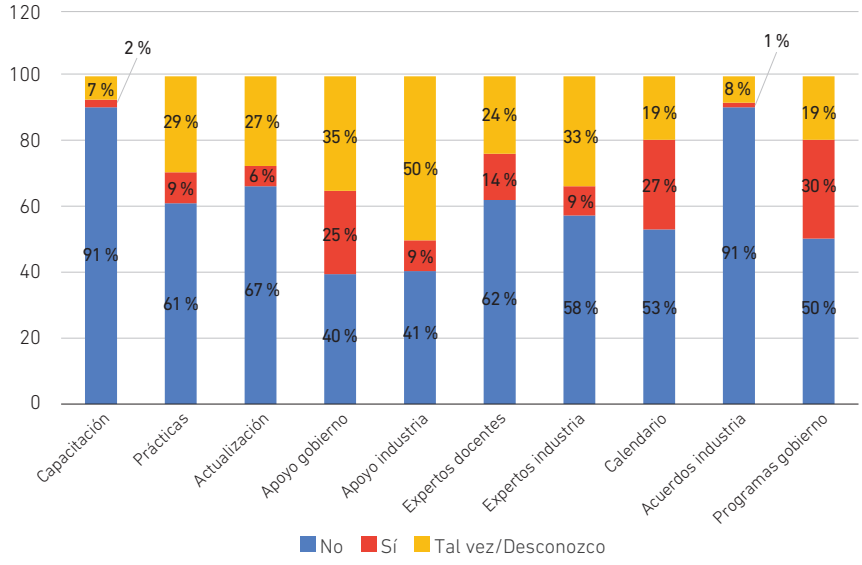
*Opinión general manifestada en las entrevistas acerca de las características del vínculo entre universidad, industria y gobierno*



Al desagregar las respuestas en función de cada una de las preguntas del grupo, se obtiene la Figura 13, que se presenta a continuación.

**Figura 13**

*Opiniones acerca de las características del vínculo entre universidad, industria y gobierno*




Se ve en la Figura 13 que los grados de aprobación más bajos están relacionados con los programas de capacitación docente sobre la situación de la industria y su relación con la universidad, así como con los acuerdos de prácticas sistemáticas y orgánicas con la industria (91 % de respuestas negativas en ambos grupos, mientras que el promedio de respuestas negativas del grupo ronda, como se acaba de señalar en la Figura 12, el 62 %). Mientras tanto, las respuestas afirmativas a las mismas preguntas son, respectivamente, de 2 % y 1 % (frente al ya señalado promedio del 13 %, según se ilustra en la Figura 12), lo cual indica que los docentes entrevistados consideran ampliamente que existen falencias en ambas cuestiones.

En cambio, el porcentaje de respuestas negativas sobre el apoyo del gobierno o de la industria a la universidad ronda el 40 %, que es el más bajo de todo el grupo. Esto indicaría, por ende, el mayor grado de reconocimiento —aunque igualmente escaso— a la interrelación entre estas instancias. Por otro lado, el mayor porcentaje de respuestas afirmativas se recoge en relación con el apoyo programático del gobierno (entre 25 % y 30 %), mientras que para el apoyo efectivo de la industria se responde afirmativamente en una proporción mucho menor (entre 1 % y 9 %). Esto estaría mostrando que los docentes se encontrarían mucho más conformes con el apoyo oficial que con el privado,

toda vez que además un alto porcentaje afirmativo (27 %) considera que el calendario académico se sincroniza todo lo posible con el de la actividad industrial, para impulsar así la interacción.

En cuanto al grado de desinformación que los docentes entrevistados tienen acerca de cada uno de los ítems sobre los que se preguntó (siendo el promedio del 25 %, según lo señalado en la Figura 12), el mayor porcentaje de desconocimiento se obtiene nuevamente con el apoyo de la industria (50 %), mientras que el menor se da en relación con los acuerdos industriales (8 %) y a los programas de capacitación ofrecidos por la industria (7 %), lo cual demostraría que, si bien los entrevistados conocen la existencia de programas y acuerdos, no se sentirían lo suficientemente apoyados como para acceder a ellos.

### 4.9 Discusión de los resultados obtenidos

Del análisis previamente desarrollado y a modo de resumen, se postula un conjunto de medidas que la universidad puede implementar para reducir la brecha detectada en su relación con la industria. A saber:

- Involucrar más a los docentes. La universidad puede difundir los detalles de la UIC que se estén llevando a cabo.
- Ofrecer más variantes de carreras cortas con salida laboral inmediata, diagramadas en función de las necesidades actuales de la industria y de su planificado crecimiento futuro.
- Adecuar la planificación curricular y los programas de estudio y sus contenidos a la constante actualización de la industria, lo que aumenta su propia frecuencia de actualización, de modo que la UIC pueda contribuir efectivamente a satisfacer las demandas que la sociedad le hace de manera implícita, pero cotidiana.
- Mejorar la formación de los estudiantes en habilidades blandas, de modo tal que se consiga en el futuro una mejor inserción laboral en puestos gerenciales de la industria.
- Asignar fondos *ad hoc* dentro del presupuesto universitario para lograr una mejor integración con la industria. Para ello, la universidad debe proponer y solicitar —si es necesario y a tal efecto— una ampliación presupuestaria al Estado nacional.
- Impartir conocimientos sobre *software* más directamente relacionados con la situación y necesidades actuales de la industria. Ello implica que sería ventajoso contar con docentes que trabajen directamente en el ámbito industrial, a la par de su labor universitaria.

- Impulsar convenios con la industria para obtener ayuda económica para cursar formación de posgrado, tanto para alumnos de carreras de grado como para docentes.
- Impulsar la realización de una mayor cantidad y variedad de prácticas y pasantías (rentadas o no) *in situ*, a fin de lograr una mejor inserción de los estudiantes en el mercado laboral.
- Promover convenios para que el personal docente pueda ser certificado de forma oficial por los organismos profesionales correspondientes para el dictado de las materias relacionadas con los conocimientos requeridos por dicha certificación.
- Ofrecer una grilla de materias optativas renovada en cada periodo lectivo, orientada a impartir contenidos de *software* relacionados con las demandas actualizadas de la industria.
- Propiciar la realización de un programa empresarial de ayuda a los graduados universitarios para facilitar su inserción en el mercado laboral.
- Interactuar activamente en el quehacer cotidiano con la industria, para lograr que el mercado laboral contrate graduados en puestos acordes a su formación, que tenga en cuenta las normas regulatorias sobre el tema.
- Mejorar la formación del estudiantado en gestión empresarial en general.
- Ofrecer programas específicos de capacitación para obtener las certificaciones requeridas actualmente por la industria.
- Propiciar un proceso para medir en forma sistemática la brecha entre lo que la universidad ofrece y lo que la industria requiere.
- Tener en cuenta como valor agregado en la labor docente los vínculos que los propios profesores puedan tener con el sector empresarial.
- Fomentar la creación de un marco regulatorio institucional de la universidad, tanto en el aspecto externo como en el interno, para relacionarse positivamente con la industria.
- Mejorar la interrelación entre la universidad y las instituciones gubernamentales, considerando que los docentes manifestaron que se encontrarían mucho más conformes con el apoyo oficial que con el privado.
- Clarificar la definición de los roles en la estructura organizacional de la universidad con miras a la integración con la industria.
- Mejorar y adecuar la estructura interna de la universidad, tanto gerencial como administrativa, para vincularse con la industria.

– Mejorar el proceso de apoyo a los estudiantes para una posible salida laboral al final de la carrera universitaria de grado.

– Contar con un plantel científico abocado a la integración de la universidad con la industria.

– Ampliar el apoyo a los docentes para acceder a los acuerdos industriales y a los programas de capacitación ofrecidos por la industria.

## 5. CONCLUSIONES

Con los resultados detallados en el punto anterior, definidos a partir de ocho grupos de referencia que representan una idea general de primera instancia, y teniendo en cuenta que el presente estudio constituye una propuesta para un análisis a profundizar, se obtienen las siguientes conclusiones. Cabe mencionar que estas conclusiones se desprenden de los datos de las entrevistas, los cuales muestran, con ayuda de los gráficos, los principales hallazgos relacionados directamente con las preguntas de investigación planteadas.

- El tema analizado resulta de interés para todos los entrevistados. Esto se pudo apreciar también durante el desarrollo mismo de las entrevistas.
- La opinión generalizada es que se considera insuficiente el estado actual de la UIC en la República Argentina.
- La amplia bibliografía y la gran cantidad de artículos científicos encontrados actualmente al respecto (en particular, en Argentina) destacan asimismo la importancia del estudio y su justificación en el ambiente educativo universitario.
- Los objetivos planteados en el estudio quedan cumplidos con plenitud, no habiéndose encontrado resistencias ni dificultades.
- Se ha logrado dar, a través de las respuestas recibidas en las entrevistas, una mirada más profunda y específica al análisis de la problemática.
- Los datos obtenidos resultan provechosos y útiles para el análisis a efectuar.
- La metodología de trabajo aplicada se ha desarrollado y adaptado satisfactoriamente.

## 6. FUTURAS LÍNEAS DE INVESTIGACIÓN PROPUESTAS

A partir de las antedichas conclusiones, se propone como primera línea de investigación futura la repetición de este análisis. Para ello, sería importante realizar entrevistas sobre el mismo asunto a docentes de otras universidades, a empresarios y a investigadores, así como agregar preguntas que abarquen otros aspectos del análisis. Sería interesante

sondear, por ejemplo, si se opina que la presencialidad al cursar las materias afecta (o no) la elección de una carrera universitaria por sobre los cursos virtuales que permiten un vínculo directo y rápido con la industria. O bien si se piensa (o no) que el trabajar en la industria debería ser un requisito *sine qua non* para que un docente pueda impartir enseñanzas prácticas en las materias del área de Informática. Los autores ya se encuentran trabajando en ello, y en un futuro cercano se publicarán los resultados obtenidos.

Adicionalmente, se plantea efectuar un trabajo similar con funcionarios de Gobierno y con distintos actores sociales relevantes respecto del tema en estudio. Asimismo, se plantea continuar el análisis y ahondarlo, llevando adelante un mayor desagregado de los datos, para obtener así conclusiones más profundas, de modo que sea posible postular un modelo de universidad emprendedora que acorte la brecha existente. Sobre esto último también se encuentran trabajando los autores del presente estudio.

## REFERENCIAS

Alderete, M. V., Porris, M. S., & Verna Etcheber, R. (2020). Hacia un modelo de innovación de cuádruple hélice: experiencias con PyMEs de Bahía Blanca, Argentina. *Ciencias Económicas*, *1*(17), 67-88. https://doi.org/10.14409/rce.v1i0.9994

Arza, V., & Vazquez, C. (2010). Interactions between public research organisations and industry in Argentina. *Science and Public Policy*, *37*(7), 499-511. http://hdl.handle.net/10625/50641

Bahdanava, A., Gaydova, M., Izmailovich, S., Voronko, E., & Kostuchenko, E. (2024). International cooperation between universities and business as a condition for raising human development level. *BIO Web of Conferences*, *83*(06001), 1-14. https://doi.org/10.1051/bioconf/20248306001

Bergenholtz, C., & Bjerregaard, T. (2014). How institutional conditions impact university–industry search strategies and networks. *Technology Analysis & Strategic Management*, *26*(3), 253-266. https://doi.org/10.1080/09537325.2013.850473

Di Meglio, F. (2024). Capacidades y perfiles de vinculación científico-tecnológica en las universidades de la Provincia de Buenos Aires, Argentina. *Integración y Conocimiento*, *13*(1), 234-255. https://doi.org/10.61203/2347-0658.v13.n1.44227

Etzkowitz, H., & Leydesdorff, L. (1997). *Universities and the global knowledge economy: a triple helix of university-industry relations.* Pinter.

Furnari, S. (2014). Interstitial spaces: micro interaction settings and the genesis of new practices between institutional fields. *Academy of Management Review*, *39*(4), 439-462. https://doi.org/10.5465/amr.2012.0045

Gibson, D. V., & Foss, L. (2017). Developing the entrepreneurial university: architecture and institutional theory. *World Technopolis Review*, *6*(1), 3.1-3.15. https://doi.org/10.7165/WTR17A0809.16

Grosso, M. J. (2019). Especialización productiva y las prácticas de outsourcing y offshoring en el sector de software y servicios informáticos. *Pymes, Innovación y Desarrollo*, *7*(3), 37-62. https://revistas.unc.edu.ar/index.php/pid/article/view/28897

Hernández, M. C., Podestá, M. P., & Bedoya, B. E. (2015). *Conditions for the promotion and development of creative industries within higher education institutions* [Ponencia]. Proceedings of the 17th International Conference on Engineering and Product Design Education, Loughborough, United Kingdom. https://www.designsociety.org/publication/38436/CONDITIONS+FOR+THE+PROMOTION+AND+DEVELOPMENT+OF+CREATIVE+INDUSTRIES+WITHIN+HIGHER+EDUCATION+INSTITUTIONS

Jiménez, C. N., & Castellanos, O. F. (2008). *Desafíos en gestión tecnológica para las universidades como generadoras de conocimiento* [Presentación de paper]. I Congreso Internacional de Gestión Tecnológica e Innovación, Bogotá, Colombia.

Krepki, D. (2024). El garaje en la empresa: jóvenes trabajadorxs intra-emprendedores en la industria tecnológica argentina. El caso Globant. *Trabajo y Sociedad*, *25*(42), 5-18. http://www.scielo.org.ar/pdf/tys/v25n42/1514-6871-tys-25-42-5.pdf

Lauric, A., Scoponi, L., Torres Carbonell, C., & De Leo, G. (2024). Vinculación DCAUNS-AERINTA Bahía Blanca-sector productivo: indicadores para la gestión de la sustentabilidad de Pymes agropecuarias en ambientes frágiles de Argentina. En C. Garrido-Noguera y D. García Pérez de Lema (Coords.), *Universidades, Economía Circular y los ODS en el espacio birregional ALCUE* (pp. 131-148). ALCUE, FAEDPYME, Unión de universidades de América Latina y El Caribe. https://dialnet.unirioja.es/servlet/libro?codigo=976325

Massaro, F. (2016). *Un nuevo modelo teórico sobre los procesos de Spinout. Aplicación y validación estadística para el sistema científico-tecnológico argentino (período 2005-2015)* [Tesis de doctorado, Universidad Nacional de Lomas de Zamora]. Repositorio Institucional Digital de Acceso Abierto de la UNLZ.

Ortiz, F. D. (2019). *Establecimiento del estado del arte de la gestión de riesgos en el proceso de implantación de sistemas informáticos* [Tesis de maestría, *Universidad Tecnológica Nacional Facultad Regional Buenos Aires*].

Peksatici, Ö., & Ergun, H. S. (2019). The gap between academy and industry. A qualitative study in Turkish aviation context. *Journal of Air Transport Management*, *79*, 101687. https://doi.org/10.1016/j.jairtraman.2019.101687

Sales, M. (2021). *Diagrama de pareto*. EALDE Business School. https://www.gestiopolis.com/diagrama-de-pareto/

Scott, W. R. (1987). *The adolescence of institutional theory*. Administrative science quarterly.

Universidad Tecnológica Nacional. (2022, 15 de junio). *Ordenanza 1877/22 sobre diseño curricular de la carrera en Ingeniería en Sistemas de Información. Plan 2023*. https://www.frba.utn.edu.ar/wp-content/uploads/2022/12/Ordenanza-1877-Plan-ISI2023.pdf

Taucean, I. M., Strauti, A. G., & Tion, M. (2018). Roadmap to entrepreneurial university–Case study. *Procedia-Social and Behavioral Sciences*, *238*, 582-589. https://doi.org/10.1016/j.sbspro.2018.04.038

Zachman, P. P., & Redchuk, A. (2016). Singularities of the university spin-off in northern Argentina. En A. Rocha, A. Correia, H. Adeli, L. P. Reis & M. Mendonça (Eds.), *New advances in information systems and technologies.* Springer.

## APÉNDICES

**Apéndice A. Cuestionario de 88 preguntas utilizado para las entrevistas**
https://drive.google.com/file/d/11203WN-gsFiqxGOH_tjX89iKKZp6qV2I/view

**Apéndice B. Enlace a la grilla de respuestas dadas por los docentes**
https://docs.google.com/spreadsheets/d/1ak_ehw0MKj-YXAcsXxah677IPPAvKabj/edit?gid=1818122643#gid=1818122643

# Dynamic Malware Analysis Using Machine Learning-Based Detection Algorithms

Erly Galia Villarroel Enriquez
20182063@aloe.ulima.edu.pe
ORCID: 0000-0001-8566-0494
Facultad de Ingeniería, Universidad de Lima

Juan Gutiérrez-Cárdenas
Jmgutier@ulima.edu.pe
ORCID: 0000-0003-2566-4690
Facultad de Ingeniería, Universidad de Lima

Received: May 10th, 2024 / Accepted: June 8th, 2024
doi: https://doi.org/10.26439/interfases2024.n19.7097

ABSTRACT. With the increasing popularity of cell phone use, the risk of malware infections on such devices has increased, resulting in financial losses for both individuals and organizations. Current research focuses on the application of machine learning for the detection and classification of these malware programs. Accordingly, the present work uses the frequency of system calls to detect and classify malware using the XGBoost, LightGBM and random forest algorithms. The highest results were obtained with the LightGBM algorithm, achieving 94,1 % precision and 93,9 % accuracy, recall, and F1-score, demonstrating the effectiveness of both machine learning and dynamic malware analysis in mitigating security threats on mobile devices.

KEYWORDS: malware / machine learning / detection

## Análisis dinámico de *malware* mediante algoritmos de detección basados en *machine learning*

RESUMEN. Con la creciente popularidad del uso de teléfonos celulares, el riesgo de infecciones por malware en dichos dispositivos ha aumentado, lo que genera pérdidas financieras tanto para individuos como para organizaciones. Las investigaciones actuales se centran en la aplicación del aprendizaje automático para la detección y clasificación de estos programas malignos. Debido a esto el presente trabajo utiliza

la frecuencia de llamadas al sistema para detectar y clasificar malware utilizando los algoritmos XGBoost, LightGBM y random forest. Los resultados más altos se obtuvieron con el algoritmo de LightGBM, logrando un 94.1% de precisión y 93.9% tanto para exactitud, recall y f1-score, lo que demuestra la efectividad tanto del uso del aprendizaje automático como del uso de comportamientos dinámicos del malware para la mitigación de amenazas de seguridad en dispositivos móviles.

PALABRAS CLAVE: malware / machine learning / detección

## 1. INTRODUCTION

According to the FortiGuard Labs threat report (Fortinet, 2022) for Latin America and the Caribbean, cyberattack attempts surged by 600 % in 2021 compared to 2020, constituting 10 % of the total global attempts. FortiGuard Labs, Fortinet's threat intelligence laboratory, also noted that Peru was the third most targeted country, with 11,5 billion attack attempts, following Mexico and Brazil.

Duo et al. (2022) argue that cyberattacks pose security challenges in cyber systems, potentially affecting system performance or causing additional damages. These attacks can take various forms, including denial of service, phishing, or malware.

Several authors (Saravia et al., 2019, as cited in Ashik et al., 2021) define malware as malicious software embedded in lawful programs to perform criminal activities. With the rapid spread of the Internet and the proliferation of connected devices, malware attacks have increased significantly, jeopardizing user privacy.

Regarding the causes of malware infections, Ashik et al. (2021) point out that these are primarily due to the download of free software such as games, web browsers, or free antivirus programs. Three common methods by which malware can infiltrate a device are highlighted: download attack, where malware is hosted on a web server to infect devices visiting the page; update attack, which modifies a benign application to include malware characteristics; and repackaging attack, where malware is embedded in a benign application (Felt et al., 2014, as cited in Surendran & Thomas, 2022).

Current research focuses on enhancing malware detection through machine learning algorithms. For instance, Mahindru and Sangal (2020) developed a framework to safeguard Android devices using a dataset containing benign and malicious samples of Android Package Kit (APK) files collected from sources including Google Play Store, Android, Panda.App, among others. They employed a feature selection approach, extracting specific application features (permissions, system calls, number of app downloads, and app ratings) for model training using various machine learning algorithms such as logistic regression, support vector machines (SVM), or random forest (RF), to compare them with existing models trained on all APK file characteristics. They found that models employing a feature selection approach outperformed those using the entire set of extracted features.

The selection of dataset features to be used as input for machine learning models is also crucial, as noted by Wu et al. (2021). Their study focused on devising why an application is classified as malware by machine learning algorithms, using application programming interface (API) calls and permissions from APK files. The dataset collected by the authors consisted of 20 120 benign applications and 15 570 malicious apps. They found that these two characteristics alone are insufficient to fully explain malware behavior.

Surendran and Thomas (2022) proposed a novel malware detection system in Android using graph centrality measures composed of system calls from APK applications. The RF algorithm exhibited the highest performance, with 0,98 accuracy for detecting obfuscated malware. Additionally, they suggested further research to reduce false positive rate (when an application is incorrectly classified as malware).

For this study, a data science approach will be adopted, utilizing the RF, XGBoost, and LightGBM algorithms, which have been employed in recent research. Louk and Tama (2022) applied these algorithms to efficiently detect malware, achieving precision and accuracy above 99,2 % for all three algorithms.

Based on the algorithms currently used in the field of data science, the goal of our present study is to evaluate the effectiveness of machine learning algorithms such as RF, LightGBM, or XGBoost on the CICMalDroid 2020 dataset, which utilizes dynamically observed malware behaviors (actions performed by malware while in execution) to identify the most suitable model for preventing malware cyberattacks.

This article is organized as follows: Section 2 presents a survey of experimental studies on major malware detection techniques and machine learning algorithms to be trained for malware detection scenarios. Section 3 discusses the methodology and describes the dataset employed. Section 4 presents the experimentation, followed by the results in Section 5. Section 6 provides discussions and finally Section 7 presents the conclusions and future works.

## 2. BACKGROUND

There are two main techniques for analyzing malware: static and dynamic analysis. The former involves analyzing malware without executing it (Alosefer Y., 2012, as cited in Aslan & Samet, 2020), while the latter involves analyzing malware as it runs in a real or virtual environment (Bhat & Dutta, 2019, as cited in Liu et al., 2020). For the present study, malware execution will not be performed directly; instead, a dataset where dynamic analysis was previously conducted will be employed.

### 2.1 Dynamic Analysis

Surendran et al. (2020) claim that to detect malicious activity, dynamic analysis predominantly considers data originating from the running application, including system call traces and sensitive API calls. It is worth noting that system calls contain more relevant information about malware behavior compared to API calls, as misclassifications can occur when a benign application invokes API calls frequently used in malicious applications. The authors conclude that certain types of legitimate applications, such as face detection or weather prediction apps, request more system privileges during execution. In such cases, these benign apps may generate system calls that resemble those found

in malware apps, potentially causing these goodware apps to be mistakenly flagged as malware. To reduce these false positives, the authors suggest future research should consider both the timing and frequency of malicious system calls in an application.

Feng et al. (2018) highlight that system calls indicate how applications request services from the operating system to perform important functions such as power management, device security, and hardware resource access, among others. However, they can also be used for malicious purposes; the authors note that malware tends to use more system calls than benign applications. Examples of such system calls include syscall operations like "fork," "fchmod," and "wait4," which indicate changes in file ownership containing sensitive information or the creation of child processes to perform hidden malicious behavior. The authors conclude that despite the effectiveness of their dynamic analysis framework called EnDroid, which extracts system-level behavior traces and common application-level malicious behaviors, it is necessary to improve the coverage of the dynamic analysis for future research.

Surendran and Thomas (2022) focused on tracking system calls to detect malicious activity in applications, noting that malware applications automatically invoke sensitive APIs (such as making calls) to execute privileged actions like collecting information from contacts. For their research, they first gathered system call traces and organized them into an ordered graph where system calls are vertices and their edges represent adjacency relations. This ordered graph enables the extraction of central values from the system call graph, such as "rename" and "open" system calls, using centrality measurements like eigenvector, betweenness, and closeness. The authors concluded that their proposed system outperforms existing malware detection mechanisms based on system calls with an accuracy of 0,99. However, they noted the necessity of using new tools for collecting system calls from malware applications. Due to the mechanism employed by the authors, some malware apps did not exhibit malicious behavior during the collection of syscall traces because of the limited code coverage problem in automated test case generation tools like monkeyrunner.

As a result of the review, some gaps were identified like the necessity of using new tools for collecting system calls, improving the coverage of dynamic analysis, and considering the frequency of occurrence of system calls when analyzing malware applications. Therefore, in this article, we will use a dataset that includes the number of occurrences of dynamic malware behaviors and has been processed with a tool that adequately reconstructs them.

### 2.2 Supervised Machine Learning Techniques for Malware Detection

Several authors have employed machine learning algorithms to detect malware, utilizing feature extraction from various datasets. The most relevant works reviewed are mentioned below.

Aebersold et al. (2016) used RF and SVM algorithms on three different datasets: the first consisting of the complete list of JavaScript samples available on jsDelivr (a content delivery network for open-source projects), the second including the Top 500 Alexa websites (a page collecting information from websites), and the third consisting of a set of malicious JavaScript samples gathered from the Swiss Reporting and Analysis Centre for Information Assurance (MELANI). All precision scores were above 99 %. The authors concluded that a more representative dataset is needed to perform malware detection because the dataset of malicious scripts they used was much smaller than the dataset of benign scripts, making it unclear whether the classifier is capturing the actual syntactic characteristics correlated with the malicious behavior.

Chen et al. (2018) also used RF and SVM, as well as k-nearest neighbors (KNN), and applied them to a dataset they collected. The authors focused on two metrics, false negatives, and accuracy, concluding that the best algorithm was RF with an accuracy of 96,35 % and a false negative rate of 2,50 % (the lowest among the three algorithms applied).

Another example of using RF and SVM algorithms is provided by Kim et al. (2018), who collected macro applications in Excel and Word, then extracted 15 characteristics from them, such as the number of characters in the code (excluding comments) and average word length. The models employed ranged from SVM, multi-layer perceptron (MLP), or RF, resulting in relatively good performance. RF achieved a precision of 98,2 %, SVM achieved 88,1 %, and MLP achieved 93,8 % in the same metric.

Choudhary and Sharma (2020) also applied these machine learning algorithms—using datasets from other authors to evaluate malware detection with KNN, SVM, decision tree, and MLP—obtaining accuracies higher than 87 %, 89 %, 92 %, and 90 %, respectively. They concluded that machine learning algorithms have greater potential for malware detection compared to traditional techniques employed by current antivirus software (signature analysis).

Current research reflects the use of new machine learning algorithms for both detection and classification compared to past research. Chen et al. (2021) used the LightGBM algorithm on the Drebin dataset, using the frequency of API calls made by APK files within the dataset and achieved an accuracy of 99,54 %. Other works that used LightGBM include those by Gao et al. (2022), Onoja et al. (2022), and Chen et al. (2023).

New algorithms were also employed by Urooj et al. (2022), who collected 56 000 features from 100 000 Android applications using static analysis. For feature extraction, they used Androguard and performed feature selection to reduce the number of features, then input them into machine learning models such as KNN, naive Bayes (NB), radial basis function (RBF), decision tree, SVM, and AdaBoost

with decision trees. They achieved an accuracy of 96,26 % with the AdaBoost model and a false positive rate of 0,3 %. The authors concluded that ensemble and strong learner algorithms perform comparatively better when dealing with classifications and high-dimensional data. They also highlighted that their research approach was restricted in terms of static analysis, thus it is important to use a dataset with dynamic features for future research.

Şahın et al. (2022) used algorithms such as KNN, NB, RBF, decision tree, SVM, and linear methods like Linear Regression, obtaining the best result from the combination of SVM and decision trees, with an F-measure of 96,95 % for the AMD dataset. They concluded that the application of popular classification algorithms positively benefits malware classification. However, they noted that their static analysis approach used APK permissions as features, emphasizing the need to expand the research using dynamic behaviors of malware.

Palsa et al. (2022) used the XGBoost algorithm to detect malware in a dataset they collected through VirusShare, achieving 96,54 % accuracy using dynamic analysis features. Other authors who used the XGBoost algorithm include Dhamija and Dhamija (2021) and Kumar and Geetha (2020), who efficiently detected and classified malware with this novel algorithm. They concluded that the use of machine learning algorithms potentially benefits the detection of malicious software.

Finally, Louk and Tama (2022) applied RF, XGBoost, CatBoost, gradient boosting machine (GBM), and LightGBM algorithms to three different datasets containing features of malware that could be run portably (portable executable [PE] files) on Windows, obtaining precision and accuracy results above 99,2 % for both metrics. They concluded that the algorithms that performed best are those based on decision trees, with performance differences between algorithms being not statistically significant.

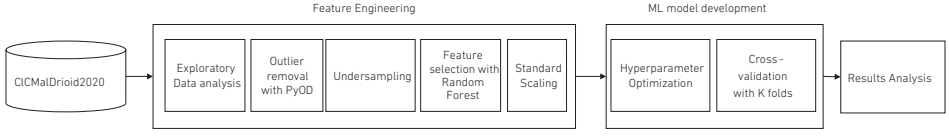
As a result of the literature review on machine learning models used for malware detection, it is evident that the most recent studies indicate the need to expand research using dynamic behaviors. It is also noted that the best-performance algorithms are those using decision trees (including RF), XGBoost, and LightGBM (a variation of gradient boosting to make it more efficient). This research will therefore experiment with these three models.

## 3. METHODOLOGY

This research proposes a methodology based on comparing machine learning algorithms applied to the detection and classification of malware families, specifically RF, LightGBM, and XGBoost. These algorithms will be applied to the CICMalDroid 2020 dataset, as described in the Experimentation section. Figure 1 presents the block diagram outlining the activities to be carried out in the experimentation.

**Figure 1**

*General Overview of the Proposed Approach*




Below is a brief explanation of the methodology:

1. Acquisition of Dataset With System Call Frequencies

   The chosen dataset for this research was CICMalDroid 2020—developed by Mahdavifar et al. (2020)—which contains the elements outlined in Table 1. It is important to emphasize that this research will only use the 'feature_vectors_syscallsbinders_frequency_5_Cat.csv' file. This dataset was generated through dynamic analysis using the CopperDroid tool, a system based on virtual machine introspection (VMI) that automatically reconstructs specific low-level Android behaviors and specific operating system activities from Android samples. According to the authors, out of 17 341 Android samples, only 13 077 executed successfully while the remainder failed due to issues such as timeouts, invalid APK files, and memory allocation difficulties. Additionally, 12 % of the JSON files (CopperDroid output results in this format) from the successful executions were not uploaded to the Canadian Institute for Cybersecurity website—where the CICMalDroid 2020 dataset is stored—due to "unfinished strings" (records that should be in double quotes but lack a closing quote).

