Inteligencia Artificial: Un estudio de su impacto en Ciberseguridad

María Lourdes Martín Martín
Ingeniería Informática
Inteligencia Artificial

Dr. David Isern Alarcón
Dr. Friman Sánchez Castaño

23 de junio del 2024

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Inteligencia Artificial: Un estudio de su impacto en Ciberseguridad*
Nombre del autor:	*María Lourdes Martín Martín*
Nombre del consultor/a:	*Dr. David Isern Alarcón*
Nombre del PRA:	*Dr. Friman Sánchez Castaño*
Fecha de entrega (mm/aaaa):	06/2024
Titulación:	*Ingeniería Informática*
Área del Trabajo Final:	*Inteligencia Artificial*
Idioma del trabajo:	*Castellano*
Palabras clave	*Artificial Intelligence, Cybersecurity, Attacks*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

En este trabajo de investigación, se tiene como finalidad estudiar de forma detallada dos campos: la Inteligencia Artificial y la Ciberseguridad, además, de la innegable conexión entre ellos. Ambas áreas están intrínsecamente vinculadas y se busca proporcionar una visión exhaustiva sobre ellas.

Para empezar, se hace una exposición incluyendo los conceptos más importantes, a la vez, que una pequeña introducción histórica desde sus comienzos hasta la actualidad. El propósito de este estudio no es solo realizar un análisis teórico, también, se busca captar la atención del lector, así como, facilitar su comprensión elaborando una estructura y desarrollando el contenido de una forma ordenada y didáctica.

Además, se detallan varias de las numerosas aplicaciones que hay de la IA en Ciberseguridad. Así también, se tiene como objetivo aclarar la teoría con casos de estudio reales. De igual manera, se destacan tanto las limitaciones y los desafíos, como la normativa legal al respecto. Para continuar, se repasan las tendencias emergentes, finalizando con una serie de reflexiones.

Abstract (in English, 250 words or less):

The purpose of this research is to study in detail two fields: Artificial Intelligence and

Cybersecurity, in addition to the undeniable connection between them. Both fields are intrinsically linked, and the aim of this research is to provide a comprehensive overview of them.

Initially, a presentation is made including the most important concepts of these fields. Next, a short historical introduction from its beginnings to the present. The purpose of this study is not only to carry out a theoretical analysis, but also to capture the reader's attention and facilitate their understanding by developing a structure and the content in an orderly and didactic way.

Finally, several of the numerous applications of AI in Cybersecurity are presented. Furthermore, the objective is to clarify the theory with real case studies. Likewise, both the limitations and challenges are highlighted, along with the legal regulations in this regard. Bringing it all together, the emerging trends are reviewed, ending with a series of reflections.

Agradecimiento

Me gustaría expresar mi más sincero agradecimiento a todas las personas que han hecho posible la realización de este Trabajo de Fin de Grado.

En primer lugar, deseo agradecer a mi tutor, Dr. David Isern Alarcón, por su orientación y consejos valiosos, así como su paciencia y comprensión durante el desarrollo de este proyecto. Su conocimiento y dedicación han sido esenciales para el progreso de esta investigación.

También, me gustaría agradecer a todos los profesores del Departamento de Estudios de Informática, Multimedia y Telecomunicaciones de la UOC. Sus enseñanzas, dedicación y compromiso a lo largo de estos años han sido determinantes para mi formación académica.

Y finalmente, un especial agradecimiento a mi familia. A mi hijo Yeray y a mi nuera Alejandra por su amor y apoyo incondicional; a mi marido Bruce, por su paciencia, cariño y por estar siempre a mi lado, brindándome el apoyo emocional necesario para alcanzar este logro.

La presencia y respaldo de todos ellos han sido cruciales en este proceso.

Índice

Índice de Gráficos

Índice de Tablas

1. Introducción

1.1. Contexto y justificación del Trabajo

A medida que el mundo avanza en la digitalización, la Inteligencia Artificial (IA) crece a un ritmo sin precedentes, haciéndose presente en campos como son la sanidad, la banca, el comercio, la energía, la sostenibilidad y muchos otros, en los cuales el uso de IA es una realidad para la mejora de problemas, la toma de decisiones y el autoaprendizaje.

Ahora bien, así como la IA ha tenido grandes avances y su aplicación ha revolucionado diversos ámbitos y áreas, la ciberdelincuencia es un mal que sigue en crecimiento. En un mundo tan conectado por avances tecnológicos, el cibercrimen se ha convertido en una de las mayores amenazas para el mundo digital. Según un estudio publicado por EY Global (EY), en el 2022 los ciberataques tuvieron un coste total de 7 billones de euros. Este dato nos confirma cómo es de rentable el uso de tecnología para perpetrar delitos (Ernst & Young Global, 2023).Y, además, en ese mismo año, España registró un total de 374.737 ciberdelitos, teniendo un crecimiento del 26% respecto al año anterior. Este crecimiento convierte a este tipo de delitos en una problemática social, haciéndose necesario el uso de nuevas capacidades tecnológicas, como es el de la IA para hacer frente a este problema (Ministerio del Interior, 2022).

Como se puede ver, la ciberdelincuencia está evolucionando en su complejidad y sofisticación, por lo que es de vital importancia la protección de la información y de los Sistemas Informáticos.

La ciberseguridad es una diciplina muy amplia y en constante evolución que en combinación con IA han dado nuevas formas de protección en contra de la ciberdelincuencia.

La IA puede usarse para servir de apoyo a los profesionales en seguridad, para resolver los problemas de complejidad y entre otras muchas cosas, para crear sistemas con una infraestructura fuerte y menos vulnerable a los ataques (Ayerbe, 2020).

Por lo que, la IA se está convirtiendo en un arma muy poderosa a la hora de afrontar los desafíos en ciberseguridad. El aprendizaje automático en ciberseguridad: “Es la técnica que sigue siendo la más investigada entre todas las técnicas de inteligencia artificial en los proyectos actuales de innovación en ciberseguridad a nivel europeo’’ (INCIBE, 2023).

IA destaca por su capacidad para aprender de los ataques pasados, detectar los patrones de amenazas y responder a incidentes en tiempo real. Desempeña un papel muy importante en la protección contra los ciber-actores maliciosos (National Security Agency USA, 2021).

Sin embargo, su implementación aún enfrenta desafíos técnicos, éticos y legales. Además, también, hay una clara necesidad de mejorar la concienciación y la educación en ciberseguridad y en IA tanto en individuos, como en organizaciones. Ya que, de esta forma, entendiendo las bases sobre las que se construyen las amenazas y utilizando las tecnologías de la IA se podrá mejorar en la selección de las medidas de protección y de defensa.

Por otro lado, hay que tener en cuenta que el desafío de implementar IA en las diferentes etapas de ciberseguridad es un reto que ya se ha abordado y que está en continuo desarrollo e investigación. Por lo cual, es importante realizar un análisis de las nuevas tendencias, el desarrollo, la protección, la respuesta y las estrategias para responder a las amenazas cibernéticas. En el mundo digital, es imprescindible poder ser conocedor de dichos progresos, fomentar la formación continua para que sea mucho más fácil detectar y neutralizar los nuevos y sofisticados ataques, amenazas y los puntos de vulnerabilidad de los sistemas.

Por lo que, con este proyecto de investigación y estudio se busca hacer una aportación significativa con la finalidad de poder llegar a ser una fuente de información y análisis de una forma clara y sencilla a la vez que técnica de los actuales avances del aprendizaje automático en ciberseguridad. Para conseguir dicha finalidad, además, vamos a analizar varios escenarios de estudio, tales como, por ejemplo, la respuesta a *ransomware* automatizado, la detección de ataques de día cero, la simulación de adversarios, entre otros.

1.2. Objetivos del Trabajo

1.2.1. Objetivo Principal

Identificar las principales características, usos y aplicaciones de IA en el campo de la ciberseguridad.

1.2.2. Objetivos Específicos

- Analizar las amenazas y vulnerabilidades en el panorama actual de la ciberseguridad.
- Reconocer los fundamentos de la IA.

- Descubrir y estudiar cuáles son las aplicaciones de la IA empleadas en ciberseguridad.
- Analizar cómo los algoritmos de la IA que se utilizan para prevenir y detectar ataques cibernéticos también sirven para responder ante esas amenazas.
- Estudiar como la IA analiza las vulnerabilidades, las identifica y clasifica para proponer soluciones y verificar su reparación.
- Proponer recomendaciones para el uso de la IA en ciberseguridad, además de establecer limitaciones y desafíos.
- Desarrollar las principales conclusiones de la investigación, que indiquen los aspectos tanto positivos como negativos de las aplicaciones de la IA en ciberseguridad.

1.3. Estado del Arte

Una vez contextualizados el objetivo de la investigación, cabe plantear la pregunta de análisis a resolver:

¿Cuáles son las principales características, usos y aplicaciones de la IA en el campo de la ciberseguridad?

Para abordar esta cuestión de estudio se ha hecho una búsqueda y análisis de las teorías relacionadas con ciberseguridad e IA que contribuirán a una mejor comprensión del actual trabajo de investigación.

Para empezar, cabe destacar la Teoría del Empujón (*Nudge Theory*). Nació en el año 2008 y su autor es el economista *Richard Thaler*, ganador de un premio nobel. Esta teoría hace referencia al impulso que guía a las personas a tomar decisiones con beneficios a largo plazo. Estos impulsos van más allá de las rutinas diarias. En relación con la ciberseguridad, esta teoría se basa en destacar riesgos para guiar a los usuarios del ciberespacio hacia un comportamiento que busque mitigar las vulnerabilidades, promoviendo hábitos de protección ante los riesgos del cibercrimen (Muñoz, 2023). Impulsa al usuario a tomar acciones de protección como: el uso de antivirus en todos los dispositivos electrónicos, el seguimiento de las recomendaciones a la hora de crear y gestionar contraseñas de forma segura, el estar informados sobre las nuevas tendencias y formas de atacar la vulnerabilidad de los usuarios del ciberespacio.

Por otro lado, es relevante mencionar la Teoría General de Sistemas (TGS). Su origen concreto data de 1950 por el biólogo australiano *Ludwin Von Bertalanffy*, quién planteó las bases para el desarrollo de la teoría. Más adelante, se han ido haciendo contribuciones y TGS ha ido evolucionando y cambiando, adaptándose al modelo actual.

Esta teoría refleja el estudio multidisciplinario de los sistemas en general. Su principal propósito es el estudio de todos los principios aplicables a los sistemas, sea cual sea su nivel y en cualquier campo de investigación. Por lo que se puede aplicar en la presente investigación debido a que tiene relación, ya sea con el campo de la ciberseguridad, como también con el desarrollo de IA. Es digno de mención que el objetivo de esta teoría es el descubrimiento de las dinámicas, las restricciones y las condiciones de un sistema. Y esto es lo que se ha visto en los últimos años, tanto en el desarrollo y como en el avance de la ciberseguridad y la IA. La TGS promueve la integración de IA para desarrollar sistemas que se ajusten y puedan responder de forma efectiva a los continuos desafíos, con una capacidad más fortalecida a la hora de resistir o recuperarse de forma rápida de los incidentes cibernéticos, y con la habilidad de prever y responder a amenazas. (Gonzales, Dormido, & Sánchez, 2019).

De igual manera, conviene destacar como la Teoría de la Computabilidad se relaciona directamente con IA y también, con la denominada Teoría de la Recursión. Dicha teoría estudia los problemas que se pueden resolver por medio de algoritmos. Fue desarrollada en los años 30 gracias a la aportación de trabajos de autores como *Church*, *Gödel, Klenner, Post* y *Turing*. Su desarrollo ha tenido gran influencia a lo largo del tiempo en aspectos de práctica computacional, interpretación de programas, dualidad entre software y hardware, así como en otras aplicaciones del campo tecnológico. Esta teoría ha sido una de las impulsoras de la IA (Gallardo, Lesta, & Arques, 2003).

Finalmente, otra teoría que tiene una fuerte relación con la IA y que es de importancia para esta investigación es la Teoría de la Información desarrollada por *Claude Shannon* en 1948. Este trabajo estableció las bases para la Teoría Matemática de la Comunicación (*Shannon-Weaver Model*), en la que *Shannon* colaboró con *Warren Weaver*. Se utilizaron conceptos de la Teoría de la Información, como la entropía, para conseguir optimizar los canales de comunicación. De esta forma, se garantizaba una trasmisión de datos de forma segura. Por lo que, vista la conexión anterior, se puede afirmar que dichos estudios son las bases teóricas para crear herramientas de IA que puedan llegar a ser armas eficaces capaces de analizar, detectar y responder amenazas cibernéticas (Aqib, Samreen, Sania, & Munawar, 2023).

Ahora, una vez hecho un amplio desarrollo sobre las teorías más importantes relacionadas con ciberseguridad e la IA, se responderá a la pregunta base de este estado del arte:

¿Cuáles son las principales características, usos y aplicaciones de IA en el campo de la ciberseguridad?