2. Feature Engineering

   An exploratory data analysis was performed to ensure that the dataset does not contain null values. Given that the dataset is imbalanced, undersampling will be employed to balance the number of records. The most relevant features will be selected using the RF algorithm to determine the relative importance of each feature in predicting the target. The final dataset will be standardized to equalize the scales of the numbers before being used in the model building process.

3. Application of Machine Learning Models (XGBoost, LightGBM, RF)

   Cross-validation with grid search will be employed to select the best hyperparameters for the machine learning models. Subsequently, the models will be executed using k-folds cross-validation, and metrics such as accuracy, precision, and F1-score will be obtained for each model, in addition to the confusion matrix.

4. Results Analysis

For the analysis of the results, precision and recall metrics will be employed as they are effective to compare multi-class datasets. The results obtained will be compared with previous research that used the same dataset or datasets that included only system call frequencies. Possible reasons for differences in results will be discussed.

## 4. EXPERIMENTATION

### 4.1 Description of the Dataset Employed in the Experimentation Section

The components of the CICMalDroid 2020 dataset are described in Table 1.

**Table 1**

*Description of the Components of the CICMalDroid 2020 Dataset*

| Component | Description |
|---|---|
| APK Files | 17 341 Android samples categorized into five groups: riskware, banking malware, benign samples, SMS malware, and adware. |
| Capturing Logs | 13 077 samples were analyzed and the results were categorized into five groups: riskware, banking malware, SMS malware, adware, and benign samples. |
| Comma-Separated Values (CSV) Files | 'feature_vectors_syscallsbinders_frequency_5_Cat.csv':<br>1. Contains 470 characteristics, including binders, composite behaviors, and system call frequencies, retrieved from 11 598 APK files.<br>2. Contains 139 features, including system call frequencies, retrieved from 11 598 APK files.<br>3. Contains 50 621 features retrieved from 11 598 APK files, including static data such as sensitive APIs, files, method tags, intent actions, permissions, packages, and receivers. |

*Note.* These components and their descriptions were derived from research by Mahdavifar et al. (2020). The CSV files and dataset components were downloaded from the Canadian Institute for Cybersecurity website[1]. The '.csv' file that will be used for the creation of machine learning models in the Experimentation section is 'feature_vectors_syscallsbinders_frequency_5_Cat.csv'[2]. The file contains 470 dynamically observed behaviors and their frequencies of occurrence during dynamic application analysis. The content of the extracted '.csv' file is presented in Table 2, and the descriptions of the observed columns are presented in Table 3.

1 https://www.unb.ca/cic/datasets/maldroid-2020.html
2 https://drive.google.com/file/d/1CuLCATUoxK42LsJhFkV1Vk85Wi7vFXGc/view?usp=sharing

**Table 2**

*Visualization of the First Five System Calls From the First Five Records Contained in the 'feature_vectors_syscallsbinders_frequency_5_Cat.csv' File With Their Respective Frequencies*

| ACCESS_PERSONAL_INFO___ | ALTER_PHONE_STATE___ | ANTI_DEBUG_____ | CREATE_FOLDER_____ | CREATE_PROCESS`_____ |
|---|---|---|---|---|
| 1 | 0 | 0 | 3 | 0 |
| 3 | 0 | 0 | 6 | 0 |
| 2 | 0 | 0 | 4 | 0 |
| 1 | 0 | 0 | 4 | 0 |
| 3 | 0 | 0 | 11 | 0 |

**Table 3**

*Description of the First Row in Table 2*

| System Call Name | Description |
|---|---|
| ACCESS_PERSONAL_INFO___ | Permits access to personal information |
| ALTER_PHONE_STATE___ | Modifies the phone's state variable |
| ANTI_DEBUG_____ | Protects against debugging techniques |
| CREATE_FOLDER_____ | Creates a folder or directory |
| CREATE_PROCESS`_____ | Creates a new process |

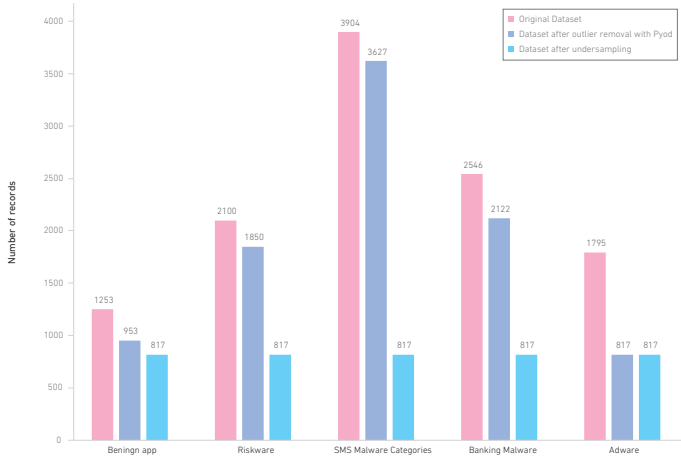
### 4.2 Feature Engineering

The Google Colab programming environment was used, with Python as the programming language. For the removal of the most representative outliers, the PyOD library (a Python library for detecting outlier objects in multivariate datasets) was employed, utilizing the KNN class for outlier detection. According to the documentation of the PyOD KNN model, the distance to the nearest neighbor of an observation can be viewed as its outlying score. The model also includes a parameter called contamination, which represents the proportion of outliers in the dataset. For the purposes of this study, the contamination value of 0,02 was selected.

Subsequently, class balancing was performed using the imblearn library with the undersampling method. Figure 2 shows the comparison of the number of records in the original dataset and after processing with the PyOD library and class balancing with the imblearn library.

**Figure 2**

*Comparison of the Number of Samples in the Original Dataset Versus the Processed Dataset*




Finally, variable selection based on importance was conducted using the SelectFromModel[3] library and the RF algorithm. This technique involves identifying the most relevant features in a dataset by assessing their importance scores. These scores are determined by the RF model, which evaluates how much each feature contributes to increasing or reducing impurity across all the trees in the forest.

To achieve this, various threshold values were tested to evaluate the models and compare accuracy metrics. The threshold corresponding to the best result in Table 4 was selected.

**Table 4**

*Experimentation Carried Out With Different Thresholds to Determine Which Value to Choose*

| Threshold | Number of Features | Accuracy |
|---|---|---|
| 0,001 | 153 | 0,9204 |
| 0,002 | 127 | 0,9241 |
| 0,003 | 107 | 0,9216 |
| 0,004 | 89 | 0,9253 |
| 0,005 | 75 | 0,9192 |
| 0,006 | 62 | 0,9204 |
| 0,007 | 54 | 0,9228 |

(*continues*)

3 https://scikit-learn.org/stable/modules/feature_selection.html#select-from-model

(*continued*)

| Threshold | Number of Features | Accuracy |
|---|---|---|
| 0,008 | 49 | 0,9155 |
| 0,009 | 42 | 0,9228 |
| 0,01 | 34 | 0,9216 |

Since the threshold yielding the highest accuracy result is 0,004 (corresponding to an accuracy of 0,9253), a selection will be made of the 89 attributes whose importance values exceed the selected threshold.

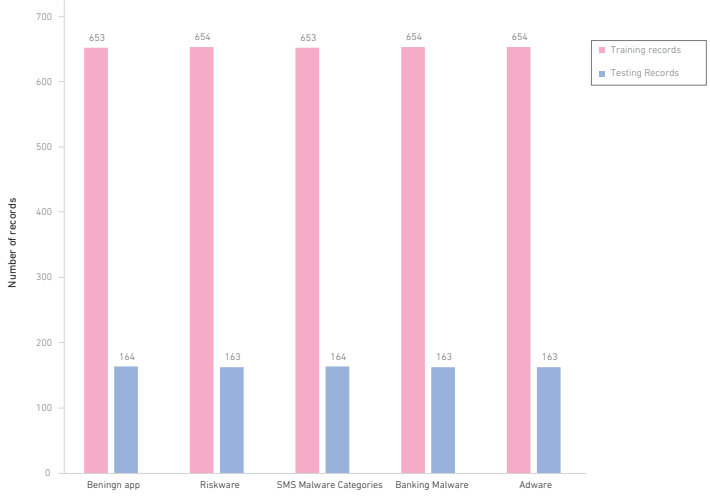
## 5. Results

### 5.1 Training and Test Samples

Training with the machine learning models will use 80 % of the records, while the remaining 20 % will be allocated for testing purposes. Figure 3 illustrates the distribution of training and testing records for each category.

**Figure 3**

*Comparison of the Number of Training and Testing Records*




### 5.2 Machine Learning Models

Table 5 presents the results obtained for each algorithm, including their standard deviation across folds.

**Table 5**

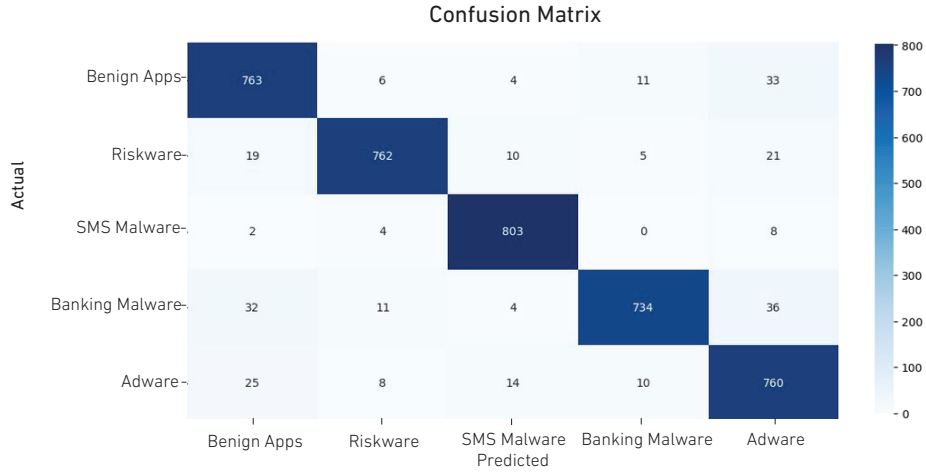
*Comparative Table of Precision, Recall, Accuracy, and F1-Score Results for Machine Learning Algorithms With Standard Deviation Across Folds*

| Algorithms | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Random Forest | 0,9356 ± 0,0140 | 0,9379 ± 0,0129 | 0,9356 ± 0,0140 | 0,9358 ± 0,0139 |
| XGBoost | 0,9349 ± 0,0118 | 0,9362 ± 0,0111 | 0,9349 ± 0,0118 | 0,9350 ± 0,0116 |
| Light GBM | 0,9395 ± 0,0105 | 0,9410 ± 0,0102 | 0,9395 ± 0,0105 | 0,9396 ± 0,0105 |

The data within Table 5 presents the precision, recall, accuracy, and F1-score outcomes achieved by three distinct machine learning algorithms: RF, XGBoost, and Light GBM. Despite relatively minor differences, Light GBM emerged as the top performer, attaining the highest accuracy of 0,9395 ± 0,0105, the highest precision score of 0,9410 ± 0,0102, and the highest recall rate of 0,9395 ± 0,0105. XGBoost followed closely with the second highest accuracy of 0,9349 ± 0,0118, the second best precision score of 0,9362 ± 0,0111, and the second best recall rate of 0,9349 ± 0,0118. RF trailed slightly behind, achieving an accuracy of 0,9356 ± 0,0140, a precision score of 0,9379 ± 0,0129, and a recall rate of 0,9356 ± 0,0140, which was the lowest among the three algorithms.

**Figure 4**

*Confusion Matrix of the RF Algorithm After Applying Stratified K-Fold Cross-Validation (10 Folds)*




Or the Benign Apps class, the model achieved an excellent accuracy by correctly identifying 763 instances. However, it misclassified 6 instances as riskware, 4 as SMS malware, 11 as banking malware, and 33 as adware.

In the riskware class, the model accurately predicted 762 instances but misclassified 19 as benign apps, 10 as SMS malware, 5 as banking malware, and 21 as adware.

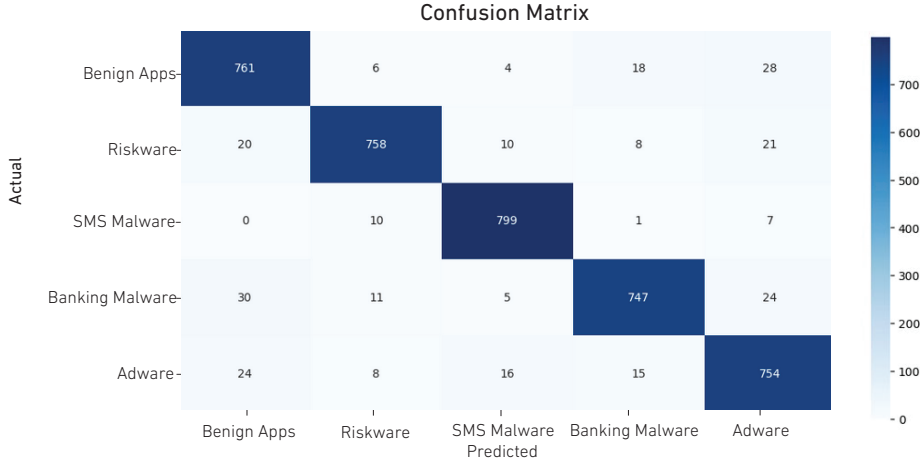
The model exhibited high accuracy for the SMS malware class, correctly classifying 803 instances. Misclassifications were minimal, with 2 instances classified as benign apps, 4 as riskware, and 8 as adware.

For banking malware class, the model correctly predicted 734 instances but misclassified 32 instances as benign apps, 11 as riskware, 4 as SMS malware, and 36 as adware.

Finally, the Adware class achieved high accuracy with 760 instances correctly classified. Misclassifications were relatively low, with 25 instances classified as benign apps, 8 as riskware, 14 as SMS malware, and 10 as banking malware.

**Figure 5**

*Confusion Matrix of the XGBoost Algorithm After Applying Stratified K-Fold Cross-Validation (10 Folds)*




As shown in Figure 5, for the benign apps class, the model correctly identified 761 instances but misclassified 6 as riskware, 4 as SMS malware, 18 as banking malware, and 28 as adware.

In the riskware class, 758 instances were accurately predicted, while 20 were misclassified as benign apps, 10 as sms malware, 8 as banking malware, and 21 as adware.

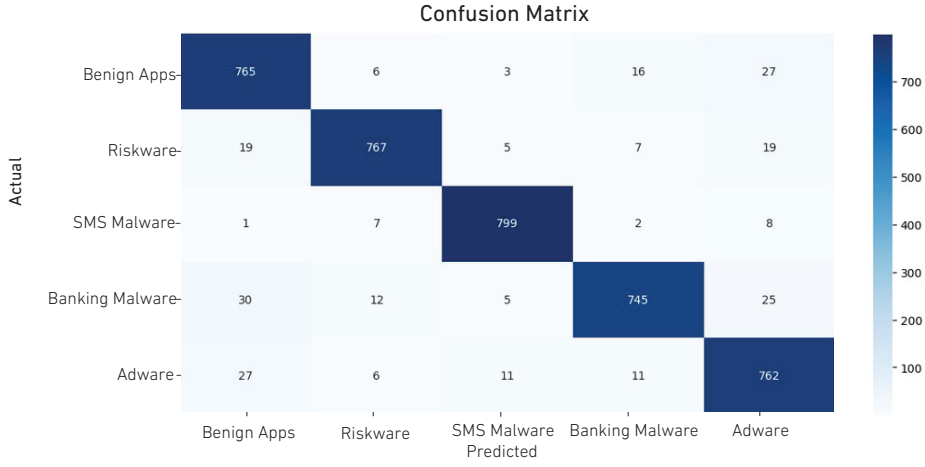
The model exhibited a high accuracy for the SMS malware class, correctly classifying 799 instances, with relatively low misclassifications: 10 as riskware, 1 as banking malware, and 7 as adware.

For the banking malware class, 747 instances were accurately predicted, but 30 were misclassified as benign apps, 11 as riskware, 5 as SMS malware, and 24 as adware.

Finally, the adware class achieved high accuracy with 754 instances correctly classified. Misclassifications were relatively low: 24 as benign apps, 8 as riskware, 16 as SMS malware, and 15 as banking malware.

**Figure 6**

*Confusion Matrix of the LightGBM Algorithm After Applying Stratified K-Fold Cross-Validation (10 Folds)*




As illustrated in Figure 6, for the benign apps class, the model correctly identified 765 instances but misclassified 6 as riskware, 3 as SMS malware, 16 as banking malware, and 27 as adware.

In the riskware class, 767 instances were accurately predicted, while 19 were misclassified as benign apps, 5 as SMS malware, 7 as banking malware, and 19 as adware.

The model exhibited high accuracy for the SMS malware class, correctly classifying 799 instances, with relatively low misclassifications: 1 as bening apps, 7 as riskware, 2 as banking malware, and 8 as adware.

For the banking malware class, 745 instances were accurately predicted, but 30 were misclassified as benign apps, 12 as riskware, 5 as SMS malware, and 25 as adware.

Finally, the adware class achieved high accuracy with 762 instances correctly classified and relatively low misclassifications: 27 as benign apps, 6 as riskware, 11 as SMS malware, and 11 as banking malware.

## 6. DISCUSSION

As detailed in the Results section, all models achieved weighted accuracies above 93 %, which are considered good overall. Furthermore, as observed in Table 5, the highest accuracy of 93,95 %, the best precision of 94,1%, and the best recall of 93,95 % were obtained with the LightGBM algorithm. These results could be attributed to the unique features of the LightGBM algorithm.

According to Ke et al. (2017), LightGBM is a decision tree-based model designed to work efficiently with large datasets. It employs a variant of the gradient boosting technique, focusing on leaf-wise tree growth, which improves its efficiency and scalability. Moreover, LightGBM utilizes advanced techniques like gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). GOSS helps to reduce the variance of the model by focusing on instances with larger gradients, leading to more accurate predictions. EFB bundles mutually exclusive features, reducing the number of splits needed and improving computational efficiency. These features likely contributed to a better fit of the model to the data, avoiding issues such as overfitting and improving its generalization ability.

The processing of the CICMalDroid 2020 dataset showed different results compared to previous studies that employed the same dataset. For instance, Sönmez et al. (2021) used machine learning algorithms for malware family classification, achieving a maximum precision of 90,2 % and an average recall of 89,54 % with the KNN algorithm. In contrast, our study achieved a maximum precision of 94,1 % and a recall of 93,95 % with the LightGBM model. This variance can be explained by our use of Decision Tree-based and boosting algorithms, leading to better outcomes.

On the other hand, Bhatia and Kaushal (2017) researched malware detection using system call frequencies, achieving a precision of 88,9 %. In contrast, our study obtained a precision of 94,1 %. This discrepancy can be attributed to differences in dataset size; while the authors used 100 records, this research employed a dataset of 4 085 records. Additionally, the dataset employed in this research was processed using CopperDroid, a tool that enables high-level system call extraction in Android applications, whereas the authors used the "strace" command for their system call collection.

Kshirsagar and Agrawal (2022) focused their study on feature selection using methods traditionally employed in malware detection systems. They achieved a higher precision of 97,46 %; in contrast, our study achieved the highest precision of 94,1%. This difference in results can be attributed to the feature selection methods used. The authors selected 80 features from 470 original features by using the ReliefF method, which identifies the most relevant features in a dataset based on their ability to distinguish between instances from different classes. In contrast, our study employed 89 features obtained from SelectFromModel, a wrapper-based method that selects features based on their importance in a pre-trained machine learning model.

It should be noted that some limitations were found in this research: the impossibility of extracting the dynamic behaviors of the APKs with CooperDroid due to its discontinuation, which needed the use of the '.csv' file mentioned in the Methodology section. It is also worth mentioning that the obtained data was unbalanced, as the class related to SMS malware was the majority class. Consequently, undersampling had to be performed, reducing the amount of data compared to the original records to achieve balance.

## 7. CONCLUSIONS AND FUTURE WORK

In the present research, the effectiveness of machine learning algorithms (RF, LightGBM, and XGBoost) was evaluated to identify the most appropriate model to prevent malware attacks, using a dataset of dynamically observed malware behaviors. The results indicate that the more data used to train the machine learning models, the better the classification between families. After preprocessing, 817 records per family were achieved, totaling 4 085 records. Increasing the amount of data is expected to have a positive impact by improving the accuracy rate.

Having compared the results of this study with previous research in the Discussion section, it is concluded that dynamically observed behaviors of malware can be successfully employed in malware family classification with the assistance of machine learning models. Dynamic behaviors offer a more detailed insight into malware characteristics, enabling a finer classification.

Finally, for future research, the plan is to expand the study by reducing the limitations discussed in the previous section, specifically by processing similar numbers of malware applications to avoid class imbalance. Likewise, we will seek to implement methodologies employed by different authors to reduce the rate of false positives in the classification.

## REFERENCES

Aebersold, S., Kryszczuk, K., Paganoni, S., Tellenbach, B., & Trowbridge, T. (2016). Detecting obfuscated JavaScripts using machine learning. *ICIMP 2016 the Eleventh International Conference on Internet Monitoring and Protection: May 22-26, 2016*, Valencia, Spain, *1,* 11–17. https://doi.org/10.21256/zhaw-3848

Ashik, M., Jyothish, A., Anandaram, S., Vinod, P., Mercaldo, F., Martinelli, F., & Santone, A. (2021). Detection of malicious software by analyzing distinct artifacts using machine learning and deep learning algorithms. *Electronics*, *10*(14), 1694. https://doi.org/10.3390/electronics10141694

Aslan, Ö. A., & Samet, R. (2020). A comprehensive review on malware detection approaches. *IEEE Access*, *8*, 6249-6271. https://doi.org/10.1109/ACCESS.2019.2963724

Bhatia, T., & Kaushal, R. (2017). Malware detection in android based on dynamic analysis. *2017 International Conference on Cyber Security and Protection Of Digital Services (Cyber Security)*, London, UK, 1-6. https://doi.org/10.1109/CyberSecPODS.2017.8074847

Chen, S., Xue, M., Fan, L., Hao, S., Xu, L., Zhu, H., & Li, B. (2018). Automated poisoning attacks and defenses in malware detection systems: An adversarial machine learning approach. *Computers and Security, 73*, 326-344. https://doi.org/10.1016/j.cose.2017.11.007

Chen, Y. C., Chen, H. Y., Takahashi, T., Sun, B., & Lin, T. N. (2021). Impact of code deobfuscation and feature interaction in android malware detection. *IEEE Access*, *9*, 123208-123219. https://doi.org/10.1109/ACCESS.2021.3110408

Chen, Z., & Ren, X. (2023). An efficient boosting-based windows malware family classification system using multi-features fusion. *Applied Sciences*, *13*(6), 4060. https://doi.org/10.3390/app13064060

Choudhary, S., & Sharma, A. (2020, February). Malware detection & classification using machine learning. *2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3)*, Lakshmangarh, India, 1-4. https://doi.org/10.1109/ICONC345789.2020.9117547

Dhamija, H., & Dhamija, A. K. (2021). Malware detection using machine learning classification algorithms. *International Journal of Computational Intelligence Research (IJCIR)*, *17*(1), 1-7. https://www.ripublication.com/ijcir21/ijcirv17n1_01.pdf

Duo, W., Zhou, M., & Abusorrah, A. (2022). A survey of cyber attacks on cyber physical systems: Recent advances and challenges. *IEEE/CAA Journal of Automatica Sinica*, *9*(5), 784-800. https://doi.org/10.1109/JAS.2022.105548

Feng, P., Ma, J., Sun, C., Xu, X., & Ma, Y. (2018). A novel dynamic android malware detection system with ensemble learning. *IEEE Access*, *6*, 30996-31011. https://doi.org/10.1109/ACCESS.2018.2844349

Fortinet. (2022, February 8). *América Latina sufrió más de 289 mil millones de intentos de ciberataques en 2021* [Press release]. https://www.fortinet.com/lat/corporate/about-us/newsroom/press-releases/2022/fortiguard-labs-reporte-ciberataques-america-latina-2021

Gao, Y., Hasegawa, H., Yamaguchi, Y., & Shimada, H. (2022). Malware detection using LightGBM with a custom logistic loss function. *IEEE Access*, *10*, 47792-47804. https://doi.org/10.1109/ACCESS.2022.3171912

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA. https://

proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Kim, S., Hong, S., Oh, J., & Lee, H. (2018, June). Obfuscated VBA macro detection using machine learning. *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, Luxembourg, Luxembourg, 490-501. https://doi.org/10.1109/DSN.2018.00057

Kshirsagar, D., & Agrawal, P. (2022). A study of feature selection methods for android malware detection. *Journal of Information and Optimization Sciences*, *43*(8), 2111-2120. https://doi.org/10.1080/02522667.2022.2133218

Kumar, R., & Geetha, S. (2020). Malware classification using XGboost-Gradient boosted decision tree. *Advances in Science, Technology and Engineering Systems Journal*, *5*(5), 536–549. https://doi.org/10.25046/aj050566

Mahdavifar, S., Kadir, A. F. A., Fatemi, R., Alhadidi, D., & Ghorbani, A. A. (2020). Dynamic android malware category classification using semi-supervised deep learning. *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, Calgary, AB, Canada, 515-522. https://doi.org/10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00094

Mahindru, A., & Sangal, A. L. (2021). MLDroid—Framework for Android malware detection using machine learning techniques. *Neural Computing and Applications*, *33*, 5183-5240. https://doi.org/10.1007/s00521-020-05309-4

Liu, K., Xu, S., Xu, G., Zhang, M., Sun, D., & Liu, H. (2020). A review of android malware detection approaches based on machine learning. *IEEE Access*, *8*, 124579-124607. https://doi.org/10.1109/ACCESS.2020.3006143

Louk, M. H. L., & Tama, B. A. (2022). Tree-based classifier ensembles for PE malware analysis: A performance revisit. *e*, *15* (9), 332. https://doi.org/10.3390/a15090332

Onoja, M., Jegede, A., Blamah, N., Abimbola, O. V., & Omotehinwa, T. O. (2022). EEMDS: Efficient and effective malware detection system with hybrid model based on xceptioncnn and lightgbm algorithm. *Journal of Computing and Social Informatics*, *1*(2), 42-57. https://doi.org/10.33736/jcsi.4739.2022

Palša, J., Ádám, N., Hurtuk, J., Chovancová, E., Madoš, B., Chovanec, M., & Kocan, S. (2022). MLMD—A malware-detecting antivirus tool based on the XGBoost machine learning algorithm. *Applied Sciences*, *12*(13), 6672. https://doi.org/10.3390/app12136672

Şahın, D. Ö., Akleylek, S., & Kiliç, E. (2022). LinRegDroid: Detection of Android malware using multiple linear regression models-based classifiers. *IEEE Access*, *10*, 14246–14259. https://doi.org/10.1109/ACCESS.2022.3146363

Sönmez, Y., Salman, M., & Dener, M. (2021). Performance analysis of machine learning algorithms for malware detection by using CICMalDroid2020 dataset. *Düzce University Journal of Science and Technology*, *9*(6), 280-288. https://doi.org/10.29130/dubited.1018223

Surendran, R., & Thomas, T. (2022). Detection of malware applications from centrality measures of syscall graph. *Concurrency and Computation: Practice and Experience*, *34*(10). https://doi.org/10.1002/cpe.6835

Surendran, R., Thomas, T., & Emmanuel, S. (2020). On existence of common malicious system call codes in android malware families. *IEEE Transactions on Reliability*, *70*(1), 248-260. https://doi.org/10.1109/TR.2020.2982537

Urooj, B., Shah, M. A., Maple, C., Abbasi, M. K., & Riasat, S. (2022). Malware detection: a framework for reverse engineered android applications through machine learning algorithms. *IEEE Access*, *10*, 89031-89050. https://doi.org/10.1109/ACCESS.2022.3149053

Wu, B., Chen, S., Gao, C., Fan, L., Liu, Y., Wen, W., & Lyu, M. R. (2021). Why an android app is classified as malware: Toward malware classification interpretation. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *30*(2), 1-29. https://doi.org/10.1145/3423096

# Determining the Scale of The Most Prominent Asset Management Firms Using Data Analytics

Erick Leonel García Ibáñez
egarciasc87@gmail.com
https://orcid.org/0000-0001-8952-9143
Peter the Great St. Petersburg Polytechnic University
St. Petersburg, Russia

Received: April 30th, 2024 / Accepted: June 1st, 2024
doi: https://doi.org/10.26439/interfases2024.n19.7072

ABSTRACT. Peter Phillips and other prominent authors shed light on the existence of hidden superpowers that exert influence over the world's most important corporations. This power also extends to major global communications entities. Phillips highlights that 17 of the so-called 'Giants' manage portfolios with a staggering $1 trillion or more in assets. Additionally, there is a small group of institutions called the 'Big Three,' globally recognized asset management giants comprising BlackRock, Vanguard Group, and State Street Corporation. The discourse surrounding these 'Giants' and the 'Big Three' prompted the application of data analytics techniques to scrutinize and validate these assertions. Data, drawn from diverse sources such as Yahoo Finance and StockAnalysis.com, underwent meticulous cleaning to mitigate errors and address missing information. Subsequently, three hypotheses were subjected to rigorous evaluation through hypothesis testing. As a result of this process, only one of the three hypotheses proposed was rejected, leading us to conclude that the 'Big Three' indeed hold a very strong position among companies that belong to the most important indices (S&P 500, NASDAQ 100, DJIA) of the American financial markets.