Para empezar, en este estado del arte se examinan los estudios actuales más relevantes sobre los usos de IA en la ciberseguridad. Esta investigación seguirá como referencia diversas guías. Las cuales, servirán como guía central para orientar el desarrollo del trabajo. Seguidamente, se aportarán datos encontrados en diversas fuentes (artículos científicos, *journals*, etc), que enriquecerán el área de estudio. Además, también se incluirán casos de estudio en los que se demuestra cómo se ha aplicado la teoría en situaciones prácticas. Y con estas fuentes se busca organizar y clasificar de forma clara toda la literatura encontrada sobre los temas a tratar.

Lo primero, como guía en este estudio, se destaca la importancia y se sigue la estructura de las cinco funciones clave derivadas del marco de ciberseguridad del Instituto Nacional de Estándares y Tecnología (NIST) (2024):

1- Identificación (*Identify*)
2- Protección (*Protect*)
3- Detección (*Detect*)
4- Respuesta (*Respond*)
5- Recuperación (*Recover*)

Además, también como guía, se enlaza con el Manual de Buenas Prácticas del Centro Criptológico Nacional. Esto garantiza una estrategia de cercanía sobre el contexto donde ambas áreas de estudio, IA y ciberseguridad se juntan.

La combinación de estas guías y fuentes asegura que el estado del arte no solo sea exhaustivo y sólidamente respaldado, sino también relevante y actualizado. Ofrece una base sólida para futuras investigaciones y aplicaciones en el campo de la ciberseguridad impulsado y fortalecido por IA. Su estructura y desarrollo facilita a los lectores una visión comprensiva del potencial de la IA para optimizar la ciberseguridad en diferentes contextos.

1.4. Enfoque y método seguido

El enfoque y método seguido en este estudio es una Revisión Sistemática de Literatura (SLR) para proporcionar una investigación robusta, fiable y bien fundamentada. Con ello se consigue una exposición exhaustiva del núcleo de estudio, así como un conocimiento profundo del contexto y el estado actual de la investigación.

Primero, se identifican y recopilan los estudios relevantes de diversas fuentes (artículos científicos, *journals*, conferencias, etc) sobre la temática a tratar. Se evalúa e interpreta la información obtenida con la finalidad de identificar todos los aspectos relevantes de los campos a tratar. De esta forma, se hace una revisión exhaustiva y de los trabajos

publicados. Después, se hará un análisis crítico, una vez se hayan estructurado los datos de una forma lógica identificando patrones. Se clasifican los estudios según su enfoque (prevención, detección, etc) y los contextos en los que se aplican (dispositivos IoT, redes corporativas, etc). Esta metodología ayuda a tener una visión comprensiva, amplia y profunda sobre el campo de estudio. Ofrece una exposición detallada de los conocimientos, técnicas efectivas, avances y tendencias actuales en el área de estudio a investigar. Por lo que, con esto se pueden detectar áreas de conocimiento poco investigadas que serán posibilidades perfectas para desarrollar futuros estudios (Kitchenham & Charters, 2007).

Seguido de lo anteriormente mencionado, en la presente investigación se iniciará con una revisión de las teorías aplicadas en ciberseguridad y a la IA. Se continúa con la descripción detallada de los conceptos más relevantes en temas de IA y ciberseguridad, a fin de ofrecer al lector cuál es el estado actual de los campos a investigar, finalizando con conclusiones y futuras líneas de investigación.

Para finalizar, se ofrecerá al lector datos estadísticos actuales de la realidad de los dos campos mencionados y respectivos anexos para aclarar y esclarecer la investigación y su comprensión.

1.5. Planificación del Trabajo

En lo que se refiere a la planificación de este trabajo de investigación se procede a realizar su organización de la siguiente manera:

Nombre de la tarea	Fecha de inicio	Fecha de finalización	Feb	Mrz	Abr	May	Jun	Jul
Entrega del perfil de TFG	28.02.2024	04.03.2024						
Entrega PECO	28.02.2024	11.03.2024						
Desarrollo de título	28.02.2024	28.02.2024						
Desarrollo de palabras clave	28.02.2024	28.02.2024						
Desarrollo de la temática	29.02.2024	01.03.2024						
Problemática	02.03.2024	04.03.2024						
Objetivos	05.03.2024	09.03.2024						
Bibliografía	09.03.2024	09.03.2024						
Correcciones	10.03.2024	11.03.2024						
Entrega PEC 1	12.03.2024	26.03.2024						

Descripción del TFG	12.03.2024	14.03.2024						
Corrección de obj. generales y específicas	15.03.2024	18.03.2024						
Planificación con hitos y temp.	19.03.2024	22.03.2024						
Contenido de las PECs y entrega final	23.03.2024	24.03.2024						
Estructura de la memoria del TFG	25.03.2024	25.03.2024						
Entrega PEC2	27.03.2024	03.05.2024						
Identificación trabajo y fechas infor.	27.03.2024	29.03.2024						
Descripción del avance del proyecto	30.03.2024	10.04.2024						
Evaluar el grado de cumplimiento objetivos	30.03.2024	02.04.2024						
Justificación de cambios necesarios	03.04.2024	10.04.2024						
Realizar actividades	11.04.2024	21.04.2024						
Realización actividades previstas PT	11.04.2024	15.04.2024						
Realización actividades no previstas	16.04.2024	21.04.2024						
Realizar las desviaciones actualizar c.	22.04.2024	26.04.2024						
Resultados parciales obtenidos	27.04.2024	29.04.2024						
Comentarios	30.04.2024	02.05.2024						
Entrega final de PEC2	03.05.2024	03.05.2024						
Entrega PEC3	04.05.2024	29.05.2024						
Identificación del trabajo e informe	04.05.2024	05.05.2024						
Descripción avance del proyecto	04.05.2024	05.05.2024						
Evaluar el grado de cumplimiento obj	06.05.2024	06.05.2024						
Justificación de cambios necesarios	07.05.2024	07.05.2024						
Realizar actividades	08.05.2024	08.05.2024						
Realización actividades previstas PT	09.05.2024	10.05.2024						
Realización actividades no previstas	11.05.2024	11.05.2024						
Realizar las desviaciones actualizar	12.05.2024	15.05.2024						
Resultados parciales obtenidos	16.05.2024	20.05.2024						
Correcciones	21.05.2024	28.05.2024						
Entrega final de PEC3	29.05.2024	29.05.2024						
Entrega PEC4	30.05.2024	16.06.2024						
Memoria final	30.05.2024	13.06.2024						
Correcciones	14.06.2024	15.06.2024						

Entrega final de PEC4	16.06.2024	16.06.2024						
Entrega PEC5	17.06.2024	23.06.2024						
Elaboración presentación	17.06.2024	23.06.2024						
Síntesis del trabajo	17.06.2024	20.06.2024						
Entrega de síntesis y presentación	21.06.2024	23.06.2024						
Entrega PEC5B	24.06.2024	30.06.2024						
Presentación y defensa	24.06.2024	30.06.2024						
Actividad de evaluación	24.06.2024	30.06.2024						

Tabla 1. Planificación del trabajo

Elaboración Propia

1.6. Breve sumario de productos obtenidos

La presente investigación se estructura de la siguiente manera:

El capítulo 1, expone una breve introducción de los campos a investigar, así como un análisis a las principales teorías científicas que tienen relación con el tema. De igual manera contiene una descripción de la metodología a utilizar como también, la planificación de trabajo.