KEYWORDS: hypothesis testing / asset management firms / the Big Three

## Determinación de la escala de las firmas de gestión de activos más prominentes mediante el análisis de datos

RESUMEN. Peter Phillips y otros autores destacados iluminan la existencia de superpotencias ocultas que ejercen influencia sobre las corporaciones más importantes del mundo, este poder también abarca a importantes entidades de comunicaciones a nivel global. Phillips resalta que 17 de los llamados 'Gigantes' gestionan portafolios de inversión con al menos $1 trillón en activos. De la misma manera, existe un grupo pequeño de gestores de activos denominado los 'Tres Grandes' ('Big Three'), gigantes

reconocidos a nivel mundial en la gestión de activos que incluyen a BlackRock, Vanguard Group y State Street Corporation. El discurso en torno a estos 'Gigantes' y los 'Tres Grandes' motivó la aplicación de técnicas de análisis de datos para examinar y validar las afirmaciones realizadas sobre ellos. Los datos, recopilados de diversas fuentes como Yahoo Finance y StockAnalysis.com, fueron sometidos a una limpieza meticulosa para mitigar errores y abordar la falta de información. Posteriormente, tres hipótesis fueron sometidas a una evaluación rigurosa mediante pruebas de hipótesis (hypothesis testing). Como resultado de este proceso, solo una de las tres hipótesis planteadas fue rechazada, lo cual nos llega a concluir que los 'Tres Grandes' efectivamente poseen una posición muy fuerte en las empresas que pertenecen a los índices más importantes (S&P 500, NASDAQ 100, DJIA) de los mercados financieros norteamericanos.

PALABRAS CLAVE: prueba de hipótesis / firmas de gestión de activos / los Tres Grandes

## 1. INTRODUCTION

The arrival of globalization meant that certain brands became famous worldwide, some of which include Starbucks, Google, Facebook or Coca Cola. However, there are other big, and perhaps unknown, companies who possess a disproportionate stake in these popular brands. Peter Phillips (2018) identifies the top 17 asset management firms, which he calls "Giants". These firms hold very strong positions among companies that belong to the most important indices (S&P 500, NASDAQ 100, DJIA) of the American financial markets. Accordingly, a consensus among various authors such as Lund and Robertson (2023) and McLauglin and Massa (2020) reinforce the acknowledgment of the widely recognized concept known as the 'Big Three'. These mega companies are BlackRock, Vanguard Group and State Street Corporation.

Both the Giants and Big Three are entities that maintain extensive holdings in numerous companies across various industries and countries. One might wonder if they truly are of such immense scale or whether it beneficial for a handful of companies to wield such substantial power and influence worldwide. However, this article does not delve into ethical considerations; rather, its primary focus is to assess the global relevance and influence of the Big Three in international business. This relevance will be measured by confirming or rejecting various hypotheses derived from statements made by Phillips (2018), Lund and Robertson (2023) and McLauglin and Massa (2020), utilizing data analytics techniques. Data will be collected from different sources, cleaned, transformed and analyzed. In the following section, the basics of investments and data analytics will be explained.

### Basic concepts in asset management

According to Smart et al. (2017), a common stock offers income and capital gains. An example of income gain are dividend distributions from the company's profits, whereas a capital gain refers to the stock's increase in price. Common stocks cannot be bought directly from the issuer, but have to be traded through a centralized entity that connects buyers and sellers, called stock exchange. Furthermore, an asset can be understood as anything that delivers value to the organization or stakeholders (Canadian Network of Asset Managers, 2018). An investment is the action of acquiring an asset with the main purpose of generating profit. Organizations whose goal is to execute investing activities and manage investment portfolios are usually called asset management firms. The Canadian Network of Asset Managers (2018) describes asset management as an integrated process with the aim of effectively managing assets in order to deliver services to customers.

Collin (2023) identifies the steps of the asset management process. The first step consists on setting the goals and market assumptions for the portfolio. The strategy for asset allocation will then be developed along with capital deployment. The company should pick the financial assets in which the funds shall be invested: stocks, mutual funds, exchange traded funds (ETF) and others. Smart et al. (2017) assert that both

mutual funds and ETFs,are similar in that they allow the creation of well-diversified portfolios by holding a variety of securities. Nonetheless, ETFs can be bought or sold at current market price in regular stock exchanges. while mutual funds are traded through the fund itself or through a financial intermediary (Smart et al., 2017; U.S. Securities and Exchange Commission, 2021)

### Basic concepts in data analytics

The relevance of data and data strategies for the organization is undeniable. Kambatla et al. (2014) refer to data as a resource, while Arora and Goyal (2016) point out that data is the biggest asset inside an organization. But data itself is not enough: it must be analyzed to derive value from it. This is where data analytics comes into place as a set of techniques that focus on gaining actionable insight for smart decision-making (Duan & Da Xu, 2021).

Arora and Goyal (2016) refer to four types of data according to its source or precedence, while other authors such as Kambatla et al. (2014), Vashisht and Gupta (2015) and Vanani and Majidian (2019) identify 3 categories and Stevens (2023) establishes only 2 categories. In this research, the 3-category classification is suggested: structured (e.g. Excel files, relational databases), semi-structured (e.g. text, emails) and unstructured (e.g. videos, audios).

The classification of data analytics methods is also a matter of debate. Duan and Da Xu (2021) describe 3 categories and Kelley (2020) details 6 categories, but the most popular is one with 4 categories described in Oracle (s/f), Stevens (2023) and Mathur (2023). Tom March (2020) describes the 4 main types of data analytics as follows:

- Descriptive Analysis: Valuable for recognizing past events.
- Diagnostic Analysis: Convenient for explaining the reasons behind occurrences.
- Predictive Analysis: Utilized to forecast future trends based on historical data.
- Prescriptive Analysis: Capable of predicting probable outcomes and offering decision recommendations.

Section 2 provides an in-depth examination of the techniques and methods available and explains the process utilized to prove or disprove the hypotheses. In Section 3, the project outcomes are elucidated. Section 4 offers a concise discussion on the hypothesis and collected data. Ultimately, Section 5 encapsulates the research's conclusions.

## 2. MATERIALS AND METHODS

### Instruments

In data analytics, various approaches exist concerning the processes, methods and techniques employed. This section undertakes a comparative analysis of these approaches.

Notably, researchers such as Kelley (2020), Stevens (2023) and Peck et al. (2008) present different perspectives on the phases involved in data analytics projects, ranging from 5 to 6 phases, which is laid out in Table 2.

**Table 1**

*Phases in a data analytics project*

| Phase | Kelley (2020) | Stevens (2023) | Peck et al. (2008) |
|---|---|---|---|
| #1 | - Data Requirement Gathering | - Defining the question | - Understanding the nature of the problem |
| #2 | - Data Collection | - Data Collection | - Deciding what to measure and how to measure it. |
| #3 | - Data Cleaning | - Data Cleaning | - Data Collection |
| #4 | - Data Analysis | - Data Analysis | - Data summarization and preliminary analysis |
| #5 | - Data Interpretation | - Data Visualization and sharing findings | - Formal data analysis |
| #6 | - Data Visualization | | - Interpretation of results |

Numerous criteria exist for classifying data analytics methods and techniques (Duan & Da Xu, 2021; Kelley, 2020; Stevens, 2023; Taherdoost, 2022; Vashisht & Gupta, 2015). Some of them are highlighted in Table 2. One prevalent classification method involves categorizing methods into two broad groups: qualitative and quantitative. As elucidated by Kelley (2020), qualitative analysis concentrates on handling unstructured data, while quantitative analysis involves the collection and processing of raw data into numerical formats.

**Table 2**

*Classification of data analytics methods*

| Qualitative Analysis | Quantitative Analysis | | | | |
|---|---|---|---|---|---|
| | Descriptive | Inferential | Exploratory | Predictive | Prescriptive |
| - Content analysis | - Mean | - Regression analysis | - Cluster analysis | - Support vector machines | - Genetic algorithm |
| - Sentiment analysis | - Median | - Time series analysis | - Factor analysis | - Neural networks | |
| Qualitative Analysis | Quantitative Analysis | | | | |
| | Descriptive | Inferential | Exploratory | Predictive | Prescriptive |
| - Narrative analysis | - Mode | - Hypothesis testing | | - Decision tree | |
| - Grounded theory | - Standard deviation | - ANOVA | | | |
| | - Skewness | | | | |

**Methodology design**

We employed inferential analysis and hypothesis testing to validate certain statements outlined by Phillips (2018), Lund and Robertson (2023) and McLauglin and Massa (2020). The project unfolded in five phases, following the approach detailed by Stevens (2023). In hypothesis testing, a hypothesis is a proposition regarding the value of one or more population characteristics (Peck et al., 2008). The null hypothesis ($H_0$), often the initial assumption, is the claim presumed to be true. Conversely, the alternative hypothesis ($H_a$), which represents an alternative scenario, stands as the second option. The outcome of the process may lead to either rejecting the null hypothesis or failing to reject it, signifying that the alternative hypothesis is deemed correct.

In Figure 1 you can see some of the most important elements in hypothesis testing. First, confidence level refers to the probability that the null hypothesis will be accepted. Conversely, significance level ($\alpha$) is in the rejection area of the null hypothesis. Both confidence and significance level add up to 100% (or 1). P-value is referred to as the probability of rejecting the null hypothesis. If the p-value is lower than $\alpha$, the null hypothesis is rejected. Otherwise, the test fails to reject null hypothesis.

The example in Figure 3 shows a two-tailed test as it has two rejection areas on both sides. A one-tailed test has only one rejection area, on the right or left.

**Figure 1**

*Elements in hypothesis testing*

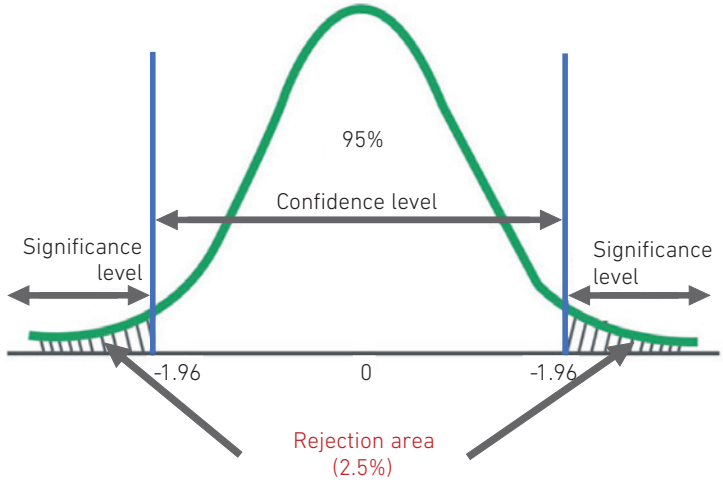
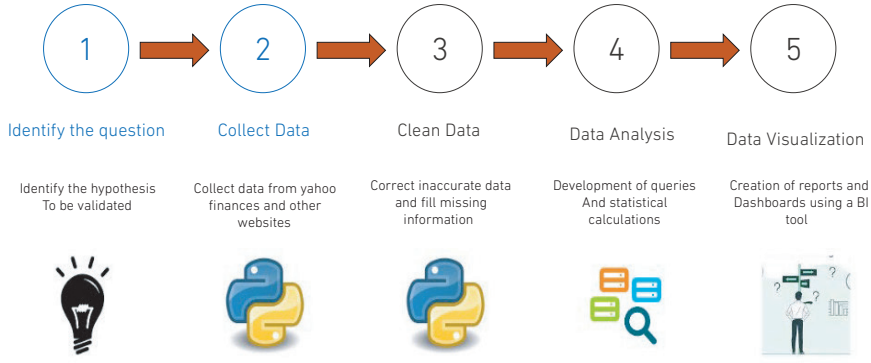



Figure 2 illustrates the stages of the data analytics project. Next, we will elaborate on each step:

**Figure 2**

*Phases in data analytics project*




**Identify the question**

In his book *Giants*, Peter Phillips (2018) states his intention to direct our concern towards powerful networks that affect our lives and society, making the following statements regarding this powerful grid:

- 17 of the most prominent asset management firms possess over $1 trillion in assets as of 2017, as seen in Table 3.
- These top firms also tend to hold positions in each other, creating a solid network with shared investments worldwide.
- There are a total of 199 people in the boards of directors of the top 17 asset management firms, who control and manage an accumulation of $41.1 trillion in assets.

While they do not agree on the exact numbers, Lund and Robertson (2023) and McLauglin and Massa (2020) acknowledge BlackRock as the largest global asset manager draw the following deductions about the Big Three:

- They represent the largest owners of most companies included in the S&P 500, owning at least 22 % in 2019.
- Collectively, they hold an average 25 % of the votes in companies included in S&P 500.
- They have the power to influence the outcome of shareholder proposals at companies included in the Fortune 250.
- The Big Three are closely and exclusively related to passive investment.

Phillips (2018), Lund and Robertson (2023) and McLauglin and Massa (2020) expounded on a set of statements regarding some of the largest asset management

firms globally. The first and second statements center around the Big Three mentioned in Lund and Robertson (2023) and McLauglin and Massa (2020), bringing up some important facts about their relevance and investing strategy. The last statement pertains to the top 5 US asset management firms selected from the list of 17 companies outlined by Phillips (2018), which includes BlackRock, Vanguard Group, JP Morgan Chase, Bank of America, and State Street. Table 3 contains the 3 tested hypotheses.

**Table 3**

*Main statements made by Phillips (2018), Lund and Robertson (2023) and McLauglin and Massa (2020)*

| Order | Statement | Null hypothesis ($H_0$) | Alternative hypothesis ($H_a$) |
|---|---|---|---|
| #1 | The Big Three own 22 % of companies included in the S&P 500, each one owning 7.33 %. | p = 7.33 % | p <> 7.33 % |
| #2 | The "Big Three" are closely and exclusively related to passive investment (at least 90 % of the portfolio). | p ≥ 90 % | p < 90 % |
| #3 | Global media are controlled by the "Global Power Elite". The Big Three own at least 15.56 % of the largest global media corporations. | p ≥ 15.56 % | p < 15.56 % |

**Collect Data**

Data collection was performed using web scraping in the Python programming language. Collected data includes: company data (e.g. country, sector) from the Stock Analysis (2023a) website, top 10 mutual fund holders from the Yahoo Finance website, and the list of companies that compound the most important indices in the American exchanges: Standard & Poor 500 (Stock Analysis, 2023b), NASDAQ 100 (Stock Analysis, 2023c), Dow Jones Industrial Average (DJIA) (Stock Analysis, 2023d).

Neufeld (2023) singled out the leading 25 global stock exchanges, which boast a cumulative market capitalization of around $107 trillion. Most notably, the New York Stock Exchange (NYSE) at $25 trillion and NASDAQ at $21.7 trillion collectively command nearly 44 % of this market value, for which the sample will include companies listed in both exchanges. As part of the analysis process, for each company under examination we systematically gathered data on the top 10 individual shareholders with direct investments in the organization, along with the top 10 mutual funds holding shares in the company through passive investment. Data was collected on March 5th, 2024.

Figure 3 shows an example of the data collection process, where top 10 holders (direct holders and mutual funds) of Apple (AAPL) are being web scrapped. Data was collected on January 28th, 2024.

**Figure 3**

*Data collection process*

```
STEP 2: get data from the web...
Stock 1 of 519 --> AAPL
                                                 Holder      Shares Date Reported  % Out         Value           Type
0                                    Vanguard Group Inc  1299997133  Sep 29, 2023  8.36%  251679452883  Direct Holder
1                                        Blackrock Inc.  1031407553  Sep 29, 2023  6.63%  199680508556  Direct Holder
2                                  Berkshire Hathaway, Inc   915560382  Sep 29, 2023  5.89%  177252495543  Direct Holder
3                                State Street Corporation   569291690  Sep 29, 2023  3.66%  110214874658  Direct Holder
4                                             FMR, LLC   298321726  Sep 29, 2023  1.92%   57755087974  Direct Holder
5                           Geode Capital Management, LLC   296103070  Sep 29, 2023  1.90%   57325556159  Direct Holder
6                           Price (T.Rowe) Associates Inc   216307878  Sep 29, 2023  1.39%   41877206501  Direct Holder
7                                      Morgan Stanley   206732960  Sep 29, 2023  1.33%   40023502317  Direct Holder
8                             Northern Trust Corporation   168874976  Sep 29, 2023  1.09%   32694196384  Direct Holder
9                         Norges Bank Investment Management   167374278  Dec 30, 2022  1.08%   32403661242  Direct Holder
10                  Vanguard Total Stock Market Index Fund   462496298  Sep 29, 2023  2.97%   89539286115     Mutual Fund
11                              Vanguard 500 Index Fund   353157634  Sep 29, 2023  2.27%   68371320097     Mutual Fund
12                              Fidelity 500 Index Fund   170161953  Oct 30, 2023  1.09%   32943355139     Mutual Fund
13                               SPDR S&P 500 ETF Trust   163961069  Sep 29, 2023  1.05%   31742863959     Mutual Fund
14                            iShares Core S&P 500 ETF   138678373  Sep 29, 2023  0.89%   26848133859     Mutual Fund
15                           Vanguard Growth Index Fund   128896004  Sep 29, 2023  0.83%   24954267161     Mutual Fund
16                Invesco ETF Tr-Invesco QQQ Tr, Series 1 ETF   124636013  Sep 29, 2023  0.80%   24129532877     Mutual Fund
17  Vanguard Institutional Index Fund-Institutional Index Fund    98610773  Sep 29, 2023  0.63%   19091046254     Mutual Fund
18                Vanguard Information Technology Index Fund    76972129  Aug 30, 2023  0.49%   14901804644     Mutual Fund
19                      Select Sector SPDR Fund-Technology    64668259  Sep 29, 2023  0.42%   12519775337     Mutual Fund
```

### Clean Data

Python programming language was also utilized for the data cleaning process. This phase consisted of validating data and handling missing values. Some of the tasks executed during the phase included performing data exploration, converting text into proper numeric format, dismissung duplicate data related to the total list of companies included in the analysis and standardizing the names of the mutual funds to more easily identify the holder.

### Data Analysis

First, we used hypothesis testing to prove or disprove the thesis that each of the Big Three, as stated in Lund and Robertson (2023) and McLauglin and Massa (2020), own an average of 7.33 % (22 % in total) of the companies included in the S&P 500. Since we are working with proportions and our sample is very limited, t-test was used. Hypothesis was tested with a confidence level of 95 % and alpha value of 0.05. The hypotheses are:

**Null hypothesis ($H_0$):** p = 7.33%.

**Alternative hypothesis ($H_a$):** p <> 7.33%

The second statement under scrutiny points out that the Big Three are closely and exclusively associated with passive investment. Because authors Lund and Robertson (2023) and McLauglin and Massa (2020) do not provide an exact quantification of how much of the portfolio is deemed to consist 'close or exclusively' of passive investment, an assumed figure of 90 % is adopted for analysis purposes. Passive investment, by definition, entails an investment strategy crafted to minimize risk and streamline portfolio selection by favoring assets like mutual funds and ETFs. Additionally, the

t-test was employed for further evaluation, with a confidence level of 95 % and significance level equal to 0.05. The hypotheses are:

**Null hypothesis ($H_0$):** $p \geq 90\ \%$

**Alternative hypothesis ($H_a$):** $p < 90\ \%$

Peter Phillips (2018) states that global media are owned and controlled by the Global Power Elite. Table 4 contains 5 of the companies identified by Phillips (2018), of which two are not listed in any exchange (Time Warner and Viacom/CBS). Next, we calculated what share of the media firms is owned by the Big Three in terms of money and percentage. It is important to highlight that 21st Century Fox is now called Fox Corporation, and Time Warner is now called Warner Bros Discovery. According to the table in Table 4, the Big Three possess an average 15.56 % of the world's largest media as of 2017: 18.16 % of Comcast Corp., 17.81 % of Disney, and 10.72 % of 21st Century Fox.

**Table 4**

*Top transnational news and entertainment corporations (Phillips, 2018)*

| Asset Management Firm | Money invested Media Company (Billion USD) | | | | |
|---|---|---|---|---|---|
| | Comcast Corp. | Disney | Time Warner | Viacom and CBS | 21st Century Fox |
| *BlackRock* | *14.400* | *9.300* | *4.000* | *1.280* | *0.852* |
| *Vanguard Group* | *12.300* | *10.700* | *4.500* | *1.390* | *0.997* |
| *State Street* | *7.300* | *7.100* | *2.700* | *0.861* | *0.588* |
| Bank of America | 2.300 | 1.800 | 0.517 | 0.154 | 0.064 |
| e | 2.160 | 2.500 | 0.707 | 0.252 | 0.135 |
| JPMorgan Chase | 2.100 | 1.800 | 0.636 | 0.476 | 0.162 |
| Capital Group | 2.100 | -- | 0.907 | 1.740 | 0.000 |
| UBS | 1.400 | 0.927 | 0.395 | 0.168 | 0.083 |
| Goldman Sachs Group | 1.190 | 0.921 | 0.908 | 0.109 | 0.207 |
| Prudential Financial | 0.737 | 0.310 | 0.000 | 0.068 | 0.070 |
| Morgan Stanley | 0.663 | 2.700 | 0.624 | 0.215 | 0.569 |
| Total owned by firms | 46.65 | 38.06 | 15.89 | 6.71 | 3.73 |
| Owned by 'Big Three' | 34.00 | 27.10 | 11.20 | 3.53 | 2.44187.19 |
| Market Cap in 2017 | 187.19 | 152.14 | (Not listed) | (Not listed) | 22.74 |
| **% Owned by 'Big Three'** | **18.16 %** | **17.81 %** | **(Not listed)** | **(Not listed)** | **10.72 %** |
| **% Average Owned by 'Big Three'** | | | **15.56 %** | | |

Dellatto (2023) wrote an article listing world's largest media companies as of 2023. The task at hand is to verify the notion that the Big Three still own at least 15.56 % of the largest global media corporations in 2023 using t-test. The sample will include the following publicly traded media companies: Comcast Corp. (NASDAQ: CMCSA), Fox Corp. (NASDAQ: FOX), Disney (NYSE: DIS), Omnicom Group (NYSE: OMC), Charter Communications (NASDAQ: CHTR), and Warner Bros Discovery (NASDAQ: WBD). These are the hypotheses:

**Null hypothesis ($H_0$):** $p \geq 15.56\ \%$

**Alternative hypothesis ($H_a$):** $p < 15.56\ \%$

**Data Visualization: Power BI**

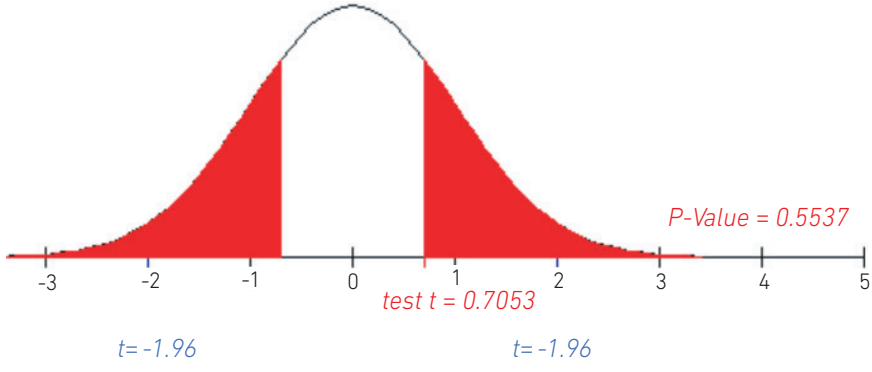
Once data has been properly collected, formatted and analyzed it is time to share the results. There are many data visualization tools in the market today (Tripathi & Bagga, 2020; Srivastava et al., 2022) that can be evaluated and selected according to the user's needs. Some of the most popular are Power BI, Zoho Analytics, Tableau, Locker and Qlik. Power BI was the selected BI tool as it is supported by most platforms. In addition, Power BI works in the cloud and on-promise (Tripathi & Bagga, 2020).

## 3. RESULTS

The first hypothesis states that the "Big Three" hold an average of 7.33 % (22 % in total) of all the companies included in the S&P 500 index. As shown in Figure 4, for a confidence level of 95 % and alpha equal to 0.025 for a two-tailed test, the calculated p-value is 0.5537. Since the p-value is higher than , we fail to reject the null hypothesis ($H_0$): p = 7.33 %.

**Figure 4**

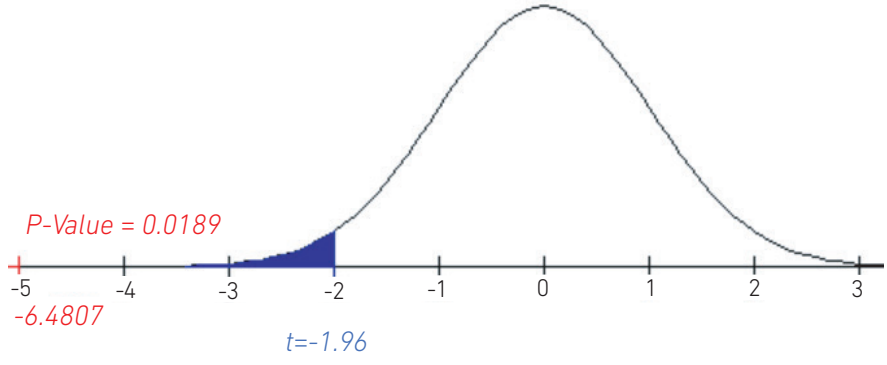
*Illustration of the results of hypothesis testing related to ownership in companies in S&P 500 using t-test*



In regards to the second hypothesis about the Big Three's association with passive investment, the test results are delineated below. As illustrated in Figure 7, the proportion of the portfolio allocated to mutual funds or ETFs was computed under the assumption that it constitutes a minimum of 90 % (≥90 %) of the entire portfolio. Employing a 95% confidence level for this test, corresponding to a significance value of 0.05 for a left-tailed test, the calculated p-value is 0.0189 as shown inc Figure 5. Since the p-value is lower than $\alpha$, we reject the null hypothesis ($H_0$): $p \geq 90$ %.

**Figure 5**

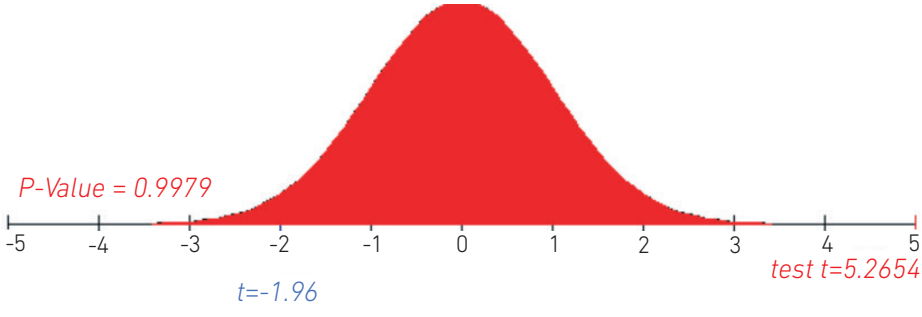
*Illustration of the results of hypothesis testing related to passive asset management using t-test*




The third hypothesis refers to the control of the largest media corporations by certain "Giants" in the asset management sector, confirming that the "Big Three" own at least 15.56 % of big media companies. After considering a confidence level of 95 % and a significance level of 0.05 for a left-tailed test, the p-value is 0.9979 (see Figure 6). Since the p-value is not lower than $\alpha$, we fail to reject the null hypothesis ($H_0$): $p \geq 15.56$ %.

**Figure 6**

*Illustration of the results of hypothesis testing related to ownership in big media companies using t-test*



## 4. DISCUSSION

In the previous chapter three hypotheses were tested using statistical techniques. Here, some charts will be employed to strengthen our results.

McLauglin and Massa (2020) stated that 22 % of the shares in the S&P 500 were held by the Big Three, up from 13 % in 2008. Figure 7 shows the share of the companies in the most important indexes in US exchanges that they own: 28.31 % in S&P 500, 24.54 % in NASDAQ 100, and 28.24 % in DJIA.

**Figure 7**

*Value of ownership of the Big Three in US indices*

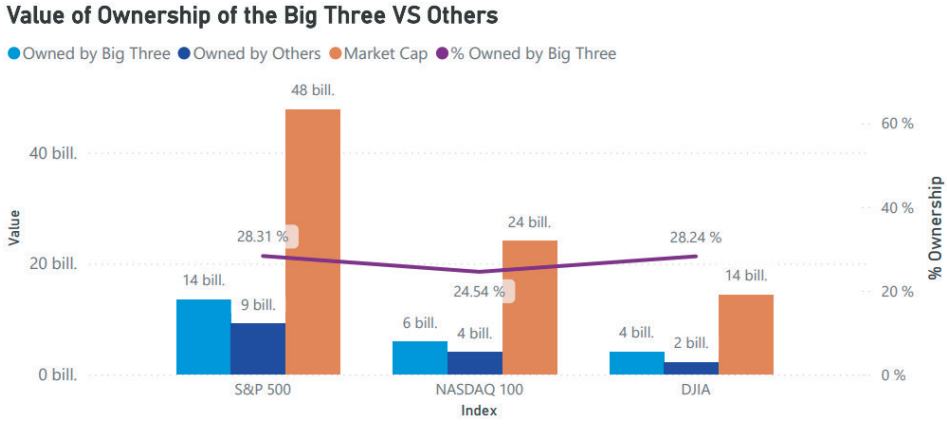



Figure 8. Illustration of the results of hypothesis testing related to the ownership in big media companies using t-test.

Furthermore, individually the Big Three are each among the Top 5 holders in the S&P 500, which are listed in Figure 8. Vanguard Group Inc is the largest investor with $7.3 trillion, followed by BlackRock with $3.7 trillion and State Street Corporation with $2.6 trillion.

Lund and Robertson (2023) refer to the Big Three as passive investors. Figure 9 illustrates the type of investment that each of the Big Three has in the S&P 500 companies. Direct Holder refers to a company investing in a particular asset as an active investor, whereas Mutual Funds is related to a passive investment strategy. Contrary to the author's statement, data shows that the largest assets managers in the world prefer active investing over passive investing. The asset manager with the highest proportion of its portfolio invested using active investment is BlackRock with 88.72 %, followed by State Street Corporation with 75.06 % and lastly Vanguard Group Inc with 55.29 %. Based on the data, Lund and Robertson's statement on big asset managers and passive investment lacks support.

**Figure 8**

*Top 5 holders in S&P 500*

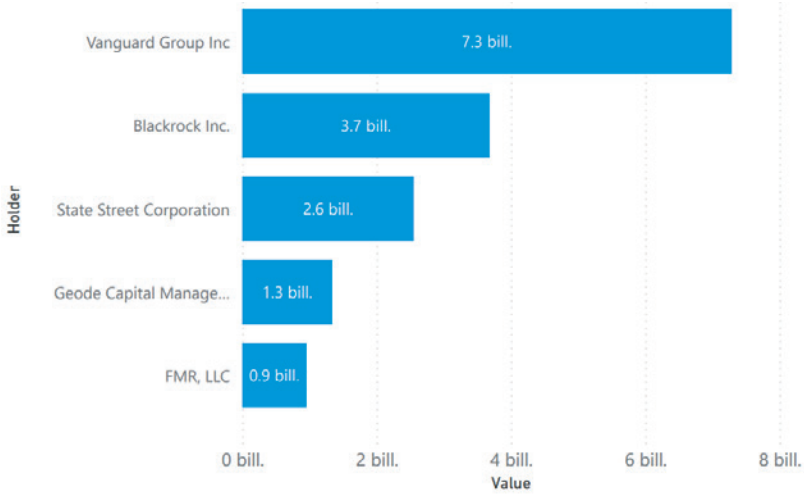
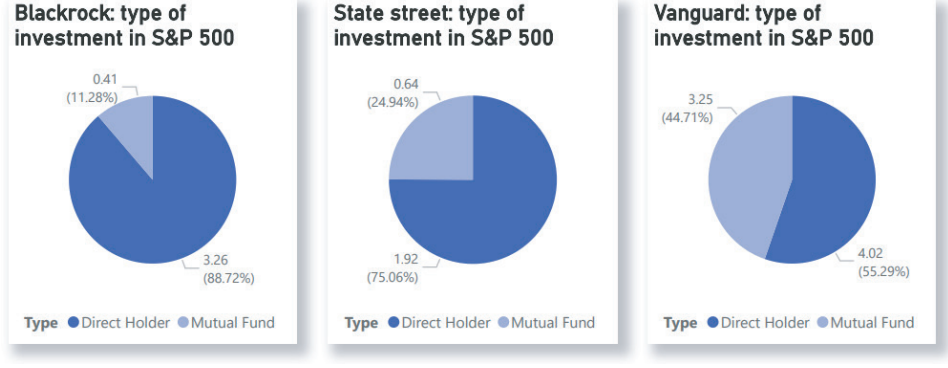



**Figure 9**

*Type of investment developed by the Big Three in S&P 500 companies*




Phillips's (2018) research pointed out that the global power elite controls global corporate media, such as Comcast Corp. or Disney. Figure 10 lists the biggest media corporations and calculates how much of them is owned by the Big Three. It is clear that the Big Three hold important positions in companies like Comcast Corp. (CMCSA: 29.12 %), Disney (DIS: 28.73 %), Warner Bros Discovery (WBD: 31.65 %), and Omnicom Group Inc. (OMC: 40.87 %).

**Figure 10**

*Ownership of the Big Three on Big Media companies*

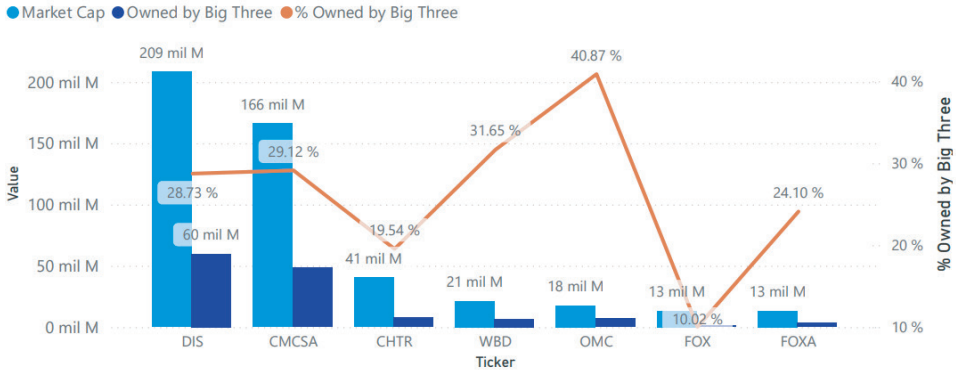



Table 6 not only shows individual investment in big media corporations, but also the total value invested by the Big Three as a whole. The Top 3 asset managers possess 28.73 % of all big media firms, worth $135 billion.

**Table 6**

*Total ownership of the Big Three in big media companies*

| Ticker | Market Cap ▼ | Owned by Big Three | % Owned by Big Three |
|---|---|---|---|
| DIS | 208.54 bn | 60 bn | 28.73 % |
| CMCSA | 166.22 bn | 48 bn | 29.12 % |
| CHTR | 40.53 bn | 8 bn | 19.54 % |
| WBD | 21.15 bn | 7 bn | 31.65 % |
| OMC | 17.66 bn | 7 bn | 40.87 % |
| FOX | 13.40 bn | 1 bn | 10.02 % |
| FOXA | 13.39 bn | 3 bn | 24.10 % |
| **Total** | **480.89 bn** | **135 bn** | **28.01 %** |

## 5. CONCLUSION

This article delves into insights provided by key sources such as Phillips (2018), Lund and Robertson (2023) and McLaughlin and Massa (2020) regarding the most prominent asset managers, often referred to as 'Giants' or the 'Big Three'. Distilling information from these references, three hypotheses were formulated and scrutinized through hypothesis testing, as detailed in Table 3. The results of the data analytics process unfold as follows: Hypothesis #1, positing that the Big Three collectively own an average of 7.33 % of

companies within the S&P 500, was supported. Conversely, the hypothesis #2 suggesting that the Big Three control over 90 % of the portfolio in passive investment was rejected. Meanwhile, the hypothesis #3 asserting that the 'Global Power Elite' commands approximately 15.56 % of the largest media corporations did not face rejection.

As of March 2024, the Big Three command formidable influence, holding 28.31 % of the total market value of S&P 500, 24.54 % of NASDAQ 100, and 28.24 % of DJIA. This substantial ownership underscores the undeniable level of power and influence they wield to advance their interests.

## REFERENCES

Arora, Y., & Goyal, D. (2016). Big data: a review of analytics methods & techniques. *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pp. 225-230. doi: 10.1109/IC3I.2016.7917965.

Canadian Network of Asset Managers. (2018). *Asset Management 101. The what, why, and how for your community*. https://www.assetmanagementbc.ca/wp-content/uploads/Asset-Management-101-The-What-Why-and-How-for-Your-Community-CNAM.pdf

Collin, V. (2023, September 4). *What is asset management*. Financial Edge. https://www.fe.training/free-resources/asset-management/what-is-asset-management/

Dellatto, M. (2023, June 8). *The world's largest media companies in 2023: Comcast and Disney stay on top*. Forbes. https://www.forbes.com/sites/marisadellatto/2023/06/08/the-worlds-largest-media-companies-in-2023-comcast-and-disney-stay-on-top/?sh=1b990b4654c6/

Duan, L., & Da Xu, L. (2021). Data analytics in industry 4.0: a survey. *Information Systems Frontiers: A Journal of Research and Innovation*. https://doi.org/10.1007/s10796-021-10190-0

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing*, *74*(7), 2561–2573. https://doi.org/10.1016/j.jpdc.2014.01.003

Kelley, K. (2020, 27 de mayo). What is data analysis? Pprocess, types, methods and techniques. Simplilearn. https://www.simplilearn.com.cach3.com/data-analysis-methods-process-types-article.html

Lund, D. S., & Robertson, A. (2023). Giant asset managers, the big three, and index investing. *USC CLASS Research Paper* (23-13). https://doi.org/10.2139/ssrn.4406204

March, T. (2020, January 10). *4 types of data analytics for educators*. Tom March.com. https://tommarch.com/2020/01/4-types-data-analytics-for-educators/

Mathur, G. (2023, September 19). *Data science vs data analytics: unpacking the differences*. IBM Blog. https://www.ibm.com/blog/data-science-vs-data-analytics-unpacking-the-differences/

McLauglin, D., & Massa, A. (2020). *The hidden dangers of the great index fund takeover*. Bloomberg. https://www.bloomberg.com/news/features/2020-01-09/the-hidden-dangers-of-the-great-index-fund-takeover

Neufeld, D. (2023, October 18). *Mapped: the largest stock exchanges in the world*. Advisor Channel. https://advisor.visualcapitalist.com/largest-stock-exchanges-in-the-world/

Oracle. (s/f). *What is data analytics?* Oracle.com. https://www.oracle.com/business-analytics/data-analytics/

Peck, R., Olsen, C., & DeVore, J. L. (2008). *Introduction to statistics and data analysis* (3a ed.). Wadsworth Publishing.

Phillips, P. (2018). *Giants: the global power elite*. Seven Stories Press.

Vanani, I. R., & Majidian, S. (2019). Literature review on big data analytics methods. In: A. Cano (Ed). *Social Media and Machine Learning*. IntechOpen. https://doi.org.10.5772/intechopen.86843

Smart, S. B., Gitman, L. J., & Joehnk, M. D. (2017). *Fundamentals of investing* (13a ed.). Pearson.

Srivastava, G., Muneeswari, S.,Venkataraman, R., Kavitha, V. & Parthiban, N. (2022). A review of the state of the art in business intelligence software. *Enterprise Information Systems*, *16*(1), 1–28. https://doi.org/10.1080/17517575.2021.1872107

Stevens, E. (2023, May 10). The 7 most useful data analysis methods and techniques. *CareerFoundry*. https://careerfoundry.com/en/blog/data-analytics/data-analysis-techniques/

Stock Analysis (2023a). *Free online stock information for investors*. Stock Analysis. Recuperado el 20 de mayo de 2024. https://stockanalysis.com/

Stock Analysis (2023b). *S&P 500 index stocks list*. https://stockanalysis.com/list/sp-500-stocks/

Stock Analysis (2023c). *NASDAQ 100 index stocks list*. https://stockanalysis.com/list/nasdaq-100-stocks/

Stock Analysis (2023d). *Dow Jones Industrial Average Stocks List*. https://stockanalysis.com/list/dow-jones-stocks/

Taherdoost, H. (2022). Different types of data analysis; data analysis methods and techniques in research projects. *International Journal of Academic Research in Management, 9*(1),1-9. https://ssrn.com/abstract=4178680

Tripathi, A., & Bagga, T. (2020). Leading business intelligence (BI) solutions and market trends. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3568414

U.S. Securities and Exchange Commission. (2021, April 2). *Investor bulletin: characteristics of mutual funds and exchange-traded funds (ETFs)*. Investor.gov. https://www.investor.gov/introduction-investing/general-resources/news-alerts/alerts-bulletins/characteristics-mutual-funds-exchange-traded-funds/

Vashisht, P., & Gupta, V. (2015). Big data analytics techniques: a survey. *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, 264–269. doi: 10.1109/ICGCIoT.2015.7380470

Yahoo Finance. (2023). Matching Mutual Funds. Recuperado el 20 de mayo de 2024, de https://finance.yahoo.com/

# Una revisión sistemática de literatura sobre implementaciones de sistemas de control de tráfico

Eduardo Rodrigo Wong Leon
2019100646@ucss.pe
https://orcid.org/0009-0000-6544-9114
Universidad Católica Sedes Sapientiae, Perú

Marco Antonio Coral Ygnacio
mcoral@ucss.edu.pe
https://orcid.org/0000-0001-6628-1528
Universidad Católica Sedes Sapientiae, Perú

Recibido: 30 de noviembre del 2023 / Aceptado: 14 de febrero del 2024
doi: https://doi.org/10.26439/interfases2024.n19.6779

RESUMEN. La congestión vehicular es una problemática que se manifiesta frecuentemente en ciudades con alta población y puede deberse a diversos factores, como la incorrecta planificación civil o el transporte público deficiente. Esto provoca un incremento en los accidentes de tránsito, la contaminación del aire, la pérdida de combustible y el descontento ciudadano. Por ello, se considera importante la implementación de sistemas de control de tráfico que genere fluidez en el tránsito vehicular y reduzca los tiempos de viaje. Este trabajo desarrolla una revisión sistemática de la literatura con el propósito de identificar los métodos, algoritmos y modelos más eficientes para la construcción de un sistema de control de tráfico. Los resultados identifican tres métodos y tres algoritmos considerados muy eficientes para el desarrollo de estos sistemas, de los cuales se resaltan el filtro bayesiano y las redes neuronales convolucionales. También se demuestra que You Only Look Once, conocido como YOLO, es el modelo de procesamiento de imagen más eficiente para estas implementaciones.

PALABRAS CLAVE: control de tráfico / métodos / algoritmos / modelos / YOLO / implementaciones

## A SYSTEMATIC LITERATURE REVIEW OF TRAFFIC CONTROL SYSTEM IMPLEMENTATIONS

ABSTRACT. Traffic congestion frequently occurs in highly populated cities and can result from poor civil planning or inadequate public transportation. This issue increases traffic accidents, air pollution, fuel loss, and public dissatisfaction. Therefore,

implementing traffic control systems that improve traffic flow and reduce travel times becomes essential. This work conducts a systematic literature review to identify the most efficient methods, algorithms, and models for developing traffic control systems. The review identifies three methods and three algorithms that are highly efficient for these systems, highlighting Bayesian filters and convolutional neural networks. It also shows that You Only Look Once (YOLO) is the most efficient image processing model for these implementations.

KEYWORDS: traffic control / methods / algorithms / models / YOLO / implementations

## 1. INTRODUCCIÓN

La congestión vehicular se manifiesta muchas veces en ciudades en desarrollo, las cuales cuentan con poco avance o inversión en el control de tránsito (Kumaran et al., 2019). Durante las horas pico de la mañana y la noche, se convierte en un problema crónico que, si no es solucionado, seguirá afectando enormemente no solo a la eficiencia del transporte ciudadano, sino también a la calidad de vida de sus residentes.

Durante los últimos años, se han realizado esfuerzos para abordar la congestión de tráfico con nuevos métodos y tecnologías, donde el uso de la inteligencia artificial (IA) ha sido la preferida (Chabchoub et al., 2021). En esta línea, el procesamiento de imágenes utiliza dispositivos inteligentes como cámaras de alta resolución. Asimismo, existen propuestas basadas en metodologías de optimización a fin de predecir el comportamiento del tráfico para su control (Cheng et al., 2023).

La congestión del tráfico es un problema de gran relevancia debido a sus múltiples repercusiones, tales como el incremento de accidentes de tránsito, la contaminación del aire, el descontento de la ciudadanía, entre otros. La importancia del tema radica en la necesidad de optimizar la gestión del tráfico y reducir los problemas asociados, así como mejorar la calidad de vida de los residentes y fortalecer la eficiencia de la administración municipal (Joo & Lim, 2021).

En los años recientes se ha propuesto la aplicación de metodologías y modelos basados en IA como las redes convolucionales (Vélez-Serrano et al., 2021) o sistemas multiagentes (Wakkumbura et al., 2021). De este modo, se brindan soluciones precisas y eficientes, pero que en ocasiones no son adaptables a más de un contexto.

A pesar de la evolución y uso de tecnologías basadas en IA, el control de tráfico sigue teniendo deficiencias en ciudades con mucha aglomeración de personas y que carecen de la incorporación de tecnología de vanguardia en sus instrumentos de control de tránsito (Liu et al., 2020). La motivación del trabajo se centra en detectar propuestas viables para el control de tráfico en ciudades grandes y que aproveche las herramientas ya incorporadas, sin requerir un cambio masivo en estas.

Se propone una revisión sistemática de literatura para identificar las tecnologías, métodos y modelos utilizados en la construcción de sistemas de control de tráfico y cuáles de estos resultan más viables y eficientes. Esto se realiza con el objetivo de recopilar soluciones para reducir la congestión vehicular a través de un sistema de control de semáforos.

Esta revisión sistemática de literatura sigue la metodología de Kitchenham (2009), que utiliza su ciclo de recopilación y selección de bibliografía para obtener artículos relevantes al campo de estudio de investigación y conseguir una conclusión precisa.

El presente artículo se organiza en cinco capítulos. En el primero, se encuentra la introducción del artículo; el segundo detalla el estado del arte del tema de investigación;

el tercero describe la metodología y el proceso de revisión de literatura; el cuarto muestra los resultados de la investigación, y las conclusiones se detallan en el quinto capítulo.

## 2. ESTADO DEL ARTE

Las soluciones de control de tráfico muestran una transición hacia soluciones más avanzadas y adaptativas, impulsadas en gran medida por la aplicación de la IA y la tecnología (Wang et al., 2023). Estas tecnologías están diseñadas para abordar los desafíos de la congestión vehicular en ciudades grandes, optimizar la gestión del tráfico y mejorar la calidad de vida de los residentes (Chabchoub et al., 2021).

### Control de tráfico

El control de tráfico es un campo de estudio amplio y crucial en la gestión del transporte y la movilidad urbana que implica administrar los componentes viales de tránsito vehicular. De esta forma, se enfoca en lograr la optimización y evitar problemáticas de congestión y accidentes vehiculares (Li, Y. et al., 2021).

Se han desarrollado diversas estrategias a lo largo de los años para abordar la congestión vehicular y mejorar la eficiencia del tráfico. Tradicionalmente, el control de tráfico se ha basado en la programación de tiempos fijos para semáforos y el uso de sensores de tráfico en las intersecciones. Sin embargo, durante estos últimos años, con el avance de la tecnología y la incorporación de la IA, se han explorado enfoques más dinámicos y adaptativos que puedan mejorar la eficiencia de tránsito en las ciudades con aglomeración de vehículos. Por ejemplo, ciertas propuestas son el uso de redes neuronales (Shin et al., 2019) o modelos de predicción (Bao et al., 2023) y control de semáforos (Tunc & Soylemez, 2023) para la generación de "olas verdes".

### 2.1 La tecnología en el control de tráfico

La tecnología desempeña un papel fundamental en el control de tráfico (Rasheed et al., 2022), lo que ha permitido una mayor flexibilidad en la gestión del flujo vehicular. Pueden incluir sensores de tráfico avanzados, cámaras de vigilancia y sistemas de semaforización inteligente. Los sensores de vanguardia utilizan IA y aprendizaje profundo, como redes convolucionales neuronales (CNN) y redes neuronales recurrentes, que les permiten registrar nuevos datos en tiempo real y mejorar su propia base de datos, lo que aumenta su precisión. Las cámaras de vigilancia utilizadas para el control de tráfico normalmente incluyen procesamiento local, como las NVIDIA DeepStream, que permite incorporar modelos inteligentes, como YOLO (You Only Look Once). Estas tecnologías se enfocan en conectarse a un sistema general para utilizar los datos recopilados en tiempo real con el objetivo de una toma de decisiones más profunda y de mayor precisión para optimizar el tráfico (Phursule et al., 2023). No obstante, con la llegada de la IA, cada vez se requiere menos del factor humano y la toma de decisiones es delegada a agentes

inteligentes que han recibido entrenamiento o aprenden en tiempo real, lo que reduce el trabajo del personal de monitoreo y mitiga la posibilidad de error y demora humana.

En países desarrollados, se ha creado el concepto *smart cities* (Aqib et al., 2019), que incorpora de manera integral la IA para acoplar la gestión de tráfico a otros ámbitos, como la comunicación entre vehículos privados, infraestructuras y peatones conectados con el internet de las cosas (Chahal et al., 2023).

### 2.2 Aplicación de la IA en el control de tráfico

La IA ha revolucionado el campo del control de tráfico al proporcionar herramientas y técnicas avanzadas para la toma de decisiones y la optimización en tiempo real. Algunos campos clave de la IA relacionados con el control de tráfico se detallarán a continuación.

#### *2.2.1 Procesamiento de imagen*

El procesamiento de imágenes desempeña un papel crucial en la supervisión y el control del tráfico. Se utilizan cámaras de alta resolución para capturar imágenes de las intersecciones y carreteras en tiempo real, y luego se aplican algoritmos de visión por computadora para analizar el flujo de tráfico, detectar vehículos, peatones y condiciones de la carretera. Esto proporciona datos visuales esenciales que permitirán el entrenamiento y posterior toma de decisiones en tiempo real de los modelos inteligentes de control de semaforización para la optimización del flujo de tráfico (Shin et al., 2019). Aunque en ocasiones puede ser impreciso debido a eventos inusuales o específicos que generalmente no suceden, cada año se proponen mejoras de precisión con el uso de nuevos modelos de IA (Chabchoub et al., 2021).

#### *2.2.2 Predicción de tráfico*

La predicción de tráfico implica el uso de modelos de IA para anticipar el comportamiento del tránsito en un área determinada. Estos modelos reciben datos históricos de tráfico, condiciones meteorológicas, eventos especiales y otros factores relevantes a través de entradas que, para el tema de tránsito, generalmente provienen de sensores y videocámaras (Chahal et al., 2023).

Aunque principalmente se utiliza el aprendizaje guiado y supervisado para entrenar los modelos de predicción (Aqib et al., 2019), se han desarrollado propuestas para permitir que estos modelos aprendan mientras están en funcionamiento y utilizan datos en tiempo real (Hao, W. et al., 2020).

#### *2.2.3 Semaforización inteligente*

Los sistemas de semaforización inteligente utilizan algoritmos y modelos de IA para ajustar los tiempos de los semáforos de manera dinámica en función de las condiciones

actuales del tráfico (Rasheed et al., 2022). Esto permite una mejor adaptación a las necesidades del tráfico en tiempo real, lo que reduce la congestión y mejora la fluidez. Estos sistemas suelen integrarse con datos de sensores y cámaras para tomar decisiones informadas, incluso utilizan modelos de predicción que incrementen su nivel de precisión para generar olas verdes (Hao, S. et al., 2019).

## 3. METODOLOGÍA

A través de la revisión sistemática de literatura, se pueden sintetizar conocimientos, descubrimientos, métodos y tecnologías utilizadas en las implementaciones de sistemas de control de tráfico. En esta investigación, se toma de guía el marco metodológico propuesto por Kitchenham (2009), que consta de tres etapas: planificación, realización e informe de la revisión.

### 3.1 Planificación de la revisión

Para la fase de planificación, se define el tema de investigación con el propósito de determinar el alcance de la revisión. Luego, se determinan preguntas de investigación referentes al tema seleccionado. Estas se muestran a continuación.

- P1: ¿Qué algoritmos y métodos se utilizan en la construcción de sistemas de control de semaforización?
- P2: ¿Qué modelos de reconocimiento de objetos por imagen se utilizan para la identificación de densidad vehicular?
- P3: ¿Qué patrones se deben considerar para la identificación de vehículos por imagen?
- P4: ¿Qué algoritmos y métodos son los más eficientes para construir un sistema sobre la infraestructura de tráfico existente?

### 3.2 Realización de la revisión

En esta etapa, se detalla el procedimiento de revisión de literatura, así como se explica la fase de búsqueda y selección de bibliografía.

#### *3.2.1 Estrategias de búsqueda*

Para la búsqueda de fuentes bibliográficas en la base de datos Scopus se utilizan palabras clave referentes a las preguntas de investigación del tema de control de tráfico. Con estas se formula la cadena de búsqueda: (“traffic signal control” OR “smart traffic light” OR “traffic light control”) AND (“vehicle counting” OR “vehicle detection” OR sensors) AND (“image recognition system” OR “image processing” OR software OR algorithms).

Se determinan criterios que mejoren la calidad de los resultados, como que estos no tengan una antigüedad mayor a los 5 años o que cuenten con una versión en inglés. También se establece que los artículos deben contar con el Identificador de Objeto Digital (DOI) para garantizar su confiabilidad. Es esencial que los artículos seleccionados se centren en el tema de control de tráfico con propuestas de implementación o aplicación para obtener resultados en contextos reales y una conclusión precisa. Los criterios a tomar en cuenta para la búsqueda y posterior selección de artículos son los que se muestran a continuación en la Tabla 1.

**Tabla 1**

*Criterios de inclusión y exclusión*

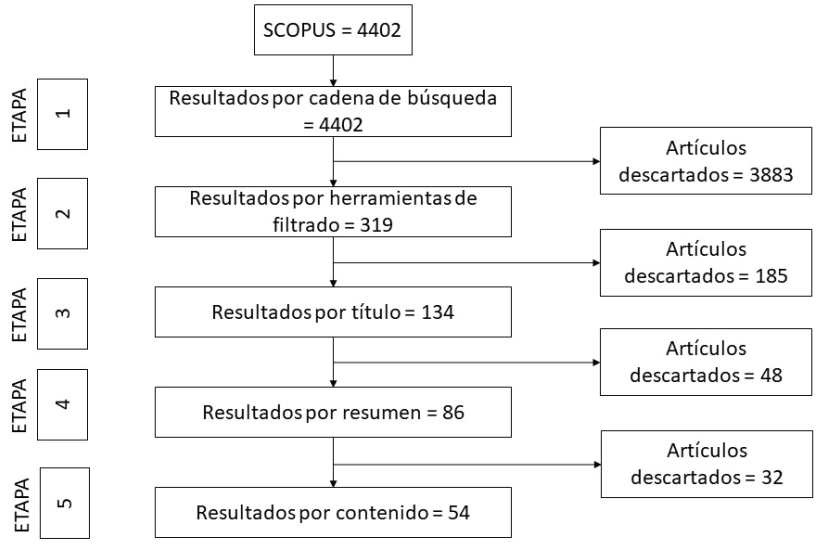
| Criterios de inclusión | Criterios de exclusión |
|---|---|
| - Artículos centrados en la investigación del tema control de semáforos | - Investigación descriptiva<br>- Sin implementación |
| - Artículos con DOI<br>- Publicación en inglés | - Artículos que no se relacionen con el tema de control de semáforos |
| - Artículos centrados en técnicas, metodologías, tecnologías o herramientas para el control de semáforos | - Informes, revistas, boletines, comentarios<br>- Antigüedad mayor a 5 años |

### *3.2.2 Selección de estudios*

A partir de la cadena de búsqueda, durante la primera etapa del cribado, se seleccionaron 4402 artículos. La segunda etapa se realizó utilizando las herramientas disponibles por la base de datos de Scopus y considerando los criterios definidos en la Tabla 1, donde se optaron por 319 artículos. Para la tercera etapa, se revisaron títulos referentes al tema de control de tráfico, donde quedaron 134 artículos. Luego, se realizó una revisión del resumen para asegurar la relevancia y utilidad de los artículos, que resultó así en 86 artículos seleccionados. Finalmente, se efectuó la lectura del texto completo del artículo, donde se eligieron 54 de ellos. A continuación, se resume el proceso de selección y cribado en la Figura 1.

**Figura 1**

*Diagrama de flujo de cribado*




*3.2.3 Extracción y análisis de datos*

Se realiza un análisis cuantitativo representado en gráficos estadísticos a partir de los datos extraídos de la realización de la revisión. En el siguiente enlace se puede encontrar el procedimiento de extracción y análisis de datos con mayor detalle: https://docs.google.com/spreadsheets/d/1rVCe81Os0tMwbYX9mGeQvW6oplaoMelK/edit?usp=drive_link&ouid=109493883837738690362&rtpof=true&sd=true

La Figura 2 muestra las publicaciones por año. A partir de ella es posible determinar que existe un interés creciente desde el 2021 en el tema del control inteligente de tráfico. Esto sugiere que existe una demanda global de soluciones para esta problemática, excepto por el año 2020 que tuvo una caída drástica en la publicación de artículos, posiblemente debido a la pandemia del COVID-19. La creciente demanda de soluciones a la problemática de control de tráfico con semaforización inteligente podría crear oportunidades para la colaboración internacional y el desarrollo de nuevas tecnologías que sean aplicables a una variedad de contextos.

La Figura 3 muestra la cantidad de artículos publicados por revistas especializadas. Se puede notar que existe una mayor cantidad de publicaciones alrededor del tema de control inteligente de tráfico en las revistas *Sensors* (7), *Applied Sciences* (6) y *Journal of Advanced Transportation* (4). Esto nos indica que en estos espacios académicos es posible encontrar información del tema de investigación en mayor cantidad y con un mayor desarrollo. Por lo tanto, esto nos permite formar un juicio previo del artículo al momento de seleccionar uno para cerciorar un punto o afirmación referente al tema de investigación.

**Figura 2**

*Artículos publicados por año*

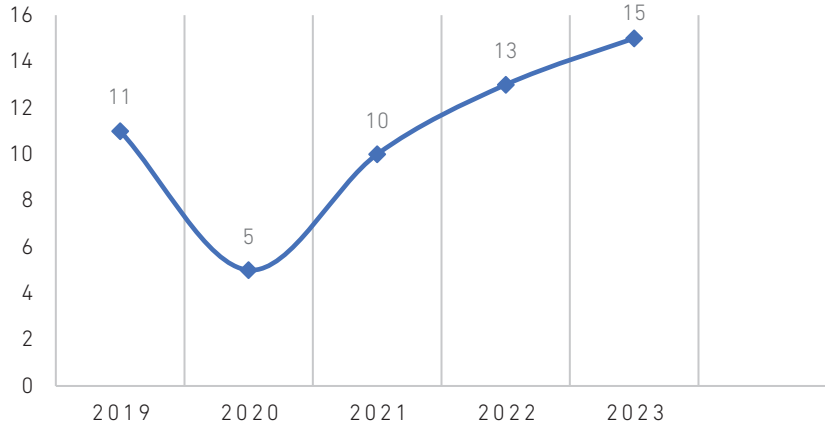
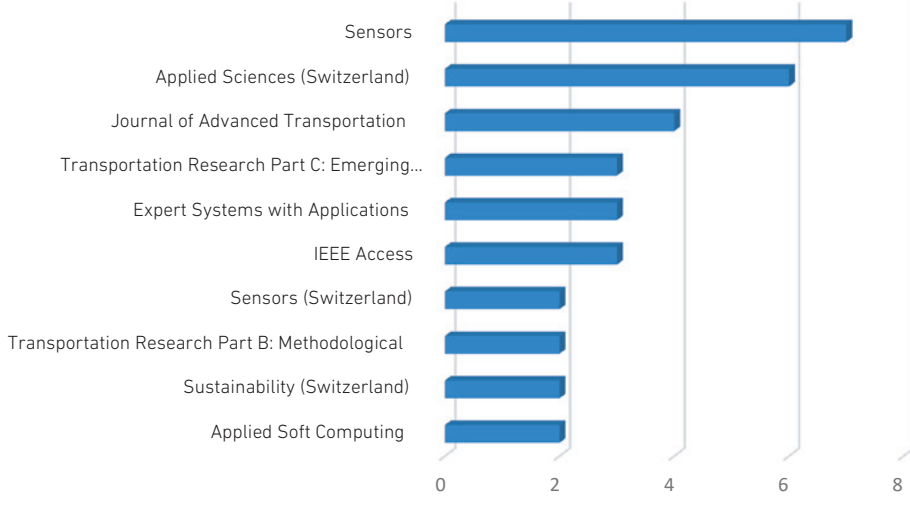



**Figura 3**

*Artículos publicados por una revista especializada*

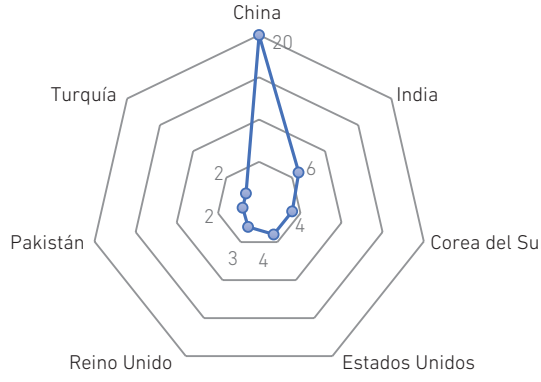



En la Figura 4 se muestran los países con más artículos publicados, en donde China (20) e India (6) son los que cuentan con una mayor cantidad de estas investigaciones. Del gráfico, se han excluido 11 países, pues solo cuentan con una sola publicación. Se determina que la problemática e investigación de soluciones del tema de control de tráfico abunda mayoritariamente en países orientales como China, debido a la sobrepoblación

en sus principales ciudades. Esto nos indica que los artículos pertenecientes a estos países desarrollan un enfoque más amplio y cercano a la problemática, además de realizarlo en una cantidad mayor que en países occidentales.