El capítulo 2, se cetra en la descripción y análisis de la cibercriminalidad y la ciberseguridad, definición, evolución y retos con la finalidad que el lector tenga un análisis detallado de este campo.

El capítulo 3, contiene un análisis de IA en la ciberseguridad. Explica y detalla los fundamentos, las definiciones, los algoritmos y más información adicional sobre el tema. Reúne datos sobre casos de estudios en los cuales se han desarrollado herramientas de IA para su aplicación en la ciberseguridad. Así también, se detallan las limitaciones, desafíos y vulnerabilidades de la IA, los ataques y las consideraciones éticas. Reúne un análisis del futuro de IA aplicada en el campo de ciberseguridad, su impacto y las investigaciones actuales.

Dentro del mismo capítulo se detallan buenas prácticas y recomendaciones en el cual se analiza la integración de una forma efectiva de la ciberseguridad y la IA. Finalmente, en el capítulo 4 se establecen las conclusiones de la investigación, así como sus futuras líneas de investigación.

El capítulo 5, contiene el glosario de la terminología usada en el proyecto, mientras en el capítulo 6 se encuentra la abreviatura utilizada.

2. Cibercriminalidad y Ciberseguridad

2.1. Cibercriminalidad: Naturaleza y su evolución con la tecnología.

La palabra Ciber-Criminología fue creada por el Profesor *K. Jaishanka* en el año 2007 (kaspersky, 2023). La cibercriminalidad es el comportamiento delictivo en el ámbito de la informática, es decir, conductas delictivas para las cuáles utilizan la tecnología.

La naturaleza de este tipo de criminalidad es muy complicada y supone un serio y continúo desafío. Partiendo de que Internet facilita un cierto anonimato a los delincuentes, lo convierte en un escenario perfecto para poder operar con impunidad desde cualquier lugar del mundo.

También, hay que considerar la diversidad de formas que hay para cometer los delitos y su impacto generalizado pudiendo afectar mucho más que a simples individuos, lo que la convierte en altamente escalable. Y después, a todo ello, hay que sumar la rapidez con la que la cibercriminalidad está evolucionando (Alenezi, Alabdulrazzaq, Alshaher, & Alkharang, 2020).

Peña (2023) en su estudio sobre ciberdelitos y cibercrimen cita que "*El auge de la ciberdelincuencia está estrechamente ligada al desarrollo tecnológico informático*". A la vez que somos testigos de los beneficiosos avances, a nivel global, en ciberseguridad y tecnología, estos también son una oportunidad para que los ciberdelincuentes innoven y perfeccionen su forma de atacar.

Las tecnologías digitales han innovado la forma en que vivimos, trabajamos y aprendemos. Pero a la vez, este veloz crecimiento y desarrollo de las tecnologías digitales ha venido junto con daños significativos ya que proporciona nuevas herramientas para el crimen (U.S. Department of States, 2024).

Los ataques cibernéticos son actividades maliciosas dirigidas a dispositivos informáticos, sistemas y redes utilizando Internet, con la única finalidad de comprometer o corromper datos reservados y sensibles. (Maad, Omega, Youssef, Indu, & Humam, 2023).

En el siguiente gráfico, se puede observar los métodos más extendidos de criminalidad informática en el año 2017 y el año 2022. Destaca el aumento de ataques por *phishing*:

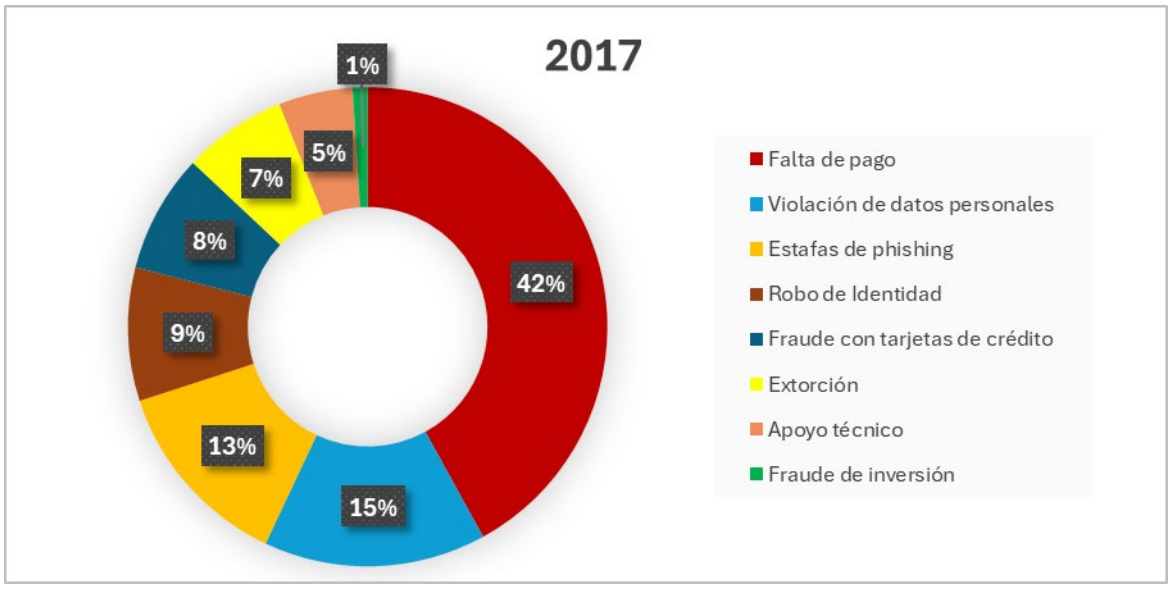

Gráfico 1. Métodos criminalidad informática 2017

Fuente: Statista (2017)