**Figura 4**

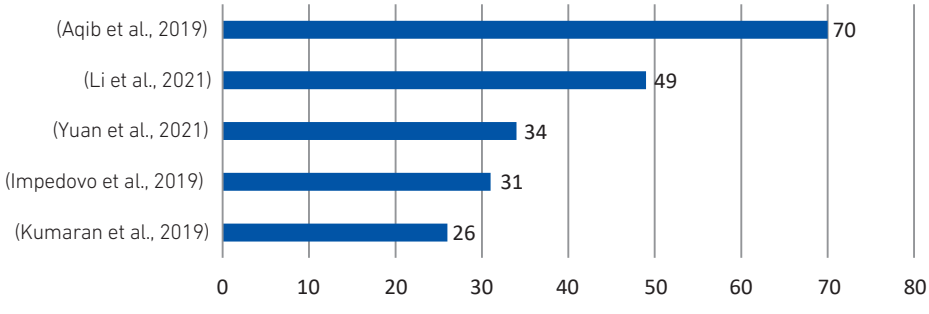
*Artículos publicados por país*




La Figura 5 expone los cinco artículos con mayor cantidad de citaciones en otros trabajos. El más citado tiene 70; el segundo lugar, 49, y el tercero, 34. Todos son citados al menos una vez por algún autor. Siete trabajos cuentan con solo una citación. Se determina a través del gráfico que existen múltiples citaciones, lo cual demuestra el interés en los temas tratados.

**Figura 5**

*Artículos con mayor cantidad de citaciones*

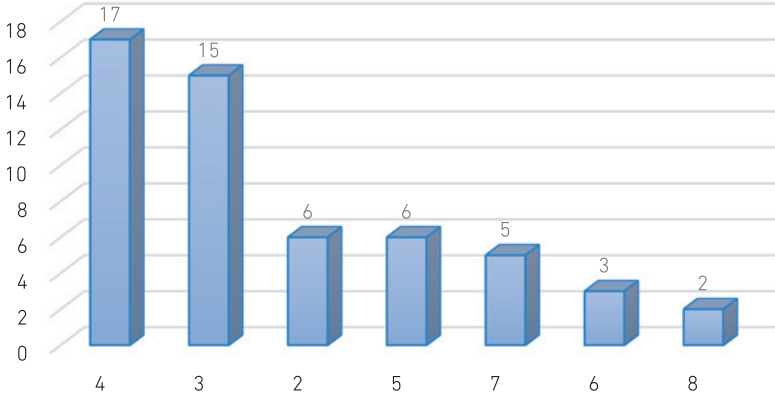



La Figura 6 presenta el número de artículos publicados según la cantidad de autores. Se evidencian 17 publicaciones con cuatro autores, 15 con tres autores y también 2 trabajos con ocho autores cada uno. Ningún trabajo cuenta con solo un autor. De esta forma, se manifiesta que alrededor del tema de control inteligente de tráfico existe una

mayor colaboración entre varios investigadores. Esto podría indicar que una solución inteligente para el problema de control de tráfico resulta lo suficientemente compleja como para necesitar varios especialistas de diversos enfoques y áreas.

**Figura 6**

*Cantidad de autores por artículos publicado*




### 3.3 Informe de la revisión

En esta etapa, se recopilan los resultados y respuestas a las preguntas de investigación definidas en la primera etapa obtenidas tras la revisión de literatura. A continuación, se responden a las cuatro preguntas.

**P1: ¿Qué algoritmos y métodos se utilizan en la construcción de sistemas de control de semaforización?**

Un algoritmo es un grupo de instrucciones organizadas y secuenciadas que en conjunto sirven para un propósito o función. En programación, la principal diferencia entre algoritmo y método es que este último es una implementación específica y concreta del primero. En la Tabla 3, se recopilan y describen los algoritmos y métodos utilizados por los autores en sus propuestas de implementación. Solo se rescatan aquellos de uso principal que son señalados y detallados en el proceso de construcción del sistema de los artículos revisados.

**Tabla 3**

*Algoritmos y métodos utilizados en sistemas de control de semaforización*

| REFERENCIA | ALGORITMO | MÉTODO | DESCRIPCIÓN |
|---|---|---|---|
| Yuan et al. (2021); Jin y Ma (2019) | | Bayesian Filter | Es un método basado en un filtro recursivo que se utiliza para estimar el estado de un sistema dinámico a partir de una secuencia de mediciones ruidosas y medir su incertidumbre. |
| Stoilova y Stoilov (2022); Chahal et al. (2023) | Bi-level Algorithm | | Es un tipo de algoritmo de optimización que se utiliza para resolver problemas con dos niveles de toma de decisiones. |
| Shin et al. (2019); Vélez-Serrano et al. (2021); Islam et al. (2022) | Convolutional Neural Network (CNN) | | Es un tipo de algoritmo de aprendizaje profundo que es particularmente adecuado para la clasificación y reconocimiento de imágenes. |
| Tan et al. (2022); Aqib et al. (2019); Bao et al. (2023); Islam et al. (2022); Impedovo et al. (2019) | | Deep Learning | Es un método de aprendizaje automático que utiliza redes neuronales artificiales para aprender de los datos procesados. |
| Mukhtar et al. (2023); Rasheed et al. (2022); Joo y Lim (2021); Li, Z. et al. (2021); Zheng et al. (2022) | Deep Q-Network (DQN) | | Es un algoritmo de aprendizaje por refuerzo que utiliza el aprendizaje profundo para aprender una política. DQN ha demostrado que es efectivo para resolver una variedad de problemas, como realizar predicciones y controlar mecanismos inteligentes. |
| Damadam et al. (2022); Rasheed et al. (2022); Tan et al. (2022); Wu et al. (2022); Wang et al. (2023); Li, Z. et al. (2021); Szoke et al. (2023); Li, W. et al. (2021) | | Deep Reinforcement Learning | Es un método de aprendizaje por refuerzo que utiliza el aprendizaje profundo para aprender una política. |
| Tunc y Soylemez (2023); Chabchoub et al. (2021); Chatterjee et al. (2019); Wakkumbura et al. (2021) | | Fuzzy Logic Control (FLC) | Es un tipo de sistema de control que utiliza lógica borrosa para tomar decisiones. FLC se utiliza para controlar sistemas complejos que son difíciles de modelar utilizando métodos de control tradicionales. |
| Yuan et al. (2021); Jin y Ma (2019) | | Gaussian Process | Es un método estadístico que se puede utilizar para la regresión, clasificación y predicción. |
| Kim et al. (2023) | Genetic Algorithm | | Es un algoritmo metaheurístico que está inspirado en el proceso de selección natural y es utilizado para resolver problemas de optimización. |

(*continúa*)

(*continuación*)

| REFERENCIA | ALGORITMO | MÉTODO | DESCRIPCIÓN |
|---|---|---|---|
| Mukhtar et al. (2023) | Graph Convolutional Network (GCN) | | Es un tipo de algoritmo de aprendizaje profundo que se puede utilizar para aprender en base de datos representados como grafos. |
| Shin et al. (2019); Chabchoub et al. (2021); Wakkumbura et al. (2021); Zhang et al. (2020); Li, Y. et al. (2021) | | Image Processing | Es un método que utiliza un conjunto de técnicas para manipular y analizar imágenes digitales. |
| Bao et al. (2023); Cheng et al. (2023); Hao, W. et al. (2020); Islam et al. (2022); Chahal et al. (2023) | | Long Short-Term Memory (LSTM) | Es un método que utiliza redes neuronales recurrentes para el procesamiento de datos secuenciales, como el procesamiento del lenguaje natural y la traducción automática. |
| Damadam et al. (2022); Korecki y Helbing (2022); Kumaran et al. (2019); Yuan et al. (2021); Wakkumbura et al. (2021); Szoke et al. (2023); Impedovo et al. (2019) | | Machine Learning | Es un método que le da a las computadoras la capacidad de aprender sin ser programadas explícitamente, se puede utilizar para resolver una amplia gama de problemas, como clasificación, regresión y predicción. |
| Wakkumbura et al. (2021); Suga et al. (2023); Mukhtar et al. (2023); Rasheed et al. (2022); Shehu et al. (2020); Li, Z. et al. (2021) | | Multi Agent Technology | Es un método que se ocupa del diseño y desarrollo de sistemas que consisten en el uso de múltiples agentes inteligentes para resolver problemas complejos que no pueden ser resueltos por un solo agente. |
| Cheng et al. (2023) | Snake Optimization Algorithm | | Es un tipo de algoritmo metaheurístico que está inspirado en el movimiento de las serpientes y es efectivo para optimizar el procesamiento de información. |

**P2: ¿Qué modelos de reconocimiento de objetos por imagen se utilizan para la identificación de densidad vehicular?**

Los modelos inteligentes son pequeños sistemas capaces de tomar decisiones y adaptarse a través del conocimiento e información adquiridos antes y durante su funcionamiento. La Tabla 4 muestra los modelos de procesamiento de imagen para reconocimiento de objetos utilizados en la construcción de sistemas de control de semaforización propuestos en los artículos revisados. En cada modelo se describe su uso en el proceso y las características que lo diferencian.

**Tabla 4**

*Modelos de reconocimiento de objetos utilizados en identificación de densidad vehicular*

| REFERENCIA | MODELO | DESCRIPCIÓN |
|---|---|---|
| Shin et al. (2019); Chabchoub et al. (2021); Wakkumbura et al. (2021); Zhang et al. (2020); Li, Y. et al. (2021) | You Only Look Once (YOLO) | Es un modelo detector de objetos de una sola pasada que identifica y clasifica objetos en una imagen en un solo paso. Puede predecir todas las cajas delimitadoras, clases y probabilidades en una sola pasada. |
| Chabchoub et al. (2021); Wakkumbura et al. (2021) | Single Shot MultiBox Detector (SSD) | Es un modelo de reconocimiento de objetos de detección rápida que identifica y clasifica objetos en una imagen en un solo paso. Es similar a YOLO, pero dado que utiliza cajas delimitadoras predefinidas puede ser más precisa, aunque también más lenta. |
| Li, Y. et al. (2021) | Faster R-CNN | Es un modelo detector de objetos de dos pasos que primero genera un conjunto de regiones candidatas y luego clasifica cada región candidata. Esto puede ser más preciso que YOLO y SSD, pero es menos eficiente. |
| Li, Y. et al. (2021) | Mask R-CNN | Es una extensión de Faster R-CNN que también predice máscaras para cada objeto detectado. Esto permite al modelo segmentar los objetos detectados, lo que es útil para aplicaciones como el reconocimiento de objetos y la visión por computadora de la robótica. |

**P3: ¿Qué patrones se deben considerar para la identificación de vehículos por imagen?**

Los patrones son un conjunto de características recurrentes que pueden ser identificadas dentro de un grupo de datos. En la Tabla 5, se recopilan los aspectos y características que los autores señalan para el entrenamiento del modelo inteligente para procesamiento de imagen en sus propuestas. Se mencionan únicamente los patrones que contribuyen a la precisión del modelo para el reconocimiento de vehículos.

**Tabla 5**

*Patrones para la identificación de vehículos por procesamiento de imagen*

| REFERENCIA | PATRONES | DESCRIPCIÓN |
|---|---|---|
| Li, Y. et al. (2021); Wakkumbura et al. (2021) | Formas | La mayoría de carros siguen formas estandarizadas con las que se puede facilitar la identificación como del tipo *muscle* o sedán. |
| Shin et al. (2019); Li, Y. et al. (2021) | Dimensiones | Los autores señalan que, tras una correcta configuración del contexto de reconocimiento, se puede identificar vehículos con mayor facilidad gracias a sus dimensiones superiores a otros objetos o personas que se desplacen en la imagen. |
| Tunc y Soylemez (2023); Chabchoub et al. (2021); Li, Y. et al. (2021) | Color | Cambios abruptos en las tonalidades facilitan la identificación de los vehículos. |
| Li, Y. et al. (2021) | Placas de matrícula | El autor señala que la identificación de una placa vehicular dentro de una de las formas facilita su reconocimiento como vehículo. |

**P4: ¿Qué algoritmos y métodos son los más eficientes para construir un sistema sobre la infraestructura de tráfico existente?**

La Tabla 6 detalla un análisis general de eficiencia de los algoritmos y métodos utilizados en las propuestas de los autores, donde se observa que las características que estos consideran afectan de manera positiva o negativa en la precisión y confiabilidad de sus sistemas de control de semáforos. Se determina el nivel de eficiencia en tres estados: ineficiente, eficiente y muy eficiente. La calificación se basa en las observaciones y resultados de optimización de las colas de tráfico presentadas en los trabajos revisados. Si la reducción de longitud de cola es menor al 20 % y presenta dificultades que afectan a la precisión del sistema, se considera ineficiente. Si la reducción de longitud de cola es menor al 20 % y las dificultades no afectan a la precisión del sistema, o la reducción es mayor y las dificultades sí afectan, se considera eficiente. Si la reducción supera el 20 % y las dificultades son ajenas a la precisión del sistema, se considera muy eficiente. También se compara su eficiencia con el resto para incrementar la confiabilidad de la calificación. La Tabla 6 muestra la rúbrica establecida para la calificación del nivel de eficiencia.

**Tabla 6**

*Rúbrica de evaluación de nivel de eficiencia*

| | Reducción de longitud de cola menor a 20 % | Reducción de longitud de cola mayor a 20 % |
|---|---|---|
| Las dificultades presentadas afectan a la precisión del sistema. | Ineficiente | Eficiente |
| Las dificultades presentadas no afectan a la precisión del sistema. | Eficiente | Muy eficiente |

**Tabla 7**

*Nivel de eficiencia de algoritmos y métodos utilizados en sistemas de control de semaforización*

| REFERENCIA | ALGORITMO/ MÉTODO | NIVEL | OBSERVACIONES |
|---|---|---|---|
| Stoilova y Stoilov (2022); Chahal et al. (2023) | Bi-level Algorithm | Eficiente | Uno de los autores señala que su funcionamiento por dos niveles incrementa su precisión y utilidad, pero su capacidad de adaptación a problemas muy específicos es inferior en comparación a otros algoritmos. |
| Tunc y Soylemez (2023); Chabchoub et al. (2021); Chatterjee et al. (2019); Wakkumbura et al. (2021) | Fuzzy Logic Control (FLC) | Eficiente | Los autores hacen observaciones de su aceptable eficiencia en sistemas de semáforos tradicionales; no obstante, debido a esto, su adaptabilidad en situaciones muy complejas no es la mejor. |

(*continúa*)

(*continuación*)

| REFERENCIA | ALGORITMO/ MÉTODO | NIVEL | OBSERVACIONES |
|---|---|---|---|
| Yuan et al. (2021); Jin y Ma (2019) | Gaussian Process | Eficiente | Los autores lo comparan con la Fuzzy Logic en cuanto a su utilidad para sistemas de control de tráfico. |
| Bao et al. (2023); Cheng et al. (2023); Hao, S. et al. (2020); Islam et al. (2022); Chahal et al. (2023) | Long Short-Term Memory (LSTM) | Eficiente | Los autores señalan que es muy útil para predicción de tráfico y registro de información en tiempo real, pero en ocasiones puede tener problemas cuando los datos de tráfico son muy ruidosos. |
| Tan et al. (2022); Aqib et al. (2019); Bao et al. (2023); Islam et al. (2022); Impedovo et al. (2019) | Deep Learning | Ineficiente | Los autores señalan que por sí sola puede resultar poco eficiente debido a su baja precisión y las grandes cantidades de datos que requiere para el control de tráfico. No obstante, puede utilizarse como base de algoritmos más complejos. |
| Kim et al. (2023) | Genetic Algorithm | Ineficiente | El autor señala que, aunque puede resolver problemas complejos de control de tráfico, también puede resultar demasiado lento para su uso en tiempo real. |
| Cheng et al. (2023) | Snake Optimization Algorithm | Ineficiente | El autor señala que puede ralentizarse en intersecciones de más de dos carriles por dirección. |
| Yuan et al. (2021); Jin y Ma (2019) | Bayesian Filter | Muy eficiente | Los autores señalaron su gran eficiencia para la predicción de tráfico en diversas situaciones, ya que puede manejar grandes cantidades de datos en tiempo real. |
| Shin et al. (2019); Vélez-Serrano et al. (2021); Islam et al. (2022) | Convolutional Neural Network (CNN) | Muy eficiente | Los autores señalan que son esenciales para la creación de un modelo inteligente en control de tráfico. |
| Mukhtar et al. (2023); Rasheed et al. (2022); Joo y Lim (2021); Li, Z. et al. (2021); Zheng et al. (2022) | Deep Q-Network (DQN) | Muy eficiente | Los autores señalan que mientras cuente con buen entrenamiento aprender una política efectiva puede ser una de las mejores opciones para el control de tráfico debido a su procesamiento por redes. |
| Damadam et al. (2022); Rasheed et al. (2022); Tan et al. (2022); Wu et al. (2022); Wang et al. (2023); Li, Z. et al. (2021); Szoke et al. (2023); Li, Y. et al. (2021) | Deep Reinforcement Learning | Muy eficiente | Los autores señalan que es esencial para el aprendizaje recursivo y la mejora constante de la capacidad de toma de decisiones de un modelo inteligente en el sistema de control de tráfico. |
| Mukhtar et al. (2023) | Graph Convolutional Network (GCN) | Muy eficiente | Los autores señalan que es muy útil y eficiente para el procesamiento de datos estructurados en el control de tráfico debido a su capacidad de lectura de grafos. |

(*continúa*)

(*continuación*)

| REFERENCIA | ALGORITMO/ MÉTODO | NIVEL | OBSERVACIONES |
|---|---|---|---|
| Wakkumbura et al. (2021); Suga et al. (2023); Mukhtar et al. (2023); Rasheed et al. (2022); Shehu et al. (2020); Li, Z. et al. (2021) | Multi Agent Technology | Muy eficiente | Los autores señalan que, si se logra superar la complejidad de su implementación, puede llegar a resultar extremadamente eficiente y estable. |

## 4. DISCUSIÓN DE LOS RESULTADOS

Los resultados hallados a través del proceso de revisión de literatura describen la existencia de 6 algoritmos y 10 métodos utilizados para la construcción de sistemas de control de semáforos. Entre estos los más utilizados son el Deep Reinforcement Learning y el Machine Learning, pues sirven como base para la construcción de modelos inteligentes para el cambio de luces y predicción de tráfico. Los artículos también describen la importancia de las CNN y la construcción de una base de datos de entrenamiento amplias para incrementar la precisión de estos sistemas (Islam et al., 2022).

Entre los algoritmos y métodos utilizados, los de mayor eficiencia para la predicción de densidad vehicular y el cambio óptimo de luces son el filtro bayesiano, la tecnología multiagente (Mukhtar et al., 2023), las CNN y el aprendizaje profundo, que demuestran buena adaptabilidad durante horas de congestión de tráfico y eventos específicos como pase de vehículos de emergencia. Los resultados sugieren que, si estos se utilizan para la predicción de la densidad vehicular, se debe iterar varias ocasiones en estratos diferentes del día para evitar incorrectas adecuaciones (Jin & Ma, 2019). Por otro lado, se descartan aquellos que para situaciones de tránsito denso resultan lentos, poco precisos o de mala adaptación como los algoritmos genéticos o el algoritmo de optimización de serpiente.

También se menciona que, para el uso de procesamiento de imagen en estos sistemas, los modelos de mayor utilidad son YOLO, SSD, Faster R-CNN y Mask R-CNN debido a su confiabilidad para el reconocimiento de objetos por imagen. Se enfatiza la superioridad de YOLO frente al resto de modelos debido a su velocidad de procesamiento con su funcionalidad de una sola pasada que resulta eficiente para el reconocimiento de densidad vehicular en tiempo real (Chabchoub et al., 2021). Se resalta que el Mask R-CNN solo se utiliza como extensión a la aplicación del Faster R-CNN (Shin et al., 2019).

Además, se describe que los patrones visuales, priorizados por los autores en el entrenamiento de sus modelos de procesamiento de imagen para la identificación de vehículos, son las formas, el color, las dimensiones y las placas de vehículos (Li, Y. et al., 2021). Se menciona que a esta última se le debe brindar mayor atención en la creación de la base de imágenes de entrenamiento para facilitar el reconocimiento de un vehículo frente al resto de objetos.

## 5. CONCLUSIONES

A través de esta revisión de literatura, se determina que el uso de modelos o algoritmos de IA se convierte en la mejor opción para el desarrollo de soluciones contra la problemática de congestión vehicular y control de semáforos, pues aportan una mayor optimización del control de luces y señales de tránsito debido a su capacidad de predicción de la densidad de tráfico.

Se puede afirmar que el modelo de procesamiento de imagen que se adecúa mejor en el proceso de reconocimiento de vehículos es YOLO debido a los resultados obtenidos por los artículos revisados. Esto se debe a su capacidad para detectar varios objetos a la vez, su rápida toma de decisiones basada en los resultados detectados y su consumo óptimo de recursos computacionales, que permite su aplicación en tiempo real en un sistema de control de semáforos.

A partir del análisis de las observaciones en los artículos revisados, es posible determinar que para el entrenamiento del modelo de reconocimiento de objetos se debe priorizar aquellas secuencias de imágenes que muestran los vehículos y su matrícula al menos una vez por cuadro. Esto se debe a que así se mejora la precisión del modelo inteligente a la hora de distinguir vehículos de otros objetos en movimiento.

De acuerdo con el análisis de los resultados, se concluye que el algoritmo más preciso para la construcción de un sistema de control de semáforos es el filtro bayesiano aplicado en las CNN, pues permite modelar la incertidumbre de los datos del tráfico para facilitar el aprendizaje y predicción de la densidad vehicular, lo que mejora la precisión de las decisiones en el cambio de luces de los semáforos.

## REFERENCIAS

Aqib, M., Mehmood, R., Alzahrani, A., Katib, I., Albeshri, A., & Altowaijri, S. M. (2019). Smarter Traffic Prediction Using Big Data, In-Memory Computing, Deep Learning and GPUs. *Sensors*, *19*(9). https://doi.org/10.3390/s19092206

Bao, Y., Huang, J., Shen, Q., Cao, Y., Ding, W., Shi, Z., & Shi, Q. (2023). Spatial–Temporal Complex Graph Convolution Network for Traffic Flow Prediction. *Engineering Applications of Artificial Intelligence*, *121*. https://doi.org/10.1016/j.engappai.2023.106044

Chabchoub, A., Hamouda, A., Al-Ahmadi, S., & Cherif, A. (2021). *Intelligent Traffic Light Controller using Fuzzy Logic and Image Processing. International Journal of Advanced Computer Science and Applications*, *12*(4), 396-399. https://doi.org/10.14569/IJACSA.2021.0120450

Chahal, A., Gulia, P., Gill, N. S., & Priyadarshini, I. (2023). A Hybrid Univariate Traffic Congestion Prediction Model for IoT-Enabled Smart City. *Information*, *14*(5). https://doi.org/10.3390/info14050268

Chatterjee, K., De, A., & Chan, F. T. S. (2019). Real time traffic delay optimization using shadowed type-2 fuzzy rule base. *Applied Soft Computing Journal*, *74*, 226-241. https://doi.org/10.1016/j.asoc.2018.10.008

Cheng, R., Qiao, Z., Li, J., & Huang, J. (2023). Traffic Signal Timing Optimization Model Based on Video Surveillance Data and Snake Optimization Algorithm. *Sensors*, *23*(11). https://doi.org/10.3390/s23115157

Damadam, S., Zourbakhsh, M., Javidan, R., & Faroughi, A. (2022). An Intelligent IoT Based Traffic Light Management System: Deep Reinforcement Learning. *Smart Cities*, *5*(4), 1293-1311. https://doi.org/10.3390/smartcities5040066

Hao, S., Yang, L., Ding, L., & Guo, Y. (2019). Distributed Cooperative Backpressure-Based Traffic Light Control Method. *Journal of Advanced Transportation*, *2019*(1). https://doi.org/10.1155/2019/7481489

Hao, W., Rong, D., Yi, K., Zeng, Q., Gao, Z., Wu, W., Wei, C., & Scepanovic, B. (2020). Traffic Status Prediction of Arterial Roads Based on the Deep Recurrent Q-Learning. *Journal of Advanced Transportation*, *2020*(1). https://doi.org/10.1155/2020/8831521

Impedovo, D., Balducci, F., Dentamaro, V., & Pirlo, G. (2019). Vehicular traffic congestion classification by visual features and deep learning approaches: A comparison. *Sensors*, *19*(23). https://doi.org/10.3390/s19235213

Islam, Z., Abdel-Aty, M., & Mahmoud, N. (2022). Using CNN-LSTM to predict signal phasing and timing aided by High-Resolution detector data. *Transportation Research Part C: Emerging Technologies*, *141*. https://doi.org/10.1016/j.trc.2022.103742

Jin, J., & Ma, X. (2019). A non-parametric Bayesian framework for traffic-state estimation at signalized intersections. *Information Sciences*, *498*, 21-40. https://doi.org/10.1016/j.ins.2019.05.032

Joo, H., & Lim, Y. (2021). Traffic Signal Time Optimization Based on Deep Q-Network. *Applied Sciences*, *11*(21). https://doi.org/10.3390/app11219850

Kim, M., Schrader, M., Yoon, H. S., & Bittle, J. A. (2023). Optimal Traffic Signal Control Using Priority Metric Based on Real-Time Measured Traffic Information. *Sustainability*, *15*(9). https://doi.org/10.3390/su15097637

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology*, *51*(1), 7-15. https://doi.org/10.1016/j.infsof.2008.09.009

Korecki, M., & Helbing, D. (2022). Analytically Guided Reinforcement Learning for Green It and Fluent Traffic. *IEEE Access*, *10*, 96348-96358. https://doi.org/10.1109/ACCESS.2022.3204057

Kumaran, S. K., Mohapatra, S., Dogra, D. P., Roy, P. P., & Kim, B. G. (2019). Computer vision-guided intelligent traffic signaling for isolated intersections. *Expert Systems with Applications*, *134*, 267-278. https://doi.org/10.1016/j.eswa.2019.05.049

Li, Y., Chen, Y., Yuan, S., Liu, J., Zhao, X., Yang, Y., & Liu, Y. (2021). Vehicle detection from road image sequences for intelligent traffic scheduling. *Computers and Electrical Engineering*, *95*. https://doi.org/10.1016/j.compeleceng.2021.107406

Li, Z., Yu, H., Zhang, G., Dong, S., & Xu, C. Z. (2021). Network-wide traffic signal control optimization using a multi-agent deep reinforcement learning. *Transportation Research Part C: Emerging Technologies*, *125*. https://doi.org/10.1016/j.trc.2021.103059

Liu, W. L., Gong, Y. J., Chen, W. N., & Zhang, J. (2020). EvoTSC: An evolutionary computation-based traffic signal controller for large-scale urban transportation networks. *Applied Soft Computing*, *97*. https://doi.org/10.1016/j.asoc.2020.106640

Mukhtar, H., Afzal, A., Alahmari, S., & Yonbawi, S. (2023). CCGN: Centralized collaborative graphical transformer multi-agent reinforcement learning for multi-intersection signal free-corridor. *Neural Networks*, *166*, 396-409. https://doi.org/10.1016/j.neunet.2023.07.027

Phursule, R., Lal, D., Waghere, S., Mughni, M. A., Ransubhe, S., & Shiralkar, C. (2023). Enhancing Traffic Flow Using Computer Vision Based - Dynamic Traffic Light Control and Lane Management. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(7S), 386-391. https://doi.org/10.17762/ijritcc.v11i7s.7014

Rasheed, F., Yau, K. L. A., Noor, R. M., & Chong, Y. W. (2022). Deep Reinforcement Learning for Addressing Disruptions in Traffic Light Control. *Computers, Materials and Continua*, *71*(2), 2225-2247. https://doi.org/10.32604/cmc.2022.022952

Shehu, H. A., Sharif, M. H., & Ramadan, R. A. (2020). Distributed Mutual Exclusion Algorithms for Intersection Traffic Problems. *IEEE Access*, *8*, 138277-138296. https://doi.org/10.1109/ACCESS.2020.3012573

Shin, J., Roh, S., & Sohn, K. (2019). Image-Based Learning to Measure the Stopped Delay in an Approach of a Signalized Intersection. *IEEE Access*, *7*, 169888-169898. https://doi.org/10.1109/ACCESS.2019.2955307

Stoilova, K., & Stoilov, T. (2022). Model Predictive Traffic Control by Bi-Level Optimization. *Applied Sciences*, *12*(9). https://doi.org/10.3390/app12094147

Suga, S., Fujimori, R., Yamada, Y., Ihara, F., Takamura, D., Hayashi, K., & Kurihara, S. (2023). Traffic information interpolation method based on traffic flow emergence using swarm intelligence. *Artificial Life and Robotics*, *28*(2), 367-380. https://doi.org/10.1007/s10015-022-00847-7

Szoke, L., Aradi, S., & Bécsi, T. (2023). Traffic Signal Control with Successor Feature-Based Deep Reinforcement Learning Agent. *Electronics*, *12*(6). https://doi.org/10.3390/electronics12061442

Tan, J., Yuan, Q., Guo, W., Xie, N., Liu, F., Wei, J., & Zhang, X. (2022). Deep Reinforcement Learning for Traffic Signal Control Model and Adaptation Study. *Sensors*, *22*(22). https://doi.org/10.3390/s22228732

Tunc, I., & Soylemez, M. T. (2023). Fuzzy logic and deep Q learning based control for traffic lights. *Alexandria Engineering Journal*, *67*, 343-359. https://doi.org/10.1016/j.aej.2022.12.028

Vélez-Serrano, D., Álvaro-Meca, A., Sebastián-Huerta, F., & Vélez-Serrano, J. (2021). Spatio-temporal traffic flow prediction in madrid: An application of residual convolutional neural networks. *Mathematics*, *9*(9). https://doi.org/10.3390/math9091068

Wakkumbura, R. T., Hettige, B., & Edirisuriya, A. (2021). Real-time traffic controlling system using multi-agent technology. *Journal Européen des Systèmes Automatisés*, *54*(4), 633-640. https://doi.org/10.18280/jesa.540413

Wang, H., Zhu, J., & Gu, B. (2023). Model-Based Deep Reinforcement Learning with Traffic Inference for Traffic Signal Control. *Applied Sciences*, *13*(6). https://doi.org/10.3390/app13064010

Wu, Q., Wu, J., Shen, J., Du, B., Telikani, A., Fahmideh, M., & Liang, C. (2022). Distributed agent-based deep reinforcement learning for large scale traffic signal control. *Knowledge-Based Systems*, *241*. https://doi.org/10.1016/j.knosys.2022.108304

Yuan, Y., Zhang, Z., Yang, X. T., & Zhe, S. (2021). Macroscopic traffic flow modeling with physics regularized Gaussian process: A new insight into machine learning applications in transportation. *Transportation Research Part B: Methodological*, *146*, 88-110. https://doi.org/10.1016/j.trb.2021.02.007

Zhang, L., Wang, L., & Zhao, Q. (2020). Traffic State Recognition of Intersection Based on Image Model and PCA Hashing. *Journal of Advanced Transportation*, *2020*(1). https://doi.org/10.1155/2020/3828395

Zhao, P., Yuan, Y., & Guo, T. (2022). Extensible Hierarchical Multi-Agent Reinforcement-Learning Algorithm in Traffic Signal Control. *Applied Sciences*, *12*(24). https://doi.org/10.3390/app122412783

Zheng, Q., Xu, H., Chen, J., Zhang, D., Zhang, K., & Tang, G. (2022). Double Deep Q-Network with Dynamic Bootstrapping for Real-Time Isolated Signal Control: A Traffic Engineering Perspective. *Applied Sciences*, *12*(17). https://doi.org/10.3390/app12178641

# Gestão do Conhecimento como Ferramenta Estratégica de Inovação nas Organizações. Uma Revisão Integrativa

Elaine Rodrigues Koller
Dr.elaine.adv@gmail.com
0009-0004-9523-0227
Universidade do Estado de Santa Catarina ( UFSC)

Paulo Roberto de Moura
Paulormoura1@gmail.com
0000-0003-3434-7443
Universidade Leonardo Da Vince (Uniasselvi).

Dra. Patrícia de Sá Freire
patriciadesafreire@gmail.com
0000-0002-9259-682X
Universidade Federal de Santa Catarina (UFSC)

Recebido: 14 de maio de 2024 / Aceito: 13 de junho de 2024
doi: https://doi.org/10.26439/interfases2024.n19.7114

RESUMO. O artigo investiga a Gestão do Conhecimento (GC) como ferramenta estratégica para fomentar a inovação nas organizações. O estudo foca em como a GC pode ser empregada para aumentar a competitividade e a eficácia organizacional por meio da inovação. Utilizando uma metodologia de revisão integrativa, a pesquisa identificou 13 categorias principais de investigação na literatura atual, ressaltando a inter-relação entre a GC e diversos aspectos organizacionais, como competitividade, estratégia, inovação e aprendizagem. O problema central do estudo questiona como a GC tem sido aplicada como ferramenta estratégica para a gestão eficaz da inovação, quanto ao objetivo geral é analisar o papel da GC na promoção da inovação dentro das organizações, visando obter vantagens competitivas e um melhor posicionamento no mercado. A pesquisa foi realizada mediante uma revisão integrativa, analisando diversas publicações acadêmicas para avaliar como a GC é abordada em termos de inovação estratégica. Os recortes das publicações são de 1990 a 2023, selecionados com base em sua relevância para a gestão do conhecimento e inovação. Os resultados indicam que organizações concentradas na GC alcançam resultados melhores em inovação, impactando diretamente em seu desempenho. Os achados enfatizam a necessidade de uma gestão estratégica do conhecimento para promover inovação contínua. O estudo contribui para a compreensão e integração

da GC como um catalisador para a inovação e a estratégia organizacional, criando um ambiente propício ao desenvolvimento de novas ideias e melhorias contínuas.

PALAVRAS-CHAVE: organização / gestão do conhecimento / inovação

## KNOWLEDGE MANAGEMENT AS A STRATEGIC INNOVATION TOOL IN ORGANIZATIONS. AN INTEGRATIVE REVIEW

ABSTRACT. The article investigates Knowledge Management (KM) as a strategic tool to foster innovation within organizations. The study focuses on how KM can be used to enhance competitiveness and organizational effectiveness through innovation. Using an integrative review methodology, the research identified 13 main categories of investigation in the current literature, highlighting the interrelationship between KM and various organizational aspects such as competitiveness, strategy, innovation, and learning. The central problem of the study questions how KM has been applied as a strategic tool for effective innovation management, while the general objective is to analyze the role of KM in promoting innovation within organizations, aiming to achieve competitive advantages and better market positioning. The research was conducted through an integrative review, analyzing various academic publications to assess how KM is approached in terms of strategic innovation. The publication excerpts range from 1990 to 2023, selected based on their relevance to knowledge management and innovation. The results indicate that organizations focused on KM achieve better innovation outcomes, directly impacting their performance. The findings emphasize the need for strategic knowledge management to promote continuous innovation. The study contributes to the understanding and integration of KM as a catalyst for innovation and organizational strategy, creating an environment conducive to the development of new ideas and continuous improvements.

KEYWORDS: organization / knowledge management / innovation

## 1. INTRODUÇÃO

Em um mundo marcado pela incerteza constante, o conhecimento se torna a chave para a competitividade organizacional (Nonaka & Takeuchi, 1997; Moraes et al., 2023). O conhecimento na sociedade deixa de ser apenas um recurso e passa a ser um diferencial competitivo, pois sem ele não há mudança no poder. Isso enfatiza que o conhecimento é a chave para toda mudança organizacional (Toffler, 1990; Wang et al., 2009).

Essa perspectiva é corroborada por Quinn (1992) e Abbas et al. (2020) que destacam a centralidade da capacidade intelectual na produção das empresas contemporâneas, a qual se traduz em produtos e serviços valorizados, impregnados de know-how. O conhecimento transcende os limites do processo gerencial, permeando e moldando profundamente a estrutura organizacional. Em um mundo cada vez mais competitivo, onde o conhecimento e a inovação se transformam em diferenciais cruciais para o sucesso, as empresas que os dominam se posicionam à frente no mercado (Nonaka & Takeuchi, 1997; Sondhi, et al., 2024).

Empresas que geram conhecimento lidam com múltiplas ideias, alimentando assim a inovação (Starkey, 1997). Essa capacidade de inovar permite que as organizações capturem as mudanças externas, transformem seus processos internos e devolvam à sociedade novos conhecimentos e inovações. Essas inovações permeiam todas as atividades humanas que se renovam e se atualizam, desempenhando um papel essencial para as empresas (Barbieri & Alvares, 2003).

Para que as organizações se tornem estratégicas no mercado, necessitam de uma reorganização completa, advinda da geração do conhecimento (Barbieri & Alvares, 2003). Posicionar-se na gestão do conhecimento, inovação e estratégia é essencial para que as organizações revisem seus métodos, mantenham o foco nos negócios e alcancem bons resultados a longo prazo. Sondhi et al. (2024) destacam que a capacidade de integração do conhecimento e a inovação são essenciais para alcançar uma vantagem competitiva sustentável, ressaltando a importância do conhecimento como um fator determinante na orientação estratégica das empresas.

A partir de 1990, a gestão do conhecimento aplicada nas organizações passou a ser vista como criação e renovação, sendo responsável pela vantagem competitiva (Venzin et al., 1998). Autores como Sondhi et al. (2024), Fitriati, et al. (2020), Pontes (2022) e Nonaka e Takeuchi (1997) definem a inovação como um elemento importante no contexto organizacional, ampliando a influência das capacidades dinâmicas sobre o sucesso empresarial, o que se relaciona diretamente com a gestão do conhecimento, deixando uma lacuna entre a GC e gestão eficaz.

Nessa perspectiva, a presente revisão integrativa pretende corroborar e busca responder, com publicações já existentes, à seguinte questão de pesquisa: Como a Gestão do Conhecimento (GC) vem sendo investigada como ferramenta estratégica para uma gestão eficaz de inovação?

O objetivo geral desta pesquisa é analisar e levantar como a Gestão do Conhecimento (GC)vem sendo estudada como ferramenta estratégica para promover uma gestão eficaz da inovação.

## 2. REFERENCIAL TEÓRICO

Entender como a competitividade, inovação e gestão do conhecimento impacta na vida das organizações torna-se um desafio. Esta investigação procura analisar o que está se produzindo nas pesquisas acadêmicas, especialmente, como a Gestão do Conhecimento (GC) é analisada como ferramenta estratégica para uma gestão eficaz da inovação. Senge (2018) destaca a importância de uma cultura organizacional que promova a aprendizagem contínua para impulsionar a inovação e o desempenho das organizações. Takeuchi e Nonaka (2008) em "A Empresa Criadora de Conhecimento", introduzem o conceito da "espiral do conhecimento", destacando como a gestão do conhecimento pode catalisar a inovação e o crescimento organizacional.

### 2.1. Organizações

A articulação entre gestão do conhecimento e aprendizagem organizacional pode ser concebida a partir de diversas atividades, sendo o foco da gestão do conhecimento o conteúdo que uma empresa cria, captura e, finalmente, usa, considerando que o foco da aprendizagem organizacional é a prática e a implantação do conhecimento (Easterby & Lyles, 2003). Essa gestão do conhecimento abrange sua criação, sua aquisição, seu armazenamento e seu uso acumulado. De acordo com Jennex e Zyngier (2007), um sistema de gestão do conhecimento sugere um ambiente corporativo que incentiva a aprendizagem organizacional.

Para Lacombe e Heilborn (2005), o mercado se torna competitivo, forçando as organizações a se reinventarem, tanto em processos quanto em produtos. De acordo com Pontes (2022), até recentemente, a maioria das empresas concentrava-se principalmente em suas concorrentes locais e regionais, com pouca atenção às concorrentes internacionais. Atualmente, competem com produtos de todo o mundo, o que exige que seus produtos tenham preços competitivos, qualidade elevada e constante inovação (Pontes, 2022). A globalização também facilitou a disponibilidade e circulação de grandes volumes de informação, tornando essencial a presença de talentos para assegurar a continuidade das organizações (Pontes, 2022).

Nas palavras de Abbas et al. (2020), uma organização com processos de aprendizagem mais avançados pode avaliar com precisão os pontos fortes e fracos de suas concorrentes. Isso permite que tais empresas se tornem mais eficientes, convertam suas falhas em sucessos e introduzam novas ideias e habilidades sustentáveis. Sobre aprendizagem organizacional, há um impacto estatisticamente significativo e positivo

na inovação, pois as empresas que têm um processo de aprendizado ativo são bem-sucedidas em fornecer a seus clientes produtos e serviços inovadores. O aprendizado aprimorado as torna capazes de não perder nenhuma oportunidade de introduzir produtos e serviços para atender às demandas do mercado em constante mudança (Abbas et al, 2020).

Segundo King (2007), as organizações podem aprender porque são compostas por indivíduos e seus colaboradores precisam, aprender antes da aprendizagem da organização. A melhoria contínua somente acontecerá se a organização estiver aberta para mudanças (Gemünden et al., 2007). Brown (1999) reforça que essas organizações somente serão bem-sucedidas se criarem um ambiente onde as mudanças aconteçam muito rapidamente, propiciando assim o compartilhamento de novos conhecimentos. Nas palavras de Alasoini et al. (2007), as organizações serão competitivas quando forem dependentes para produzir inovações.

Com o avanço da ciência e da tecnologia, o conhecimento passou a ser um componente chave não apenas em produtos e serviços, mas em toda atividade econômica. Nesse sentido, as empresas estão integrando a inovação em suas estratégias e tratando-a como uma rotina diária que precisa de organização e gestão cuidadosas (Polyakov et al., 2023). As organizações ganham mais conhecimento quando trocam esse conhecimento com o mercado. O envolvimento de todos permite o acesso a um ambiente propício para o intercâmbio de conhecimento, inovação e desenvolvimento tecnológico, resultante do relacionamento entre organizações e o mercado em que atuam (Fioravanti et al., 2023).

### 2.2. Inovação Organizacional

A inovação se configura como um elemento crucial para a sobrevivência e o sucesso das empresas no cenário competitivo atual. Definida como a "implementação de novas ideias que criam valor" (Linder et al., 2003), a inovação assume um papel fundamental na busca por vantagens competitivas e no alcance de benefícios valiosos para as organizações (Harrison & Sanson, 2002).

Damanpour e Aravind (2012) argumentam que a capacidade de aprender e de aplicar novos conhecimentos, por parte de uma empresa, é crucial, especialmente em ambientes que estão em constante mudanças. Essa capacidade de adaptação não só melhora a flexibilidade estratégica da empresa, mas também aprimora seu desempenho geral. A inovação é responsável por essa nova aprendizagem, direcionando a novas abordagens para estruturar estratégias e tarefas, modificar processos de gestão e sistemas administrativos, motivar e recompensar membros da organização, e permitir adaptações e mudanças organizacionais (Damanpour & Aravind, 2012).

Tidd e Thuriaux-Alemán (2016) propõem uma visão da inovação como um processo de gestão do conhecimento. Esse processo contínuo de criação, compartilhamento e aplicação se torna um elemento essencial para o desenvolvimento de inovações. O acesso a

um amplo espectro de conhecimentos, tanto internos quanto externos, aumenta a probabilidade de desenvolvimento de inovações. Isso, por sua vez, auxilia as organizações a alcançarem benefícios valiosos como eficiência, sustentabilidade, crescimento e prosperidade econômica (Adams & Lamont, 2003), pois, a inovação desempenha um papel vital na competitividade das empresas (Costa et al., 2023).

A inovação se manifesta em diversas formas dentro das empresas e pode ser classificada em quatro dimensões principais: Inovação de produção: refere-se à introdução de novos métodos de produção que otimizam processos e aumentam a eficiência. Inovação de processo: envolve a criação de novos processos e procedimentos que aprimoram o funcionamento da organização como um todo. Inovação de posição: refere-se ao desenvolvimento de novos produtos ou serviços que atendem às demandas dos clientes e diferenciam a empresa no mercado. Inovação de paradigma: envolve a criação de novas tecnologias ou modelos de negócios que revolucionam o setor e redefinem as regras do jogo competitivo (Wang & Ahmed, 2004).

Para os autores pesquisados, a inovação ocorre mediante a aprendizagem e aplicação de novos conhecimentos, a gestão contínua do conhecimento e o acesso a informações que permitem a criação de novas ideias e abordagens. Os propulsores podem ser classificados como a) capacidade de aprendizagem e aplicação do conhecimento (Damanpour & Aravind , 2012): habilidade de aprender e aplicar novos conhecimentos; b) disponibilidade de Conhecimento (Adams & Lamont, 2003): acesso a conhecimentos internos e externos; c) classificação da Inovação (Wang & Ahmed, 2004): produção, processo, posição e paradigma; d) Gestão do Conhecimento (Tidd & Thuriaux-Alemán, 2016): processo contínuo de criação, compartilhamento e aplicação de conhecimento.

### 2.3. Gestão do Conhecimento

A Gestão do Conhecimento é definida como um conjunto de procedimentos, métodos e técnicas claramente descritos, e empregada para encontrar informações valiosas em diferentes procedimentos administrativos, com o propósito de capacitar os negócios em perspectivas alternativas. Primeiro, para garantir suas perspectivas e realizações, e depois, para concentrar-se na construção de uma empresa ou indústria de forma sensata, considerando seus ativos de conhecimento (Takeuchi & Nonaka, 2008).

A Gestão do Conhecimento envolve o processo técnico, a abordagem cultural e a organizacional, valorizando a criação e o compartilhamento do conhecimento (Takeuchi & Nonaka, 2008). Para Valentim (2010), o conhecimento nas organizações representa a base para a formulação de estratégias e, portanto, o insumo para o desenvolvimento organizacional. Nesse sentido, a identificação dos fluxos de conhecimento facilita a aplicação de metodologias capazes de melhorar os processos de trabalho, identificando o armazenamento e a utilização da informação para aquilo que pode se configurar no alcance de vantagens competitivas para as organizações.

A gestão do conhecimento é um pilar central para o sucesso organizacional. Davenport e Prusak (1998) destacam a importância de uma administração eficaz do saber corporativo como um diferencial estratégico. Os autores argumentam que o conhecimento não é apenas um recurso valioso, mas um ativo que, se bem gerido, pode levar a uma vantagem competitiva significativa e fomentar a inovação contínua.

Por outro lado, Wiig (1994) oferece uma perspectiva detalhada sobre os princípios essenciais da gestão do conhecimento. Sugere que entender e aplicar o conhecimento de maneira estratégica é crucial para a competitividade e o crescimento. Toffler (1990) enfatiza que o conhecimento deve ser cultivado e compartilhado dentro das organizações, permitindo assim que a inovação floresça e que as empresas se mantenham à frente em mercados cada vez mais dinâmicos.

O conhecimento vira uma vantagem quando uma organização consegue extrair dados úteis e insights de seus ativos de informação durante suas atividades normais. Aprender é crucial para as organizações, mas isso não basta para melhorar o desempenho delas. O desempenho pode melhorar com a partilha de conhecimento atualizado nos processos da cadeia de suprimentos e operações, usando ferramentas de gestão do conhecimento (Jha & Karn, 2019).

O conhecimento não é apenas um extra; ele é central para definir competências, criar valor e assegurar vantagens competitivas e um bom desempenho no mercado (Polyakov et al., 2023).

## 3. METODOLOGIA

Para a abordagem desta pesquisa adotou-se o método qualitativo devido à complexidade e profundidade do tema (Gressler, 2003). A metodologia qualitativa se concentra em explorar e compreender a realidade de maneira profunda. Ela envolve técnicas que permitem aos pesquisadores capturarem e interpretarem a complexidade dos fenômenos sociais. A pesquisa qualitativa é uma abordagem exploratória que busca compreender fenômenos complexos em contextos específicos até então não observados por pesquisadores (Gressler, 2003; Raupp & Beuren, 2006).

Este estudo baseou-se em uma revisão integrativa da literatura sobre Gestão do Conhecimento como ferramenta estratégica para a gestão eficaz. Este é um método científico que permite realizar discussões sobre temas específicos, a partir de estudos já publicados sejam como abordagens quantitativas ou qualitativas. Possui fases que orientarão o pesquisador na coleta de dados e na análise dos resultados (Whittemore & Knafl, 2005; Pompeo et al., 2009). Corroborando o tema, March e Smith (1995) definem que os métodos se baseiam em um conjunto de constructos subjacentes (linguagem) e em uma representação (modelo) dentro de um espaço de solução. Essa metodologia

nos ajudou a explorar as nuances da Gestão do Conhecimento e suas implicações para a inovação nas organizações.

Também se utilizou o método de investigação indutivo. Para Lakatos e Marconi (2006), é um processo mental que, a partir da observação de dados particulares bem estabelecidos, permite inferir uma verdade geral ou universal. Inicia-se com uma premissa para, posteriormente, alcançar uma teoria.

Para investigar o problema, foram estabelecidos objetivos exploratórios e descritivos. Conforme Gil (1991), a pesquisa exploratória tem como objetivo principal aprimorar ideias ou descobrir novas instituições. Cervo e Bervian (2002) indicam que a pesquisa descritiva tem como objetivo observar, registrar, analisar e correlacionar fatos ou fenômenos (variáveis) sem que haja qualquer manipulação deles.

A pesquisa teórica foi apoiada pela técnica de coleta de dados e revisão integrativa. Inicialmente, buscou-se conhecer o tema proposto realizando uma busca nas bases de dados do Portal Capes, Scopus e Web of Science. Para garantir uma busca eficiente e relevante, definimos cuidadosamente descritores, refletindo todos os aspectos essenciais da Gestão do Conhecimento e inovação como palavras-chave e operadores booleanos ("*Knowledge management*" OR "*Strategic knowledge management*" OR, "*Innovation processes*" OR "*Impact on innovation*" OR "*Innovative processes*") AND ("*Company competitiveness*" OR "*Organizational competitiveness*" OR "*Organization*"). Utilizamos operadores booleanos nas bases de dados Scopus e Web of Science o que é uma prática recomendada para assegurar uma busca abrangente e relevante. Como critério para as palavras-chave, estas deveriam pertencer à área de gestão empresarial e possuir relevância para o estudo proposto.

Os critérios na base de dados para inclusão foram: publicações revisadas por pares, publicadas entre 1990 e 2023, escritas em português ou inglês, a Gestão do conhecimento em organização, artigos e teses, publicações gratuitas revisadas por pares. As publicações encontradas foram agrupadas em uma planilha do Microsoft Excel (®) e conforme as delimitações temporais, obteve-se um total de 124 publicações. Aplicando os critérios da base restaram 110 para análise, desses, 17 estavam em duplicidade e foram excluídos, 38 foram excluídos após leitura crítica de resumo, palavras-chave e 55 ficaram para uma leitura crítica na integra, conforme quadro 1.

**Quadro 1**

*Critério de seleção das bases de dados*

| Base | Busca | Resultado dos critérios de seleção da base | Após excluídos Arquivos duplicados | Após leitura resumo e palavra-chave | Total de artigo excluídos desta busca |
|---|---|---|---|---|---|
| Web Of Science | 35 | 30 | 20 | 12 | 8 |
| Scopus | 89 | 80 | 73 | 43 | 30 |
| Total | 124 | 110 | 93 | 55 | 38 |

*Nota*. Adaptado dos critérios de seleção das bases de dados (2023).

A exclusão aconteceu depois da leitura do resumo e das palavras-chave. Eliminou-se 38 publicações que não tratavam de forma aprofundada a GC como ferramenta estratégica para uma gestão eficaz ou abordavam de forma superficial a temática do artigo.

Após conduzir o processo de seleção e análise crítica dos resumos, conforme mencionado no quadro 1 (Critério de seleção das bases de dados), foram estabelecidos critérios para inclusão e exclusão dos artigos que compuseram a amostra final. Seguindo as orientações de Braun e Clarke (2012), essa etapa permitiu a realização de leituras aprofundadas e repetitivas para identificar ideias ou conceitos relevantes à temática da pesquisa. Adicionalmente, os códigos foram levantados mediante os dados brutos coletados das publicações por meio de uma análise minuciosa.

Os dados coletados foram organizados em categorias, possibilitando uma análise detalhada de suas interligações e uma revisão constante durante a elaboração da pesquisa. Essa abordagem serve como um guia metodológico para identificar os principais temas, a partir do conjunto de dados obtidos. Após uma análise detalhada dos resumos, foram estabelecidos os critérios de exclusão para a amostra final.

Critérios de inclusão:

a) Estudos publicados entre 1990 e 2023; b) estudos empíricos e qualitativos; c) artigos que tratavam de Gestão do conhecimento em organizações; d) inovação estratégica; e) inovação nas organizações; f) estratégias organizacionais, g) publicações no formato gratuito nas plataformas.

Nesta etapa foram eliminadas 33 publicações que não versavam sobre a pergunta norteadora, resultando no quadro 2.

**Quadro 2**

*Artigos que farão parte da amostra.*

| String | Base | Leitura crítica na integra | Eliminados após leitura integra | Total que fazem parte da amostra |
|---|---|---|---|---|
| 1 | Web Of Science | 12 | 9 | 6 |
| | Scopus | 43 | 24 | 19 |
| Total | | 55 | 33 | 25 |

*Nota*. Baseado dos critérios das publicações lidas na integra (2023).

### 3.1 Análise de Dados

Após as leituras das publicações, foram selecionados estudos que versavam sobre a GC como ferramenta estratégica para uma gestão eficaz. Realizamos a análise de conteúdo para identificar e validar temas de forma sistemática, seguindo a metodologia proposta por Bardin (1977), dividida em três etapas: 1) pré-análise, 2) exploração do material e 3) tratamento dos resultados, inferência e interpretação. Isso permitiu aos pesquisadores agruparem os artigos por categorias. 1) Pré- análise: realização da leitura dos artigos encontrados, para identificar sua aderência a pesquisa, definido o *corpus* de análise. 2) Exploração do material: seleção de palavras-chave ou frases, que apareciam com mais frequência classificando os artigos, agregando informações e categorizando-os com os temas. 3) Tratamento dos resultados, inferências e interpretações: extração dos significados e as informações, após a leitura completa dos artigos, para que os pesquisadores pudessem formalizar as categorias e gerar novos conhecimentos sobre o direcionador proposto. Após a leitura dos artigos, foram categorizados por temas correlatos que foram identificados na seção discussão e resultados.

## 4. DISCUSSÃO DOS RESULTADOS

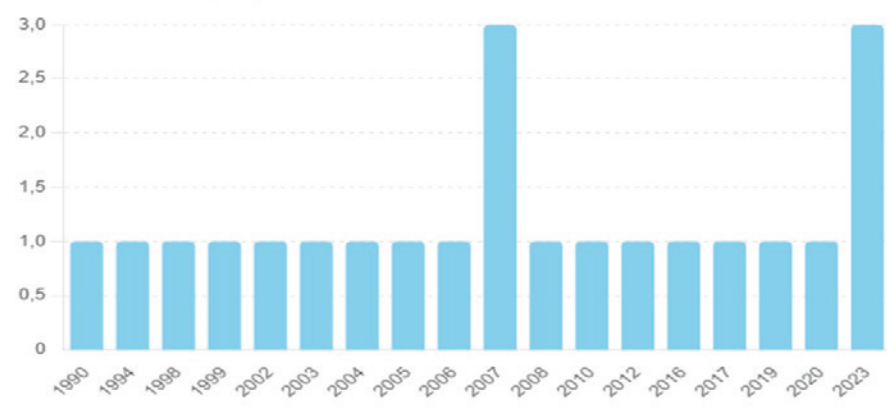
Para o autor Creswell (2018), esta seção deve expor os achados de maneira clara e direta, empregando tabelas e figuras para tornar os dados mais compreensíveis. Deve-se interpretar e analisar esses achados à luz da literatura existente, proporcionando uma análise crítica que vincule os resultados ao objetivo da pesquisa.

### 4.1 Análise Bibliométrica

Com base na pesquisa realizada, foram encontrados artigos que orientam sobre como a Gestão do Conhecimento (GC) vem sendo pesquisada como ferramenta estratégica para uma gestão eficaz de inovação nas organizações. Os artigos indexados permitiram aos pesquisadores contribuírem na comunidade científica com base nas experiências extraídas a partir da leitura dos 25 artigos. Todos estavam dentro do recorte temporal de 1990 a 2023, conforme mostra o gráfico 1.

**Grafico 1**

*Evolução da Pesquisa sobre G.C como Ferramenta Estratégica para Inovação*




*Nota*. Baseado na revisão integrativa (2023).

Conforme o gráfico, houve publicações sobre o tema pesquisado em vários anos, porém com oscilações. Percebe-se que o auge ocorreu em 2007, seguido de uma queda, com um aumento em 2023. Para melhor entender como a pesquisa sobre GC vem evoluindo como ferramenta estratégica para inovação, procurou-se identificar como elas se alinham. Essas publicações estão alinhadas da seguinte forma: "Competitividade X Inovação", "Capacidade de aprendizagem e aplicação do conhecimento" e "GC x Inovação" possuem 4 publicações. Por outro lado, "Cultura Organizacional e Aprendizagem", "Princípios e Valorização da Gestão do Conhecimento" e "Inovação Estratégica" possuem 3, 2 e 2 publicações respectivamente. Finalmente, "Sistemas de Gestão do Conhecimento", "Impacto da GC na Inovação", "Barreiras a inovação", "GC em Cadeias de Suprimentos", "GC em Produtos e Serviços", "Tecnologias da Informação e GC" e "Gestão Estratégica" têm 1 publicação cada. Pautados no objetivo da pesquisa, foram levantadas 13 categorias, conforme indicado no Quadro 3, na seção de análise de conteúdo.

### 4.2 Análise de Conteúdo

Para a análise de conteúdo, segue-se a ideia de Bardin (1977) que estabelece regras para garantir a qualidade da análise, tais como: a) exaustividade: todos os elementos do corpus foram considerados; b) representatividade: a amostra do universo inicial foi representativa; c) homogeneidade: as publicações obedeceram a critérios de escolhas; d) pertinência: as publicações foram de encontro com o problema de pesquisa. O resultado aparece no quadro 3.

**Quadro 3**

*Categorias das Publicações*

| Categorias | Autor / Ano |
|---|---|
| GC e Inovação Organizacional | Tidd e Thuriaux-Alemán (2016); Abbas et al. (2020); Jennex e Zyngier (2007); Takeuchi e Nonaka (2008). |
| Cultura Organizacional e Aprendizagem | Senge (2018); Easterby-Smith e Lyles (2003); King (2007). |
| Sistemas de Gestão do Conhecimento | Jennex e Zyngier, 2007 |
| Competitividade e Inovação | Lacombe e Heilborn (2005); Brown (1999), Alasoini et al. (2007); Costa et al. (2023). |
| Princípios e valorização da Gestão do Conhecimento | Wiig (1994); Davenport e Prusak (1998) |
| Inovação Estratégica | Valentim (2010); Toffler (1990) |
| Barreiras à Inovação | Gemünden et al. (2007) |
| Impacto da GC na Inovação | Fioravanti et al. (2023) |
| GC em Cadeias de Suprimentos | Jha e Karn (2019) |
| GC em Produtos e Serviços | Pontes (2022) |
| Tecnologias da Informação e GC | Wang & Ahmed, (2004), Wang et al. (2009) |
| Capacidade de aprender e aplicar conhecimento vantagem competitiva inovadora. | Damanpour e Aravind (2012); Harrison e Samson (2002); Adams e Lamont (2003); Linder et al. (2003) |
| GC e Gestão Estratégica | Polyakov et al. (2023). |

*Nota*. Baseada na seção fundamentação teórica (2023).