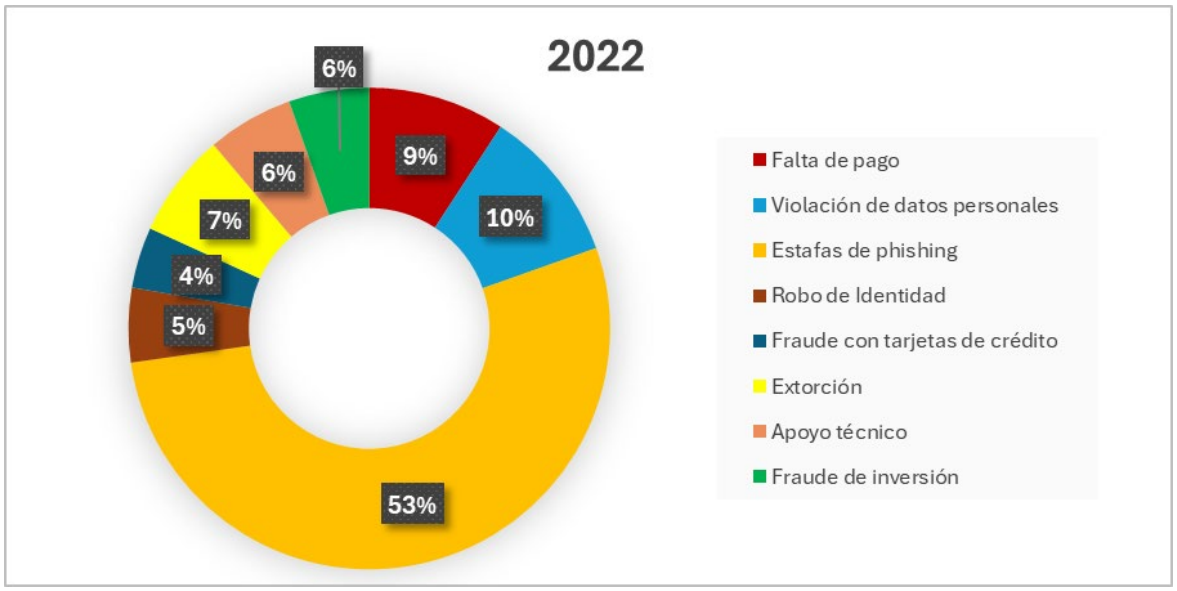

Gráfico 2. Métodos criminalidad informática 2022

Fuente: Statista (2022)

Mientras que, en la siguiente tabla, se han recopilado datos sobre el gran el incremento del cibercrimen en el periodo de un año (2021-2022). Con lo cual, con estos datos, solo se confirma la rápida velocidad con la que la cibercriminalidad está aumentando.

Aumento de la Ciberdelincuencia	300%	Desde el inicio Pandemia COVID
Filtraciones de datos sector sanitario	58%	
Correos que obstruyo Google	18.000	Malware y phishing vinculados al coronavirus

Tabla 2. Incremento del cibercrimen 2021-2022

Fuente: Stefanini Group (2022)

El origen de estos ataques puede tener una finalidad personal, política o económica. Al igual que pueden ser perpetrados tanto, por un solo individuo como por un colectivo determinado (Maad, Omega, Youssef, Indu, & Humam, 2023).

La siguiente figura muestra que los costos del cibercrimen aumentarán a más de 13 trillones de dólares para el año 2027.

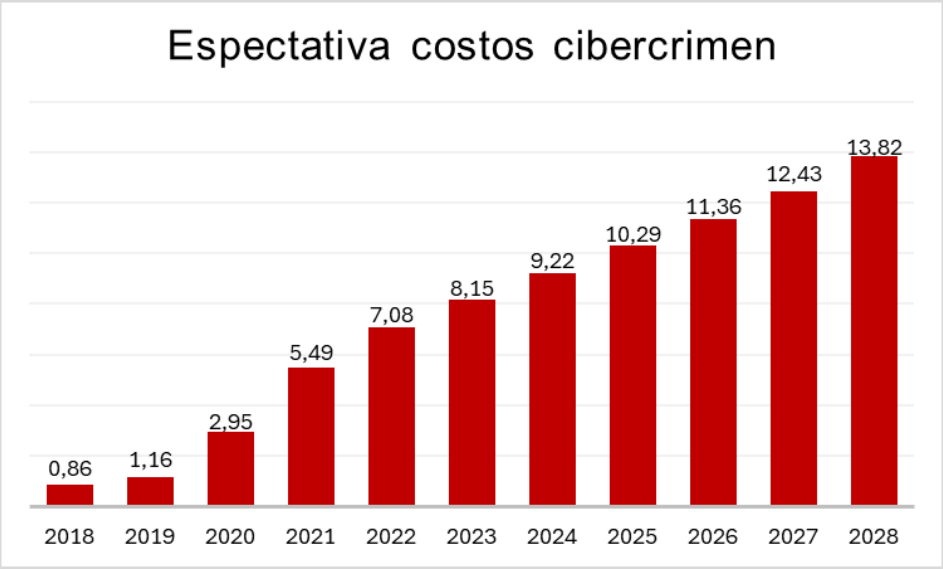

Gráfico 3. Expectativa costos cibercrimen

Fuente: Statista (2024)

2.2. Definición de Ciberseguridad y los tipos de incidentes.

Según el trabajo de *Sheldon* (2016), la ciberseguridad se fundamenta en 3 principios: Confidencialidad, Integridad y Disponibilidad (CIA):

- Confidencialidad: Solo las personas que tienen autorización son las que pueden acceder sin problemas a la información importante. Esto se implementa utilizando, por ejemplo, nombres de usuario, claves y listas de control de acceso.

- Integridad: Los datos que se envían serán los mismos que se reciben. Esto se logra con técnicas de protección y gestión de la información digital, tales como, el cifrado/descifrado y el *hash.*

- Disponibilidad: El sistema, la información y las funciones están disponibles para los usuarios autorizados según lo necesiten. Y esto se cumple revisando el *hardware*, actualizando el *software* y con una optimización de la red.

2.2.1. Definiciones

Muchos autores a lo largo de la historia han tratado de definir lo que es la ciberseguridad.

Según Fernández & Martínez (2018), la finalidad de la ciberseguridad es proteger cualquier tipo de información sensible y valiosa examinando y mitigando las posibles amenazas.

Conforme al Centro Criptológico Nacional (CNN), la ciberseguridad es el conjunto de medidas que se utilizan para proteger los sistemas, las redes y los programas de los ciberataques con finalidad delictiva (CCN, 2023).

Por otro lado, para los autores *Chandramouli*, *Raj & Bhagyaveni* (2020) la ciberseguridad es el conjunto métodos y tecnologías que protegen los dispositivos electrónicos (*hardware* y *software*), la información y las conexiones de redes de todo tipo de ataque o amenaza digital.