A primeira categoria aborda artigos que destacam a inovação decorrente da Gestão do Conhecimento. Os autores Tidd e Thuriaux-Alemán (2016), Abbas et al. (2020), Jennex, (2007), e Takeuchi e Nonaka (2008) destacam a importância de criar, capturar e utilizar conhecimento para impulsionar a inovação. Na segunda categoria são discutidos os artigos que enfatizam a importância da Cultura Organizacional e da Aprendizagem, com autores como Senge (2018), Easterby-Smith e Lyles (2017) e King (2007), discutindo a promoção de uma cultura de aprendizagem contínua que suporta a inovação e o desempenho organizacional. Em relação à categoria "Sistemas de Gestão do Conhecimento", o autor Jennex (2007) aborda a infraestrutura tecnológica e processual para a GC. Por outro lado, na categoria "Competitividade e Inovação", os autores Lacombe e Heilborn (2005), Brown (1999), Alasoini et al. (2007) e Costa et al. (2023) examinam como a inovação e a gestão eficaz do conhecimento influenciam a competitividade no mercado.

Autores como Wiig (1994), e Davenport e Prusak (1998) comentam sobre os fundamentos teóricos e práticos da GC, enfatizando como o conhecimento deve ser gerenciado como um ativo estratégico. Conforme Valentim (2010) e Toffler (1990), na categoria "Inovação Estratégica", a gestão do conhecimento é a base estratégica para o desenvolvimento organizacional. Eliminar possíveis barreiras à inovação e manter a organização

aberta às mudanças são os postulados de Gemünden et al. (2007) na categoria "Barreiras à Inovação". O impacto da GC na Inovação guia o artigo de Fioravanti et al. (2023). Na categoria "GC em Cadeias de Suprimentos e Operações", Jha e Karn (2019) salientam que a GC atrelada às cadeias de suprimentos e operações pode melhorar o desempenho e a inovação das organizações.

Pontes (2022), na categoria "GC em Produtos e Serviços", discute como a implementação da GC pode gerar impactos positivos em diversos aspectos organizacionais como preço, qualidade e inovação constante. Segundo Wang et al. (2009; Wang & Ahmed, (2004), é necessária a criação de tecnologias para redefinir as regras de competitividades organizacionais na categoria "Tecnologias da Informação e GC". Por sua vez, autores como Damanpour e Aravind (2012), Harrison e Samson (2002), Adams e Lamont (2003), e Linder et al. (2003) reflexionam sobre a "Capacidade de aprender e aplicar conhecimento como vantagem competitiva inovadora". Por fim, Polyakov et al. (2023), na categoria "GC e Gestão Estratégica", destacam que as empresas devem gerenciar e estruturar o conhecimento de maneira estratégica para que possam inovar de forma constante e eficaz. Assim, a gestão do conhecimento torna-se um aspecto crucial, ajudando as organizações a permanecerem competitivas e inovadoras em um mercado cada vez mais dinâmico.

A análise revelou que a pressão competitiva e as constantes mudanças no ambiente empresarial exigem uma adaptação dinâmica das organizações, em que a inovação se torna essencial para atender às demandas dos clientes e se manter relevante no mercado. Ao cruzar os conceitos de GC e inovação, verificou-se que uma GC bem gerida é fundamental para promover a inovação, que, por sua vez, impulsiona mudanças e estratégias organizacionais, contribuindo para uma gestão eficaz e eficiente. A amostra desta pesquisa permitiu compreender, analisar e levantar como a Gestão do Conhecimento (GC) vem sendo pesquisada como ferramenta estratégica para promover uma gestão eficaz da inovação e levando os pesquisadores a alcançarem o objetivo geral desta pesquisa.

## 5. CONCLUSÃO

Este estudo destaca a importância da Gestão do Conhecimento (GC) como uma ferramenta estratégica para promover a inovação nas organizações, resultando em vantagem competitiva no mercado. A análise revelou que a pressão competitiva e as constantes mudanças no ambiente empresarial exigem uma adaptação dinâmica, na qual a inovação se torna essencial para atender às demandas dos clientes e se manter relevante no mercado.

Ao cruzar os conceitos de GC e inovação, verificou-se que uma GC bem gerida é fundamental para promover a inovação, que por sua vez impulsiona mudanças e estratégias organizacionais, contribuindo para uma gestão eficaz e eficiente. Embora tenha sido identificada uma quantidade limitada de estudos específicos sobre o tema, a amostra desta pesquisa permitiu compreender a relação entre GC e inovação em diferentes contextos organizacionais.

A maioria dos artigos analisados enfatizou que a inovação é uma ferramenta estratégica para a competitividade das organizações. Ficou evidente que o desempenho organizacional está intrinsecamente ligado à inovação, impulsionada por experiências, troca de informações e desenvolvimento dos colaboradores. Além disso, as Tecnologias da Informação e Comunicação desempenham um papel fundamental ao facilitar a retenção de conhecimento e aprimorar processos, permitindo que as organizações entreguem produtos e serviços inovadores aos clientes de forma mais ágil e eficiente.

As categorias estudadas também destacaram a importância da GC para uma tomada de decisão eficaz e um desenvolvimento contínuo das empresas. Ao promover a retenção de colaboradores, inovação e tecnologia, as organizações podem alcançar resultados positivos e se destacar no mercado competitivo atual.

Por fim, este estudo contribui para a compreensão da relação entre GC, inovação e competitividade organizacional, destacando sua importância para o desenvolvimento empresarial. Sugere-se que futuras pesquisas explorem profundamente essa relação, investigando outras correlações e ampliando o entendimento sobre como a GC pode ser mais bem utilizada para promover a inovação. Além disso, é importante investigar como as organizações podem superar os desafios na implementação da GC e identificar as melhores práticas para gerir o conhecimento de forma eficaz. Em suma, este estudo responde ao problema de pesquisa, fornecendo insights sobre como a GC tem sido pesquisada como ferramenta estratégica para uma gestão eficaz da inovação e oferecendo sugestões para que futuras pesquisas explorem ainda mais essa relação, especialmente no contexto tecnológico e em outras áreas pouco exploradas.

## REFERÊNCIAS

Abbas, J., Zhang, Q., Hussain, I., Akram, S., Afaq, A., & Shad, M. A. (2020). Sustainable innovation in small medium enterprises: the impact of knowledge management on organizational innovation through a mediation analysis by using SEM approach. *Sustainability*, *12*(6), 2407. https://doi.org/10.3390/su12062407.

Adams, G. L., & Lamont, B. T. (2003). Knowledge management systems and developing sustainable competitive advantage. *Journal of Knowledge Management*, *7*(2), 142-154. https://doi.org/10.1108/13673270310477342

Alasoini, T., Heikkilä, A., Ramstad, E., & Ylöstalo, P. (2007). Consultation as a new way to collect information in the Tykes Program: Preliminary Results. *Workpolitical Journal*, *2*(2007), 55-71.

Bardin, L. (1977). *L'analyse de contenu* (Vol. 69). Presses universitaires de France.

Barbieri, J. C., & Álvares, A. C. T. (2003). Inovações nas organizações empresariais. Em *Organizações inovadoras: estudos e casos brasileiros* (pp. 41-63). Fundação Getulio Vargas.

Braun, V., & Clarke, V. (2012). Thematic analysis. Em. H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf & K. J. Sher (Eds.), APA handbook of research methods in psychology, Vol. 2. Research designs: quantitative, qualitative, neuropsychological, and biological (pp. 57-71). American Psychological Association. https://doi.org/10.1037/13620-004

Brown, J. S. (1999). Sustaining the ecology of knowledge. *Leader to Leader*, *12*, 31-36.

Cervo, A. L., & Bervian, P. A. (2002). *Metodologia científica* (5ª ed.). Prentice Hall.

Costa, J., Pádua, M., & Moreira, A. C. (2023). Leadership styles and innovation management: What is the role of human capital? *Administrative Sciences*, *13*(2), 47. https://doi.org/10.3390/admsci13020047

Creswell, J. W., & Creswell, J. D. (2018). *Research design: qualitative, quantitative, and mixed methods approaches* (5ª ed.). Sage.

Damanpour, F., & Aravind, D. (2012). Managerial innovation: Conceptions, processes and antecedents. *Management and Organization Review*, *8*(2), 423-454. https://doi.org/10.1111/J.1740-8784.2011.00233.X

Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Harvard Business Press.

Easterby-Smith, M., & Lyles, M. (2003). Re-reading organizational learning: Selective memory, forgetting, and adaptation. *Academy of Management Perspectives*, *17*(2), 51-55. https://doi.org/10.5465/ame.2003.10025192

Fioravanti, V. L. S., Stocker, F., & Macau, F. (2023). Knowledge transfer in technological innovation clusters. *Innovation & Management Review*, *20*(1), 43-59. https://doi.org/10.1108/INMR-12-2020-0176

Fitriati, T. K., Purwana, D., & Buchdadi, A. D. (2020). Dynamic capabilities and SMEs performance: The mediating effect of innovation (Study of SMEs in Indonesia). Em 1st International Conference on Science, Health, Economics, Education and Technology (pp. 457-464). Atlantis Press.

Gil, A. C. (1991). *Como elaborar projetos de pesquisa*. Editora Atlas SA.

Gressler, L. A. (2003). *Introdução à pesquisa: projetos e relatórios*. Edições Loyola

Gemünden, H. G., Salomo, S., & Hölzle, K. (2007). Role models for radical innovations in times of open innovation. *Creativity and Innovation Management*, *16*(4), 408-421. https://doi.org/10.1111/j.1467-8691.2007.00451.x

Harrison, N., & Samson, D. (2002). *Technology management: Text and international cases*. McGraw Hill.

Jha, P., & Karn, B. (2019). Knowledge managements' relevance in supply chain management process of Indian e-commerce companies. *International Journal of Recent Technology and Engineering*, *8*(3), 1797-1805.

Jennex, M. E., & Zyngier, S. (2007). Security as a contributor to knowledge management success. *Information Systems Frontiers*, 9, 493-504. https://doi.org.10.35940/ijrte.C4612.098319

King, W. R. (2007). A research agenda for the relationships between culture and knowledge management. *Knowledge and Process Management*, *14*(3), 226-236. https://doi.org/10.1002/kpm.281

Lakatos, E. M. & Marconi, M. A. (2006). *Metodologia do trabalho científico: procedimentos básicos, pesquisa bibliográfica, projeto e relatório publicações e trabalhos científicos* (5ª ed.). Atlas.

Lacombe, F., & Heilborn, G. (2005). *Princípios e tendências*. Editora Saraiva.

Linder, J. C., Jarvenpaa, S., & Davenport, T. H. (2003, July 15). Toward an innovation sourcing strategy. *MIT Sloan Management Review*. https://sloanreview.mit.edu/article/toward-an-innovation-sourcing-strategy/https://doi.org/10.5325/transportationj.56.4.0477

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. Decision Support Systems, *15*(4), 251-266. https://doi.org/10.1016/0167-9236(94)00041-2

Moraes, M. T. B. de, Malischeski, K., & Dandolini, G. A. (2023). Gestão do conhecimento e inovação organizacional: uma revisão integrativa. *Perspectivas Em Gestão & Conhecimento*, *13*(esp), 146-161. https://periodicos.ufpb.br/ojs2/index.php/pgc/article/view/65569

Nonaka,I., & Takeuchi, H. (1997). *Criação de conhecimento na empresa: como as empresas japonesas geram a dinâmica da inovação*. Elsevier.

Polyakov, M., Khanin, I., Shevchenko, G., & Bilozubenko, V. (2023). Knowledge management in international companies: Specific features and information tools. *Financial and Credit Activity: Problems of Theory and Practice*, 3(50), 457-467. https://doi.org. https://doi.org/10.55643/fcaptp.3.50.2023.4061

Pompeo, D. A., Rossi, L. A., & Galvão, C. M. (2009). Revisão integrativa: etapa inicial do processo de validação de diagnóstico de enfermagem. *Acta Paulista de Enfermagem*, 22, 434-438. https://doi.org/10.1590/S0103-21002009000400014

Pontes, B. R. (2022). *Planejamento, recrutamento e seleção de pessoal* (Vol. 9). LTr Editora.

Quinn, J. B. (1992). *Intelligent enterprise: a knowledge and service based paradigm for Industry*. Simon and Schuster.

Raupp, F. M., & Beuren, I. M. (2006). Metodologia da pesquisa aplicável às ciências. Em: I. M. Beuren (Ed.), *Como elaborar trabalhos monográficos em contabilidade: teoria e prática* (pp. 76-97). Atlas

Senge, P. M. (2018). *A quinta disciplina: a arte e prática da organização que aprende*. Editora Best Seller.

Starkey, K. (1997). *Como as organizações aprendem*. Futura.

Sondhi, S.S., Salwan, P., Behl, A., Niranjan, S., & Hawkins, T. (2024). Evaluation of strategic orientation-led competitive advantage: the role of knowledge integration and service innovation. *Journal of Knowledge Management*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JKM-07-2023-0660.

Takeuchi, H., & Nonaka, I. (2008). Criação e dialética do conhecimento. Em: *Gestão do Conhecimento* (pp. 17-38). Bookman

Tidd, J., & Thuriaux-Alemán, B. (2016). Innovation management practices: cross-sectorial adoption, variation, and effectiveness. *R&D Management*, *46*(S3), 1024-1043. https://doi.org/10.1111/RADM.12199

Toffler, A. (1990). *Powershift: Knowledge, wealth and violence at the edge of the 21st century*. Bantam, London.

Valentim, M. L. P. (Org.) (2010). *Gestão, mediação e uso da informação*. Cultura Acadêmica

Venzin, M., Von Krogh, G., & Roos, J. (1998). Future research into knowledge management. Em: G. von Krogh, J. Roos & D. Kleine (Eds.), *Knowing in firms: Understanding, managing and measuring knowledge* (pp. 26-66). Sage. https://doi.org/10.4135/9781446280256

Wang, H. C., He, J., & Mahoney, J. T. (2009). Firm-specific knowledge resources and competitive advantage: the roles of economic-and relationship-based employee governance mechanisms. Strategic Management Journal, *30*(12), 1265-1285. http://www.jstor.org/stable/27735491

Wang, C. L., & Ahmed, P. K. (2004). The development and validation of the organisational innovativeness construct using confirmatory factor analysis. European Journal of Innovation Management, *7*(4), 303-313. https://doi.org/10.1108/14601060410565056

Whittemore, R., & Knafl, K. (2005). The integrative review: updated methodology. *Journal of Advanced Nursing*, *52*(5), 546-553. https://doi.org/10.1111/j.1365-2648.2005.03621.x

Wiig, K. M. (1994). *Knowledge management foundations: thinking about thinking-how people and organizations represent, create, and use knowledge*. Schema Press.

# UL-Keystroke: A Web-Based Keystroke Dynamics Dataset

Aron Lo Li
20160795@aloe.ulima.edu.pe
https://orcid.org/0009-0006-2616-3950
Universidad de Lima

Juan Gutiérrez-Cárdenas
jmgutier@ulima.edu.pe
https://orcid.org/0000-0003-2566-4690
Universidad de Lima

Victor H. Ayma
vh.aymaq@up.edu.pe
https://orcid.org/0000-0002-0284-2610
Universidad del Pacífico

Received: March 13th, 2024 / Accepted: May 23rd, 2024
doi: https://doi.org/10.26439/interfases2024.n19.7009

ABSTRACT. Keystroke dynamics-based authentication systems identify individuals by analyzing their keystroke patterns when interacting with input devices such as a computer keyboard. Within the fields of Statistics and Machine Learning, several research studies have applied different techniques for recognizing keystroke patterns. This work proposes the creation of a dataset and a methodology that would allow users to capture typing patterns from students at a university in Lima, Peru, using a cloud environment and their personal devices. The cloud architecture used for the implementation and deployment of the web tool will be explained in detail. The result of this work is a dataset containing participant information, records of their keystroke patterns, and additional metadata from their web browsers, which could be used to enrich further studies. Moreover, in addition to the captured raw data, some keystroke dynamics features were generated and made available along with the dataset to facilitate the development of classification models. The dataset and methodology presented in this article can be used by other researchers to enhance existing keystroke dynamics recognition systems.

KEYWORDS: keystroke dynamics / machine learning / dataset

# UL-KEYSTROKE: UN CONJUNTO DE DATOS DE DINÁMICA DE TECLADO BASADO EN LA WEB

RESUMEN. Los sistemas de autenticación basados en la dinámica de teclado identifican a las personas analizando sus patrones de tecleo cuando interactúan con dispositivos de entrada, como un teclado de computadora. En los campos de Estadística y Aprendizaje Automático, existen varios estudios de investigación que han aplicado diferentes técnicas para el reconocimiento de patrones de tecleo. En este trabajo, se propuso la creación de un conjunto de datos, así como una metodología que permitiría a los usuarios capturar patrones de tecleo de estudiantes pertenecientes a una universidad en Lima, Perú, a través de un entorno en la nube y desde sus propios dispositivos. La arquitectura en la nube utilizada para la implementación y despliegue de la herramienta web será explicada en detalle. El resultado de este trabajo es un conjunto de datos con información de los participantes, registros de sus patrones de tecleo y metadatos adicionales de los navegadores web de los participantes que podrían usarse para enriquecer futuros estudios. Además, junto con los datos sin procesar capturados, se generaron algunas características de la dinámica de tecleo y se pusieron a disposición junto con el conjunto de datos para facilitar la generación de modelos de clasificación. El conjunto de datos y la metodología presentados en este artículo pueden ser utilizados por otros investigadores para mejorar los sistemas de reconocimiento de dinámica de teclado actuales.

PALABRAS CLAVE: dinámica de teclado / aprendizaje automático / conjunto de datos

## 1. INTRODUCTION

This article aims to comprehensively detail all aspects related to a dataset, encompassing the collection methodology, the analytical procedures conducted, the characteristics of the data, the value added by the creation of this dataset, and its potential limitations. These points will be discussed in the following sections.

## 2. SPECIFICATIONS TABLE

| Subject | Applied Machine Learning |
|---|---|
| Specific subject area | Keystroke Dynamics Authentication |
| Type of data | Raw data:<br>• Records of keystroke patterns from college students (participants) and metadata of the students' web browsers used to register the keystroke patterns.<br>• List of participants along with their personal data, as well as the chosen username and password.<br>Processed data:<br>• Time vectors generated from raw data. |
| Data collection | A web logging application, written in JavaScript, was implemented and deployed in a cloud environment. The participants were invited to cooperate virtually in the keystroke recording sessions using their own computers, where the application captured the timestamps of each pressed and released key when typing their login credentials.<br>The participants in each session were asked to perform three authentication cycles: one cycle involving the use of their own credentials and the remaining two using the credentials of randomly selected participants.<br>At the end of each session, the captured records were automatically sent to a server for storage in a non-relational database. |
| Data source location | Cloud-based collection (Heroku, MongoDB Atlas) |
| Data accessibility | Repository name: UL-Keystroke Dynamics Dataset<br>Data identification number (doi): 10.17632/9cg3c8jkh8.1<br>Direct URL to data: https://data.mendeley.com/preview/9cg3c8jkh8?a=f4e49f3e-b689-4b6c-95a3-3bf5207d1935 |
| Related research article | None |

## 3. VALUE OF THE DATA

- The scientific community has very few web-based keystroke dynamics datasets that are publicly accessible, have undergone a rigorous collection process, and have been created in uncontrolled environments.
- The published dataset contains both records of users' typing sessions as well as some metadata captured from their web browser environment. This dataset enables

researchers to leverage the gathered data to produce keystroke-related features, enhancing the effectiveness of the authentication models on which they are working.

- The generated dataset has unique characteristics compared to other public datasets. For instance, participants entered two values in the web application: a username and a password. Both values and subsequent keystrokes were captured. Unlike other datasets where participants input imposed fixed-length passwords, here the participants were given the choice to register their own password. Thus, two user groups were defined at the time of registration: some participants could choose their password but with a fixed length, while others had no length restriction.
- This keystroke dynamics dataset will allow researchers to strengthen ongoing core authentication systems by adding an additional layer of security through the application of Deep Learning and Machine Learning models or similar, which can be applied in critical systems such as banking, medical care, military, etc.

## 4. BACKGROUND

In the field of automatic authentication, several works have focused extensively on generating keystroke dynamics models. However, it is also essential to have a dataset with relevant information about keystroke patterns to generate a high-quality authentication model. In the literature, there are only a few published datasets available for researchers (Giot et al., 2009; Killourhy et al., 2009). This work aims to provide the public with access to a keystroke dynamics dataset generated from a cloud-based web approach, as well as the methodology used to construct it. This will enable future researchers to understand how to generate their own keystroke dynamics datasets and serve as a reference for conducting their collection.

## 5. DATA DESCRIPTION

The dataset comprises a collection of keystroke samples from various participants who voluntarily took part in the data collection process (experiment). These samples were captured using a web application, implemented and deployed in a cloud environment. This setup allowed the participants to conduct the tests on their computers from any location, ensuring that the experiment took place in real-world environments.

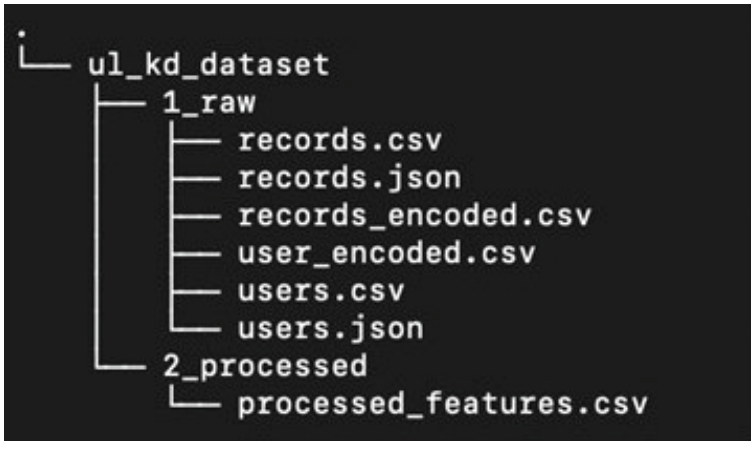
The dataset contains 10 994 keystroke patterns captured between October 2nd to November 17th, 2020. During the data collection, 66 participants were registered, with 59 % being male and 41 % female. The participants' ages ranged from 19 to 23 years old, situating the population within the young adult generation.

The dataset files are stored in Mendeley Data cloud-based communal repository and are organized hierarchically within the main folder, "ul_kd_dataset," as illustrated

in Figure 1. This folder contains two subfolders, namely "1_raw" and "2_processed". The former includes raw data exported directly from the MongoDB Atlas database used by the web application, while the latter consists of processed data derived from raw data. The "1_raw" subfolder contains two main files: "users" and "records." The "users" file includes data of the participants registered for the experiments, who signed up and completed a form on the web portal. The "records" file encompasses all keystroke patterns produced by the participants. Within the "2_processed" subfolder, there is a file containing time vectors (referred to as features), which can be directly used for training and testing classification or authentication models. Each entity mentioned above is available in both CSV and JSON formats. A more detailed description of the content of each file is presented below.

**Figure 1**

*Dataset Structure*




The "users" file, as mentioned before, contains information related to the participants, including personal data, typing style requested during registration, as well as certain metadata captured during the account creation. Table 1 presents the file fields in detail.

**Table 1**

*Structure of the users_encoded.csv File*

| Field | Description |
|---|---|
| _id | System-generated unique user identifier. |
| name | User's name. |
| lastname | User's last name. |

(*continues*)

(*continued*)

| Field | Description |
|---|---|
| age | User's age. |
| email | User's email. |
| username | Username chosen by the user to perform the login tests on the system. |
| password | Password chosen by the user to perform the login tests on the system. |
| dni | National identity card. |
| isImposedPassword | Boolean value indicating whether a fixed password was imposed at the time of registration. "True" indicates a fixed password; "False" indicates a variable password. |
| genre | User's gender. |
| handedness | User's dominant hand. |
| handDisease | Boolean value indicating if the user is experiencing motor issues with their hands. |
| date | User's registration date in the system. |
| ipAddress | User's IP address at the time they registered in the system. |
| userAgent | Browser agent used to log into the system. |

The "records" file includes keystroke samples made by users during the conducted experiments. Each pressed and released key was recorded in the dataset along with the associated timestamp at the time the participants entered their "username" and "password". Additionally, metadata of data collection sessions was stored, providing context for the tests conducted by each participant. Table 2 presents the details of each field in the file.

**Table 2**

*Structure of the records_encoded.csv File*

| Field | Description |
|---|---|
| _id | System-generated unique record identifier. |
| rawUsernameKeydown | A collection of keystroke events recorded during the input of the username, specifically when the key was pressed, including both the pressed key and the corresponding timestamp. |
| rawUsernameKeyup | A collection of keystroke events recorded during the input of the username, specifically when the key was released, including both the released key and the corresponding timestamp. |
| rawPasswordKeydown | A collection of keystroke events recorded during the input of the password, specifically when the key was pressed, including both the pressed key and the corresponding timestamp. |
| rawPasswordKeyup | A collection of keystroke events recorded during the input of the password, specifically when the key was released, including both the released key and the corresponding timestamp. |

(*continues*)

(*continued*)

| Field | Description |
|---|---|
| belongedUserId | Identity of the user who created the credentials displayed in the session. |
| performedUserId | Identity of the user who owns the records of the current sample. |
| date | Date the sample record was created. |
| sessionIndex | The user must perform three authentication cycles per session. This integer value indicates the index. |
| valid | Users could make mistakes when typing. This Boolean value indicates whether the user correctly typed the credentials displayed on the screen the first time. |
| username | Username of the associated record. |
| password | Password of the associated record. |
| ipAddress | User's IP address at the time they registered in the system. |
| userAgent | Browser agent used to log into the system. |
| token | Unique token generated within the user's browser the first time they perform the login tests. This token can be used to detect if the same computer is used by the user to perform the tests. |

In the "processed_features" file, features are generated from the difference between the captured and stored keystroke events and the raw data. Four time vectors were generated, namely ppTime, rrTime, prTime, and rpTime. Additionally, a final vector, which is a concatenation of these previous ones, is included. These vectors can be used to train and develop keystroke recognition models. Table 3 presents a detailed structure of the file.

**Table 3**

*Structure of the processed_features.csv File*

| Field | Description |
|---|---|
| _id | System-generated unique processed_features identifier. |
| belongedUserId | Identity of the user who created the credentials displayed in the session. |
| performedUserId | Identity of the user who owns the records of the current sample. |
| valid | Users could make mistakes when typing. This Boolean value indicates whether the user correctly typed the credentials displayed on the screen the first time. |
| password | Password of the associated record. |
| ppTime | Time vector generated by the difference in timestamps between two sequentially pressed keys. |
| rrTime | Time vector generated by the difference in timestamps between two sequentially released keys. |
| prTime | Time vector generated by the difference in timestamps between one pressed key and one released key. |

(*continues*)

(*continued*)

| Field | Description |
|---|---|
| rpTime | Time vector generated by the difference in timestamps between one released key and one pressed key. |
| vector | Concatenation of the four vectors (ppTime + rrTime + prTime + rpTime). |
| passLen | Length of the password. |
| vectorLen | Length of the vector created. |

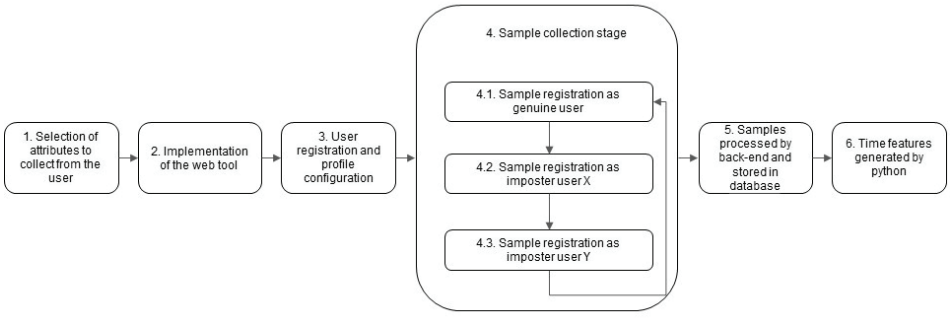
## 6. EXPERIMENTAL DESIGN, MATERIALS AND METHODS

### 6.1 Experimental Design

The process to generate the dataset of keystroke samples is detailed in Figure 2. The first phase involved implementing the web tool from scratch and identifying all attributes that could be collected from users and that would be relevant to keystroke dynamics models. Once these attributes were mapped, the second phase involved building a system with a front-end and back-end to enable the collection of these attributes. In the third phase, users who wished to participate voluntarily in the experiment were recruited and instructed to register on the web portal by filling out a form. After registration, users were required to access the web page on alternate days and complete a series of login tests displayed on the screen throughout the session. During these tests, every key pressed and released was captured as they entered the username and password. At the end of the session, all captures, which were stored locally, were sent to the service for subsequent storage. The collection phase lasted from October 2nd to November 17th, 2020, involving 66 users and recording a total of 10 994 records. The final stage involved generating the features that could be used in keystroke dynamics models.

**Figure 2**

*Methodology Design*

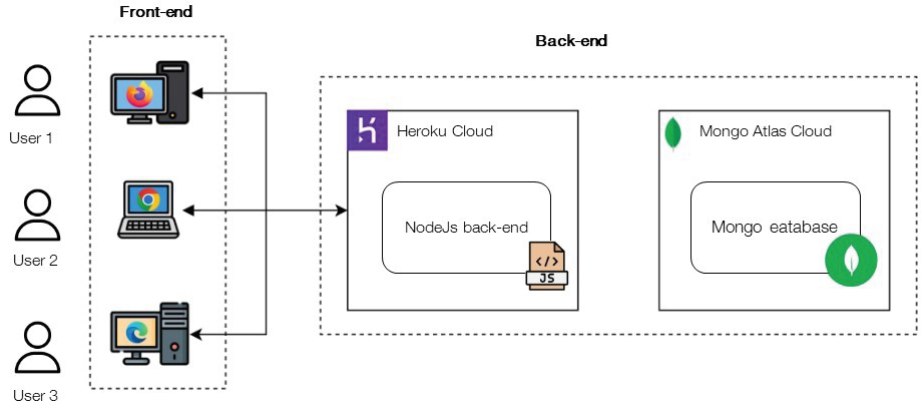


### 6.2 Materials

The collection system was implemented using web-based technology. The built solution, which includes both the front-end and back-end, is shown Figure 3.