En última instancia, ciberseguridad es un grupo de técnicas, procedimientos meticulosamente diseñados con la finalidad de resguardar los datos confidenciales y desalentar la actividad cibernética maliciosa. Tiene la capacidad de crear un entorno capaz de proteger los dispositivos, las redes y la información digital de violaciones de acceso y prevenir el robo o alteración de la información (Maad, Omega, Youssef, Indu, & Humam, 2023).

2.2.2. Tipos de incidentes

Los diferentes tipos de incidentes en ciberseguridad están aumentando de forma exponencial. Según las estadísticas del *The independent IT-Security Institute* (AV-TEST), en su análisis de incidentes en ciberseguridad del 2023, la industria de la seguridad detectó un total de 4.618 ciberataques en Europa (AVTEST, 2024).
La mayor parte de estos ataques fueron la denegación de servicio distribuida (*DDoS*) y el *ransomware*. Exactamente, fueron unos 2.525 incidentes con *DDoS* y 1.066 ataques de ransomware. Además, se registraron 1.027 incidentes no especificados tales como, el robo de datos para *hacktivismo* o espionaje (AVTEST, 2024).

Además, analizando las investigaciones de AV atlas (plataforma desarrollada por el Instituto AV-TEST), se puede confirmar el enorme incremento en estos incidentes. Solo en *malware*, la cifra no para de crecer, como se puede apreciar en la tabla que se muestra a continuación (AV ATLAS, 2023).

Nuevos Malware en el año	45.473.701
Nuevos Malware en los últimos 14 días	3.611.406
Total Malware	1.386.823.807

Tabla 3. Incremento de Malware 2024

Fuente: AV-TEST (2024)

A continuación, gracias al Instituto AV-TEST podemos ver una visión general de la situación que hay respecto a los *malwares* y los *softwares* potencialmente maliciosos. Estas muestras están clasificadas según el nombre del *malware*, el desarrollo histórico o el sistema operativo (AV ATLAS, 2023).

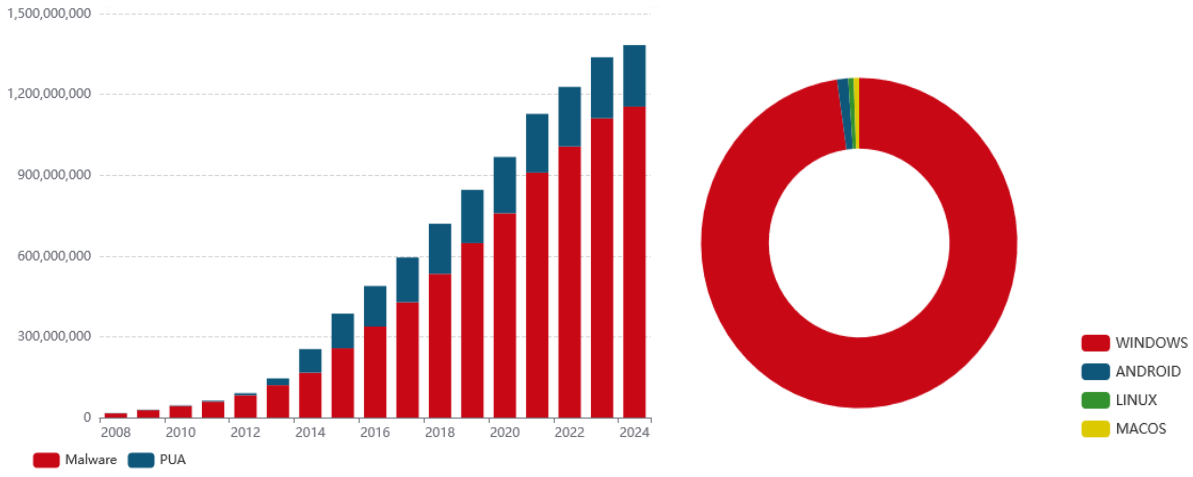

Gráfico 4. Total de malware y software malicioso según el desarrollo histórico, muestra de malware y software malicioso según el sistema operativo

Fuente: AV-TEST (2024)

Mientras que, en el siguiente gráfico, se puede ver el número total de ataques por segundo que se han sufrido en los últimos 12 meses por los nuevos *malwares* y *softwares* maliciosos. Según AV-TEST, se han contabilizado 121.106.623.

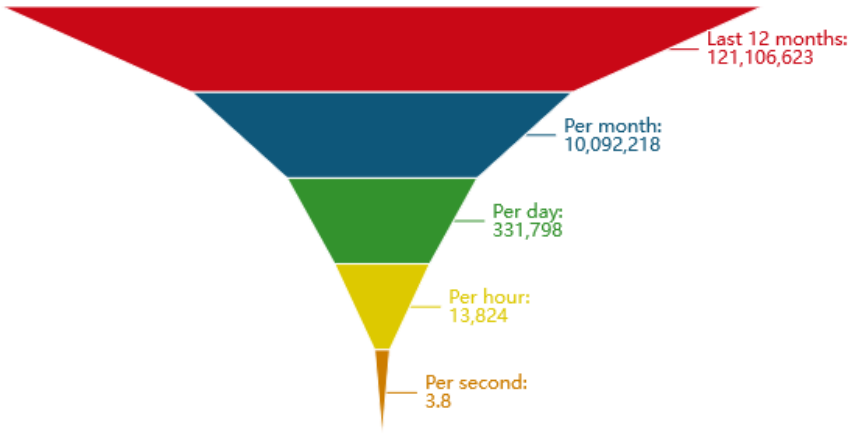

Gráfico 5. Muestra del total de malware y software malicioso por segundo

Fuente: AV-TEST (2024)

Y, por último, hay que destacar la imagen del mapa mundial de la Inteligencia de Seguridad de *Microsoft* (*Microsoft Security Intelligence*). En él, a fecha de 10 de junio del 2024, se puede observar la actividad de las amenazas globales con los países o regiones con más incidencias de *malware* en los últimos 30 días. Se muestra un total de 78.834.130 dispositivos con amenazas.

Después de analizar los datos recopilados, se concluye que para gestionar de forma efectiva la ciberseguridad es de vital importancia poder detectar y reducir la severidad tanto de los ataques y de las ofensivas que están activas como de las posibles vulnerabilidades que aún no han sido explotadas.

Por lo que, en este momento del análisis, se puede inferir que, en el ámbito de la seguridad informática, es fundamental comprender las diversas formas en que las redes y sistemas pueden llegar a ser vulnerados. Para proporcionar una visión clara y detallada de las amenazas más significativas, se realizará a continuación, una clasificación de los tipos de ciberataques más importantes según su naturaleza y efecto.

2.2.3. Los ciberataques y los riesgos de seguridad

Como comienzo, en el siguiente gráfico, se puede ver una clasificación de las diferentes amenazas más comunes que se pueden dar en el contexto de la ciberseguridad.

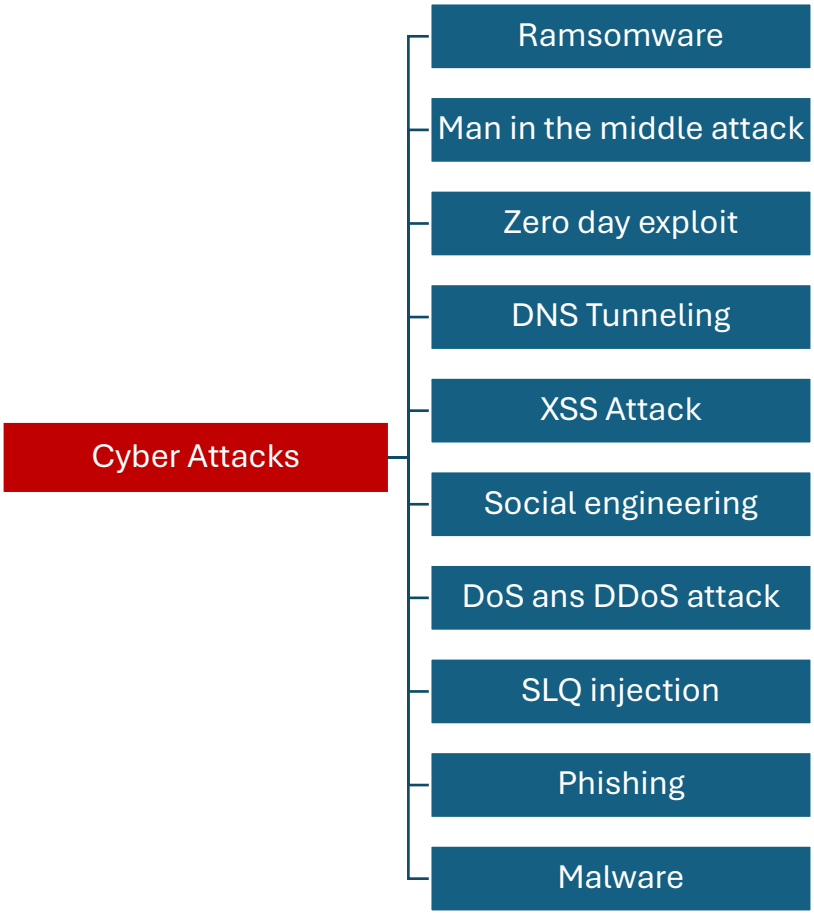

Gráfico 6. Cyber Attacks

Fuente: Sarker, Furhad & Nowrozy (2021)

Para Ahsan y otros (2022), el *malware* es un software malicioso. Su finalidad es infiltrarse, realizar acciones que no se hayan autorizado o causar daños a un sistema personal, cliente, servidor o red informática. Viola una red creando una situación vulnerable. En la mayoría de los casos, el usuario autorizado no reconoce su presencia.

Hay diversas maneras en las que un sistema puede infectarse, tales como, por ejemplo:

1- Cuando se engaña a la víctima para que instale, sin saberlo, una versión falsa de un archivo legítimo.
2- Cuando la víctima hace *clic* a un enlace no seguro y peligroso.
3- Cuando la víctima se conecta a un dispositivo infectado.

Hay diversas clasificaciones de *malwares* dependiendo del criterio que se aplique. Por ejemplo, se puede dividir al *malware* en dos categorías:

- Primera generación *malware* o *malware* estático: el *malware* no cambia su estructura después de infectar a su objetivo. Se clasifica dependiendo de la estrategia de infección.
- Segunda generación *malware* o *malware* dinámico: aquí, el *malware* sí cambia su estructura después de la infección y crea una variante nueva (Sharma & Sahay, 2019).

Seguidamente, este esquema muestra una clasificación de los *malwares* dependiendo del método de infección y de la intención del objetivo:

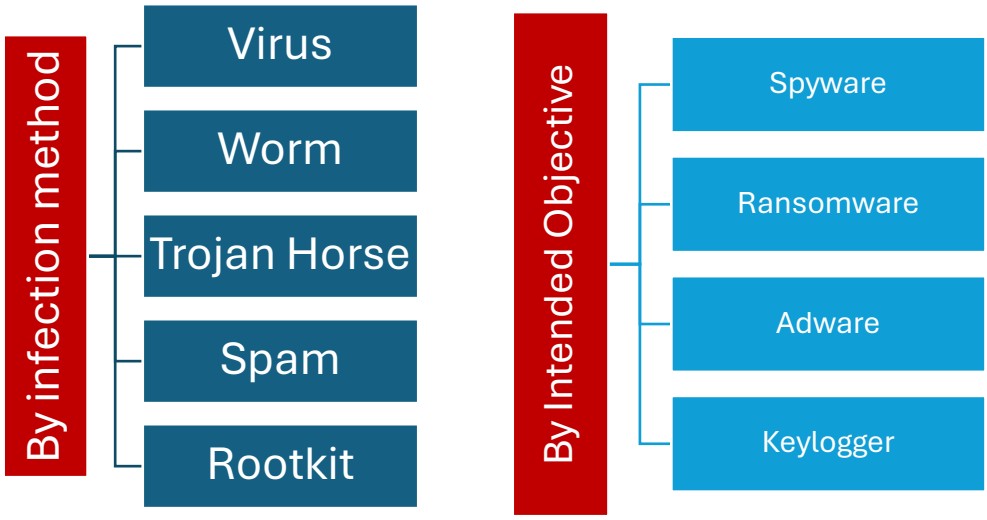

Gráfico 7. Clasificación de malware por método de infección y de intención del objeto

Fuente: Mohammed Alenezi (2020)

Para la presente investigación se destacarán los principales tipos de *malware* con la siguiente clasificación:

- *Bot* ejecutable: *Bot* es un robot. Y según se les dirija son capaces de hacer diversas funciones y no necesitan la intervención de hombre. Se ejecutan de forma automática. Las *Botnes* son conocidas como redes zombis. Diseñadas para realizar acciones maliciosas como los ataques *Distributed Denial of Service* (*DDoS*), envío de *spam*, entre otros. Se conectan a un servidor remoto que será quién le facilita las instrucciones a seguir para conseguir su objetivo. El *BonetMaster* controlará la red (Erquiaga, 2011).