**Figure 3**

*Architecture of the Web Tool*




The front-end, the visual layer where the end user interacts with the platform via a browser, was implemented using templates developed with the PUG library (version 3.0.0), JavaScript, and CSS. All screens incorporate these three components and are rendered on the back-end, then sent to the browser each time the user accesses the website URL. The logic for capturing keystrokes in the browser was written in JavaScript, the structural aspects of the web page were handled in HTML/PUG, and the visual design was crafted with CSS.

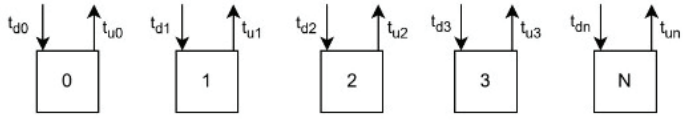
The back-end was developed in JavaScript using the Express library (version 4.17.1) and deployed on the Heroku cloud platform in a Node.js runtime environment. Different endpoints were created for each screen outlined in the user flow designed for keystroke capture sessions. When a user completes the capture flow, the front-end sends all records to the back-end, where they are received, processed, and stored in the database. Interaction with the non-relational MongoDB Atlas database was facilitated by the Mongoose library (version 5.10.5). Additionally, an email service using Mailgun was implemented to notify users of the proper storage of their session data.

Due to the unstructured nature of the keystroke capture records, a document-oriented non-relational database was chosen for storage. The MongoDB Atlas cloud was used to deploy the MongoDB Atlas database. Data was exported to monitor progress and maintain backups. The document structure used in the database is described in the Data Description section. The code used for the solution is provided in Appendix 6.

After the sampling period, various features were generated to form time vectors used for model training, including: the "down-down key feature," "down-up key feature," "up-down key feature," "up-up key feature," "hold time," and "total time." The first four features involve latency between different pressed keys, either when they are pressed ("down") or released ("up"). "Hold time" refers to the duration a single key is pressed, while "total time" encompasses the entire time taken to type a word. Classification models cannot directly process the raw data generated by the tool as it only captures the timestamp when a key is pressed and released. To make this dataset usable, the raw data is processed to generate the aforementioned features. Figure 4 illustrates the generation of these time vectors, capturing both the "down" (td) and "up" (tu) events for each key. The latencies or differences between these events allow for the construction of user-specific vectors that will subsequently be used in the models. These features were generated using a Python script, which is also available in the GitHub repository mentioned in the Appendix 7.

**Figure 4**

*Procedure to Generate Time-Based Latency Features*



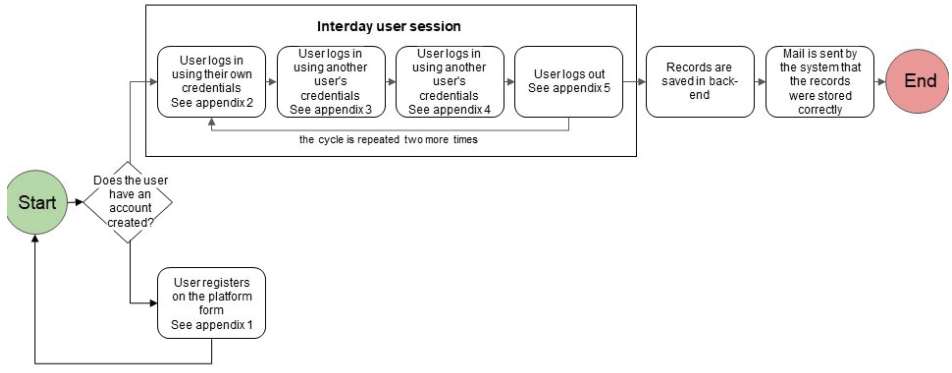

| | | | | |
|---|---|---|---|---|
| Feature down-down | $t_{d1} - t_{d0}$ | $t_{d2} - t_{d1}$ | $t_{d3} - t_{d2}$ | $t_{dn} - t_{d(n-1)}$ |
| Feature down-up | $t_{u0} - t_{d0}$ | $t_{u1} - t_{d1}$ | $t_{u2} - t_{d2}$ | $t_{un} - t_{dn}$ |
| Feature up-up | $t_{u1} - t_{u0}$ | $t_{u2} - t_{u1}$ | $t_{u3} - t_{u2}$ | $t_{un} - t_{u(n-1)}$ |
| Feature up-down | $t_{d1} - t_{u0}$ | $t_{d2} - t_{u1}$ | $t_{d3} - t_{u2}$ | $t_{dn} - t_{u(n-1)}$ |

### 6.3 Methods

Figure 5 shows the interaction flow proposed during the experimental design phase. The web tool was implemented to ensure that users could follow the proposed flow. The complete screen flow is detailed in the appendices.

**Figure 5**

*User Interaction Flowchart*




Each week, users were required to complete at least two sessions, spaced on alternate days, to prevent "over-familiarity" with typing patterns while still allowing for user variability to be recorded. Each session consisted of three tasks, which the user performed three times, as depicted in Figure 5. In the first task, the legitimate user logged in to the system with their own username and password (either self-selected or assigned, depending on the group). Subsequently, they proceeded to a second and third screen where they typed the username and password of another randomly selected user three times. Once this flow was completed, the user logged out and repeated the process two more times. It is worth mentioning that for each entry to be considered valid, the user had to input the credentials correctly on the first attempt, with no corrections allowed, such as using the backspace key to retype. The session ended when the tool recorded the required number of correct samples according to the proposed methodology. Regarding invalid samples, they were also stored as part of the dataset, as the errors could be considered features related to the user's identity and could be useful in future analyses. The data collection period lasted six weeks; however, the tool continued to capture samples beyond this period.

## 7. LIMITATIONS

In this experiment, part of the methodology involved sequentially typing the phrase that appeared on the screen three times and then repeating this process two more times. While this ensured more than 10 records per session, the experimental design, which included multiple writing sessions per user, may have introduced a bias in the data due to the repetitive nature of the tasks. This might lead to users learning and adjusting their writing patterns as the experiment progresses.

## 8. ETHICS STATEMENT

Following the ethical publishing guidelines provided by Elsevier and *Data in Brief*, the following key ethical aspects were considered**:**

> **Human Studies:** All users participated voluntarily in the experiment and were informed about both the session dynamics and the intended use of the collected data. They provided a consent by filling out a form.

## 9. DATA AVAILABILITY

Data resources can be found in the following link: https://data.mendeley.com/preview/9cg3c8jkh8?a=f4e49f3e-b689-4b6c-95a3-3bf5207d1935

## 10. CREDIT AUTHOR STATEMENT

Aron Lo Li was responsible for the conceptualization and execution of the experiments. Juan Gutiérrez-Cárdenas played an active advisory role across all project stages. Victor H. Ayma assisted in shaping and enhancing the proposed methodology.

## 11. ACKNOWLEDGEMENTS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We thank the editor, reviewers, and all participants for their involvement and contributions to this research endeavor.

## 12. DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationship that could have appeared to influence the work reported in this paper.

## 13. REFERENCES

Giot, R., El-Abed, M., & Rosenberger, C. (2009). GREYC keystroke: A benchmark for keystroke dynamics biometric systems. In *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems (BTAS 2009),* Washington D.C., United States, 1–6. https://doi.org/10.1109/BTAS.2009.5339051

Killourhy, K. S., & Maxion, R. A. (2009). Comparing anomaly-detection algorithms for keystroke dynamics. In *2009 IEEE/IFIP International Conference on Dependable Systems & Networks*, Lisbon, Portugal, 125–134. https://doi.org/10.1109/DSN.2009.5270346

## 14. APPENDICES

**Appendix 1**

User Account Creation Screen


**Appendix 2**

Legitimate User Login Screen (Task 1)

**Appendix 3**

First Screen for Logging in as an Imposter User (Task 2)


**Appendix 4**

Second Screen for Logging in as an Imposter User (Task 3)

**Appendix 5**

Third Screen for Logging in as an Imposter User (Task 3)


**Appendix 6**

Source Code of the Web Tool Implemented for Dataset Generation
https://github.com/aronlo98/tesis-keylogger

**Appendix 7**

Source Code of the Keystroke Dynamics Models
https://github.com/aronlo98/tesis

# DATOS DE LOS AUTORES


**ALEJANDRA MORALES RAMÍREZ**

Es doctora en Sistemas Computacionales. Desde mayo del 2009 se encuentra integrada como profesora de tiempo completo en el Centro Universitario Ecatepec de la Universidad Autónoma del Estado de México, donde imparte las asignaturas de Comunicación entre Computadoras, Protocolos de Red, Simulación de Redes WAN, Bases de Datos, Investigación I y Temas Selectos de Computación. Cuenta con perfil deseable PRODEP y ha publicado artículos en congresos y revistas científicas referentes a temas de educación, minería de procesos, desarrollo de software y sistemas de información, así como de circuitos analógicos. Conjuntamente, ha dirigido tesis de nivel licenciatura y maestría.

**RODOLFO GARCÍA LOZANO**

Es doctor en Ingeniería Eléctrica del Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional, Ciudad de México. Actualmente, es profesor de tiempo completo del Centro Universitario Ecatepec de la Universidad Autónoma del Estado de México y líder del cuerpo académico TICs y Dispositivos Electrónicos con registro y nivel consolidado ante la Secretaría de Educación Pública. Es miembro del Sistema Nacional de Investigadores de México.

**JUAN DE JESÚS AMADOR REYES**

Es doctor en Ciencias de la Computación. Actualmente, es profesor de tiempo completo del Centro Universitario Ecatepec de la Universidad Autónoma del Estado de México, en donde imparte clases en la licenciatura de Informática Administrativa, Ingeniería en Computación y en la Maestría en Ciencias de la Computación. Cuenta con perfil deseable PRODEP y ha publicado artículos en congresos y revistas científicas referentes a temas de realidad virtual, inteligencia artificial, computación gráfica y minería de procesos.

**CUAUHTÉMOC HIDALGO CORTÉS**

Es doctor en Sistemas Computacionales. Desde 2008 se desempeña como profesor de tiempo completo del Centro Universitario Ecatepec de la Universidad Autónoma del Estado de México, en donde imparte clases en la licenciatura de Ingeniería en Computación y en la Maestría en Ciencias de la Computación. Es integrante del cuerpo académico TICs y Dispositivos Electrónicos con registro y nivel consolidado ante la Secretaría de Educación Pública. Simultáneamente, ha dirigido tesis de nivel licenciatura y maestría.

**ROSA FLOR GOMEZ RISCO**

Es estudiante de doctorado en la mención Ciencias Matemáticas en la Universidad Nacional de Piura. Se graduó como licenciada en Matemática por la Universidad Nacional de Piura, Perú. Su área de interés se centra en base de datos.

**MADELEINE GILLIAN RABINES FLOREANO**

Es estudiante de maestría en la mención Administración de Negocios en la Universidad Nacional de Trujillo, Perú. Se graduó en Informática por la Universidad Nacional de Trujillo, Perú. Sus áreas de interés se enfocan en base de datos e inteligencia artificial.

**LOURDES RAMÍREZ CERNA**

Es magíster en Ciencia de la Computación por la Universidade Federal de Ouro Preto, Brasil, y es estudiante de doctorado en Ciencias e Ingeniería en la Universidad Nacional de Trujillo, Perú. Se graduó en Informática por la Universidad Nacional de Trujillo, Perú. Actualmente es docente en la Universidad de Lima. Sus áreas de interés son la inteligencia artificial, aprendizaje profundo, aprendizaje automático y optimización combinatoria.

**ZORAIDA MAMANI RODRIGUEZ**

Es doctora en Ingeniería por la Universidad Nacional Federico Villarreal, magíster en Computación e Informática por la Universidad Nacional Mayor de San Marcos, e ingeniera de sistemas. Es docente de la Escuela de Ciencia de la Computación de la Universidad Nacional de Ingeniería y coordinadora del grupo de investigación Ingeniería Web de la Facultad de Ingeniería de Sistemas e Informática de la UNMSM. Es investigadora en las áreas de ingeniería de software, aprendizaje automático, ciencia de datos y GovTech.

**MICHAEL DORIN**

Cuenta con un Ph. D. en Ciencias de la Computación por la Universidad de Wurzburg. Tiene más de 30 años de experiencia en desarrollo de software y ha trabajado en diversos entornos de ingeniería. Su experiencia incluye trabajos en ingeniería relacionados con comunicaciones de seguridad pública, dispositivos médicos (marcapasos), telefonía y navegación de aeronaves.

**SERGIO MONTENEGRO**

Cuenta con un doctorado y una maestría en Ciencias de la Computación por la Universidad Técnica de Berlín. Ha estado programando satélites durante los últimos 20 años. Actualmente, es profesor de tecnología de la información aeroespacial en la Universidad de Würzburg (Alemania).

**MARCELO LÓPEZ-NOCERA**

Es doctorando en Ingeniería en el Programa de Doctorado de la Universidad Nacional de Lomas de Zamora y magíster en Ingeniería en Sistemas de Información por la Facultad Regional Buenos Aires de la Universidad Tecnológica Nacional. Es docente en diversas materias de dicha universidad. Es integrante del Grupo GEMIS de investigación.

**MARÍA F. POLLO-CATTANEO**

Es doctora en Ciencias Informáticas por la Universidad Nacional de la Plata. Es investigadora categoría A de la Universidad Tecnológica Nacional y es integrante del Grupo GEMIS de investigación. Es titular de cátedra en distintas materias en diversas universidades. Ha sido ganadora de diversos premios en su trayectoria.

**FRANCISCO REDELICO**

Es doctor en Ingeniería por la Universidad de Buenos Aires. Es titular de cátedra en distintas materias en diversas universidades. Es investigador adjunto en el Instituto de Medicina Traslacional e Ingeniería Biomédica, Consejo Nacional de Investigaciones Científicas y Técnicas.

**ERLY GALIA VILLARROEL ENRIQUEZ**

Es licenciada en Ingeniería de Sistemas por la Universidad de Lima, obtenida en 2023. En 2022, participó en el Congreso Internacional de Ingeniería de Sistemas celebrado en el campus de la Universidad de Lima, donde presentó un afiche que resumía una revisión de literatura sobre ofuscación de malware. Sus intereses de investigación se centran principalmente en las aplicaciones del aprendizaje automático y el aprendizaje profundo para abordar los desafíos actuales.

**JUAN GUTIÉRREZ-CÁRDENAS**

Obtuvo su doctorado en Informática en la subdisciplina de Bioinformática en la Universidad de Sudáfrica. También tiene una maestría en Informática-Bioinformática de la Universidad de Helsinki y un Diplomado Superior en Informática de la Universidad de Witwatersrand. Es profesor a tiempo parcial en la Universidad de Lima en el área de Ingeniería de Software. Ha sido un activo defensor del desarrollo de habilidades de inves-

tigación en estudiantes de pregrado, lo que ha dado lugar a una variedad de publicaciones revisadas por pares realizadas por estudiantes en la carrera en la que actualmente trabaja. También se desempeña como evaluador de programas ABET y es miembro de la Sociedad Peruana de Informática. Entre sus intereses de investigación se encuentran los métodos de aprendizaje automático aplicados a la bioinformática, el procesamiento de lenguaje natural y la educación en ciencias de la computación.

### ERICK LEONEL GARCÍA IBÁÑEZ

Es magíster en Informática Empresarial por la Universidad Politécnica de San Petersburgo Pedro el Grande, Rusia. Es ingeniero informático por la Universidad Nacional de Trujillo y posee certificaciones en Scrum, ITIL, finanzas corporativas y gestión de proyectos. Cuenta con una amplia experiencia en el desarrollo de soluciones de TI y análisis de datos, donde destaca como un profesional apasionado y altamente capacitado en ofrecer soluciones innovadoras para problemas comerciales complejos. Trabajó 10 años como analista y desarrollador de software en la Compañía Minera Poderosa. Además, tiene más de 5 años de experiencia como inversor minorista en mercados financieros. Actualmente, se desempeña como consultor de transformación digital en London Consulting Group, una empresa con sede en México que desarrolla proyectos de innovación y gestión del cambio en Latinoamérica.

### EDUARDO RODRIGO WONG LEON

Es estudiante de la Universidad Católica Sedes Sapientiae con una pasión destacada por el desarrollo de software y el análisis de datos. Actualmente, se encuentra inmerso en su formación universitaria, donde explora activamente las tecnologías de vanguardia aplicadas sobre estas disciplinas.

### MARCO ANTONIO CORAL YGNACIO

Es candidato a doctor en Ingeniería de Sistemas por la Universidad Nacional Mayor de San Marcos y magíster en Ciencias en Ingeniería de Sistemas y Computación. Es experto en temas de transformación digital en universidades. Ha sido responsable técnico del proyecto Cero Papel en la Universidad Nacional Mayor de San Marcos, donde trabajó con sistemas de gestión documentaria con firma digital, implementación de documentos digitales tales como grados, títulos y otros, integración de sistemas y generación de servicios para la universidad. Ha sido jefe de la Unidad de Tecnología Educativa, jefe de la Unidad de Servidores y Sistemas de Información de la Red Telemática-UNMSM y jefe de la Oficina de Calidad y Acreditación Académica de la FISI-UNMSM, responsable de la acreditación de los programas de posgrado e implementación de la norma ISO 9001:2015. Es investigador en temas de transformación digital con publicaciones en revistas científicas y evaluador de proyectos de innovación tecnológica para CONCYTEC.

**ELAINE RODRIGUES KOLLER**

Es abogada con posgrado en Ciencias Penales y maestría en Licitaciones y Contratos. Actualmente, está cursando una maestría en Ingeniería y Gestión del Conocimiento en la Universidad Federal de Santa Catarina. Se desempeña en el área de Innovación y es coautora de libros, con artículos publicados en revistas, congresos y periódicos nacionales.

**PAULO DE MOURA**

Posee un posgrado en Gestión Estratégica de Personas (2011) y en Gestión en Tutoría en EAD por la UNIASSELVI (2022). Actualmente, está cursando una maestría en Gestión de la Ingeniería del Conocimiento en la Universidad Federal de Santa Catarina (PPGEGC/ENGIN). Se graduó en Administración por la Universidad Estácio de Sá (2004). Es profesor universitario, conferencista y consultor empresarial. Se desempeña como profesor en los cursos de Gestión en UNIASSELVI (SC) y es miembro de Ingeniería de la Integración y Gobernanza del Conocimiento para la Innovación (ENGIN). Es autor de capítulos de libros y tiene artículos publicados en revistas internacionales, congresos y periódicos nacionales.

**PATRÍCIA DE SÁ FREIRE**

Es doctora en Ingeniería y Gestión del Conocimiento (2013) y magíster en EGC (2010) por el Programa de Posgrado en Ingeniería y Gestión del Conocimiento de la Universidad Federal de Santa Catarina. Tiene una licenciatura en Educación, con especialización en Tecnologías de la Educación, por la PUC/RJ (1986). Es especialista en Marketing por la ESPM/RJ (1987) y en Psicopedagogía por la UCB/RJ (2006). Actualmente, es profesora del Departamento de Ingeniería del Conocimiento e investigadora del PPGEGC de la UFSC. Lidera el laboratorio de Ingeniería de la Integración y Gobernanza del Conocimiento para la Innovación (ENGIN), y es miembro de los Grupos IGTI (Núcleo de Inteligencia, Gestión y Tecnología para la Innovación/UFSC) y KLOM (Interdisciplinar en Conocimiento, Aprendizaje y Memoria Organizacional/UFSC). Es autora de libros y capítulos de libros científicos, destacando la coautoría de capítulos en la obra Interdisciplinariedad en Ciencia, Tecnología e Innovación, que obtuvo el 2.º lugar en el Premio Jabuti en 2011.

**ARON LO LI**

Es ingeniero de sistemas por la Universidad de Lima. Con más de tres años de experiencia en las áreas de desarrollo de software y big data aplicado a empresas. Se encuentra interesado en seguir desarrollando proyectos en el ámbito analítico, tanto en aprendizaje automático como en aprendizaje profundo.

**JUAN GUTIÉRREZ-CÁRDENAS**

Obtuvo su Ph. D. en Ciencias de la Computación en la subdisciplina de Bioinformática de la Universidad de Sudáfrica. También cuenta con una maestría en Ciencias de la Computación-Bioinformática de la Universidad de Helsinki y un diplomado en Ciencias de la Computación de la Universidad de Witwatersrand. Es profesor a tiempo parcial en la Universidad de Lima en el área de Ingeniería de Software. Ha sido un activo defensor del desarrollo de habilidades de investigación en estudiantes de pregrado, lo que ha resultado en una variedad de publicaciones revisadas por pares de estudiantes en la carrera en la que actualmente trabaja. También se desempeña como evaluador de programas ABET y es miembro de la Sociedad Peruana de Computación. Es reconocido como investigador de nivel 4 según lo establecido por Renacyt. Entre sus intereses de investigación se encuentran los métodos de aprendizaje automático aplicados a la bioinformática, el procesamiento del lenguaje natural y la educación en ciencias de la computación.

**VICTOR H. AYMA**

Cuenta con Ph. D. y maestría en Ingeniería Eléctrica por la Pontificia Universidad Católica de Río de Janeiro, Brasil, en 2021 y 2015, respectivamente. Recibió su título de licenciado en Ingeniería Electrónica por la Universidad Nacional de San Antonio Abad del Cusco, Perú, en 2012. Es miembro del Instituto de Ingenieros Eléctricos y Electrónicos (IEEE). Actualmente, es profesor asistente en el Departamento de Ingeniería de la Universidad del Pacífico, en Perú. Sus intereses de investigación incluyen visión por computadora, sensores remotos, aprendizaje profundo y aprendizaje automático.

# POLÍTICA EDITORIAL

## ENFOQUE Y ALCANCE

Interfases es una revista digital, gestionada por la Universidad de Lima, para la publicación de investigaciones originales en áreas temáticas relacionadas con la ciencia de la computación, ingeniería del software, sistemas de información, tecnologías de información, ciberseguridad, ciencia de datos y áreas afines. Se publican artículos científicos y avances de investigación, siempre que cumplan con el proceso de revisión por pares. La revista Interfases está indizada en CrossRef, Dialnet, Latindex y DOAJ: Directory of Open Access Journals, y se publica dos veces al año: la primera en julio y la segunda en diciembre. Sin embargo, a partir de julio del 2021, los manuscritos individuales se publicarán tan pronto como estén listos, añadiéndolos progresivamente al contenido de la edición en curso en la modalidad de publicación continua. Los artículos en publicación continua siguen el proceso de revisión por pares, y ya se pueden citar utilizando el año de publicación y el DOI.

## PROCESO DE REVISIÓN POR PARES

Los manuscritos originales e inéditos enviados a la revista Interfases siguen un proceso de evaluación en dos etapas. En la primera, el editor examina el contenido para determinar si el manuscrito está alineado con el alcance y si los autores han seguido las directrices. Si el manuscrito no es aceptado, se devuelve al autor con el detalle de las razones que motivan la decisión tomada por el editor.

Si el manuscrito es aceptado, el editor envía el trabajo a revisores externos expertos en el tema de investigación. Esta segunda evaluación corresponde a una revisión por pares doble ciego, donde el autor y los revisores son anónimos.

El revisor evalúa el contenido del manuscrito y, basándose en su experiencia y conocimiento, adopta una de las siguientes recomendaciones:

1. El manuscrito es aceptado sin cambios o con cambios mínimos.

2. El manuscrito se acepta, a condición de realizar cambios importantes, de acuerdo con las observaciones del revisor. La versión corregida del manuscrito debe ser aprobada en una segunda revisión.

3. El manuscrito no se acepta por las contribuciones limitadas del estudio u otras consideraciones informadas por el revisor.

Con base en los comentarios de los revisores, el editor informa la decisión al autor correspondiente, quien tiene hasta 30 días para realizar los cambios al manuscrito (recomendaciones 1 y 2) o argumentar por qué no se acepta (recomendación 3).

Una vez que los revisores reciben el manuscrito corregido, tienen hasta 20 días para informar el resultado de la nueva evaluación; posteriormente, emiten su recomendación final. Una vez que el editor recibe la segunda ronda de revisiones, toma una decisión para publicar el manuscrito y luego se le notifica al autor correspondiente.

Cualquier objeción del autor respecto de la decisión del editor o hacia los comentarios de los revisores será resuelta por el Comité Editorial como instancia final.

La revista se adhiere a los criterios establecidos por el Guidelines on Good Publication Practice del Committee on Publication Ethics (COPE), el cual establece las sanciones

# DIRECTRICES PARA AUTORES/AS

## ENVÍO DEL MANUSCRITO

Interfases publica tres tipos de artículos: trabajos de investigación (hasta 5000 palabras), avances en investigación (hasta 2800 palabras) y revisiones (hasta 1500 palabras).

Todos los artículos se envían del mismo modo. Una vez que el editor verifique que el contenido del manuscrito pertenece al ámbito de Interfases, lo remitirá a un proceso de revisión por pares. Este proceso (compuesto de dos rondas) toma aproximadamente 2-3 meses, pero este tiempo podría extenderse, dependiendo de la complejidad del manuscrito.

Los manuscritos enviados a Interfases no deben haberse publicado previamente ni estar en consideración para su publicación en otra revista.

### Página del título

La página del título debe incluir:

- Un título conciso e informativo (hasta 30 palabras).
- El nombre completo de cada autor, incluyendo la afiliación institucional, la dirección de correo electrónico y el código ORCiD.
- Resumen de 200-250 palabras. El resumen debe indicar la naturaleza y contribución del estudio. Evite las abreviaturas no definidas, las ecuaciones matemáticas o las referencias bibliográficas en el texto del resumen.
- Palabras clave (3-5) separadas por comas. Las palabras clave deben tomarse de la taxonomía de la IEEE Computer Society: https://www.computer.org/digital-library/journals/sc/tsc-taxonomy-list

**Texto**

Los trabajos enviados deben haber sido redactados en un documento Word (.doc o .docx), y aquellos aceptados para ser publicados deben usar la plantilla de Interfases LATEX que estará disponible pronto.

Al redactar el manuscrito, es necesario usar la opción de numeración automática para enumerar las páginas. Por favor, evite el uso de funciones de campo. Para hacer tablas, utilice la función de tabla, no coloque una hoja de cálculo pegada. Si escribe su manuscrito con Word, use el editor de ecuaciones o MathType para las ecuaciones.

**Tablas**

Las tablas son el núcleo de los nuevos hallazgos reportados en la corriente principal de la ciencia, por lo tanto, incluya las tablas que considere estrictamente necesarias. Todas las tablas se enumeran utilizando números arábigos (por ejemplo, Tabla 1, Tabla 2, ...) e incluyen un título que detalla la relevancia de los datos presentados.

Las tablas se mencionan en el orden en que aparecen en el manuscrito. Además del número, el título y los datos, las tablas pueden incluir una nota para detallar la fuente de información, así como explicaciones adicionales que no están incluidas en el manuscrito.

**Abreviaturas**

Use abreviaturas solo si son necesarias para mejorar la legibilidad de su documento. Debe definir cada abreviatura en la primera mención, para después usarla de manera consistente.

**Conclusiones**

Recuerde que las conclusiones no son la versión narrativa y textual de las tablas incluidas en la sección Resultados. Por el contrario, las conclusiones reseñan y sintetizan los principales argumentos del artículo. Las conclusiones se extraen de los hallazgos y proporcionan una respuesta adecuada a la pregunta de investigación. Además, las conclusiones incluyen las limitaciones del estudio y sugieren nuevas preguntas y aplicaciones para futuros estudios.

**Referencias**

Las citas y las referencias deberán indicarse de acuerdo con las normas APA. Según la norma señalada, las referencias, enlistadas al final de la publicación, se realizarán de la siguiente forma:

*Libros*

Apellido del (los) autor(es), letra inicial del nombre del (los) autor(es). (Año de la publicación). Título del libro (en cursiva), (número de la edición). Nombre de la editorial.

*Artículos de revistas o capítulos de un libro*

Apellido del (los) autor(es), letra inicial del nombre del (los) autor(es). (Año de publicación). Título del artículo o el capítulo. Nombre de la revista o el libro (en cursiva), número de la revista (en cursiva), páginas en las que se encuentra el artículo o el capítulo.

*Libros electrónicos*

Apellido del (los) autor(es), letra inicial del nombre del (los) autor(es). (Año de publicación). Título del texto electrónico (en cursiva). Recuperado de http://... (dirección web).

*Artículos de revistas electrónicas*

Apellido del (los) autor(es), letra inicial del nombre del (los) autor(es). (Año de publicación). Título del artículo. Nombre de la revista (en cursiva), páginas en las que se encuentra el artículo. Recuperado de http://... (dirección web).

*Ponencias en congresos o simposios*

Apellido del (los) expositor(es), letra inicial del nombre del (los) autor(es). (Año, [indicar día] de [indicar mes]). Título de la ponencia [en cursiva]. Conferencia presentada en el [nombre del evento]. Recuperado de http://... [dirección web].

**Material suplementario electrónico**

Los autores pueden incluir archivos de texto (incluyendo tablas y figuras) y hojas de cálculo como material complementario. Sin embargo, para datos de investigación, es recomendable archivarlos en repositorios de datos. Para el código de software, los autores pueden usar plataformas como GitHub o similares.

Si los originales contienen fotografías o reproducciones de obras pictóricas, estas se entregarán aparte en archivos TIFF o JPG, con 300 píxeles de resolución (dpi). Si contienen gráficos, cuadros, dibujos, flujogramas u otros elementos, estos deben entregarse igualmente en un archivo aparte y en el programa original en que fueron creados (por ejemplo: Excel, Illustrator, etcétera).

**Lista preliminar para la preparación de envíos**

Los artículos deberán respetar el siguiente formato:

- Página A4.
- Título del artículo, centrado en negrita, con letra Times New Roman de doce puntos.
- Títulos del texto, centrados en negrita, con letra Times New Roman de doce puntos, dejando dos líneas en blanco antes del párrafo.

- Texto del cuerpo con letra Times New Roman de doce puntos, con espacio y medio de interlineado.

**Declaración de privacidad**

Los nombres y las direcciones de correo electrónico introducidos en esta revista se usarán exclusivamente para los fines establecidos en ella y no se proporcionarán a terceros o para su uso con otros fines.


UNIVERSIDAD
DE LIMA