Esquema de las fases del ciclo de vida de una *Botnet*:

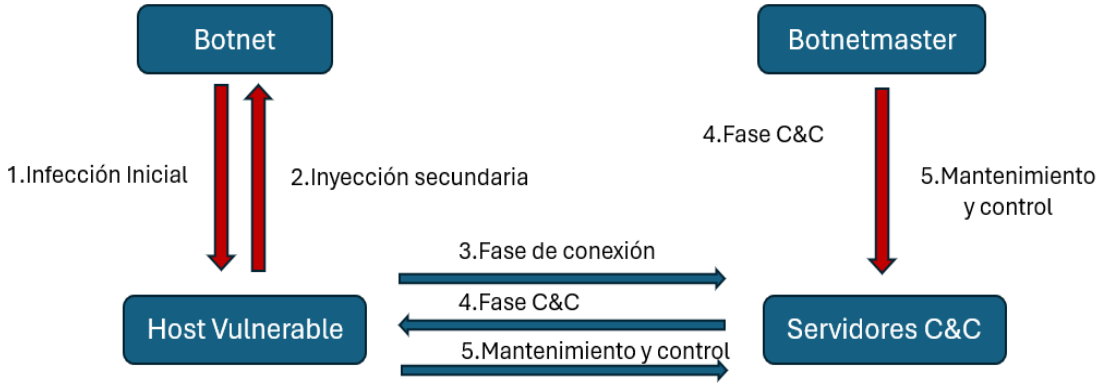

Gráfico 8. Fases del ciclo de un Botnet

Fuente: Erquiaga (2011)

- Troyanos: Se trata de programas que se disfrazan para conseguir una apariencia legítima con funciones maliciosas (Palmer, 2019).
- *Spyware*: Vigila y guarda información del usuario sin su permiso para luego enviarla sin su autorización a un tercero (Javaheri, Hosseinzadeh, & Rahmani, 2018).
- Virus: Son programas malignos que se replican y propagan de un sistema a otro sin control. Requieren de un archivo host infectado para poder expandirse. Los virus pueden infectar tanto las computadoras de escritorio como los servidores de red. Pueden, entre otras cosas, eliminar archivos, robar datos, dañar al sistema operativo (Guaña, y otros, 2022).
- *Ransomware*: Con este tipo de *malware* le cifran e incluso le roban los archivos a la víctima. Seguidamente, se le exige el pago de un rescate que suele ser en criptomonedas ya que es más difícil de rastrear. Se amenaza al cliente con la destrucción de los datos e inclusive, su publicación si estos son de naturaleza sensible (Cisco Systems, 2021).
- Gusanos: Este tipo de *malware* no necesita la acción directa de un usuario para distribuirse. Se replica a sí mismo para propagarse a otros ordenadores a través de las redes sin el uso de un archivo *host*. Aprovechan las vulnerabilidades de seguridad para infiltrarse. Roban información sensible, destruyen datos y manipulan el funcionamiento del sistema, llegando incluso a permitir que se controle de forma remota (Talukder & Talukder, 2020).

Un método común para distribuir el *malware* es el *drive-day attack*: El usuario visita una página *web* con apariencia inofensiva. Automáticamente, el sistema del usuario se infecta. El código malicioso se ejecuta en su navegador y buscará fallos de seguridad que no estén corregidos en su navegador, sus extensiones o sus complementos (*pluggins*). Si, finalmente, consigue con éxito su objetivo, el *malware* se descargará e instalará en el dispositivo sin que el usuario tenga conocimiento de ello. Y una vez allí, podrá realizar actividades maliciosas (Mostofa, 2021).

Una idea para protegerse de todas estas amenazas es blindar el perímetro del sistema instalando los controles que sean pertinentes, como, por ejemplo:

- Los sistemas de detección/prevención de intrusiones (*firewall, software* antivirus) (Mostofa, 2021).

Ahora bien, por lo que se refiere al término *phishing* es un método en el que se busca engañar al usuario solicitando una acción por su parte. Se realiza a través de correos electrónicos, sitios *web* falsificados o con mensajes de texto. Se intenta llevar a la víctima a un enlace o un archivo adjunto malicioso para que, por ejemplo, verifique una cuenta o realice un pago. De esta forma, podrán obtener información confidencial que utilizarán posteriormente en todo tipo de acciones fraudulentas, como acceder a las cuentas bancarias o robar la identidad del usuario, entre otras (Aakanksha, Ankit, & Dharma, 2017).

Otra técnica de ataque cibernético es *Man-in-the-Middle attack* (MITM**)**: Esta situación ocurre cuando los intrusos interceptan, con éxito y sin el consentimiento de ninguna de las dos partes que participan, una transacción o una comunicación entre ellas. La forma más común de entrada para estos intrusos suele ser:

1) El *wifi* público no seguro: Aquí, el atacante se posiciona entre el dispositivo de una víctima y la red. En este tipo de red que no utiliza cifrado o usa uno débil.
2) Cuando el *malware* de un atacante entra en el sistema de la víctima, puede instalar un *software* para conseguir la información sensible y privada de la víctima (Conti, Dragoni, & Lesyk, 2016).

En cuanto a los ataques de denegación de servicio (DDoS), suceden cuando los ciberdelincuentes intentan que el sistema de información o servicio no esté disponible a los usuarios legítimos, por lo que estos no son capaces de acceder. Se puede interrumpir el servicio de diferentes formas. El método más común consiste enviar una

enorme cantidad de solicitudes falsas a una página *web* para que los usuarios reales no puedan acceder a ella. Bloquean el acceso saturando el servicio. Este tipo de ataque se suele centrar en objetivos como, por ejemplo, los servidores *web* de empresas de alto perfil (plataformas comerciales, medios de comunicación, finanzas y gobiernos) (CISA, 2021)

Además, cabe destacar las diferentes formas de clasificación que tienen dichos ataques. Según el tipo del ataque, están:

1) Los ataques basados en *Host (Host-Based DoS)*: solo dirigido a un sistema o servidor.
2) Ataques Basados en Red *(Network-Based DoS)*: Se basan en la infraestructura de red.
3) Ataques Distribuidos (*DDoS)*: Incluyen a varios sistemas comprometidos (Dasgupta, Akhtar, & Sen, 2020).

Dependiendo del método de ataque, es decir, el tipo de tráfico que es lanzado a los sistemas de las víctimas, destacan tres:

4) Los ataques que se basan en volumen: se utiliza una gran cantidad de tráfico falso. Aquí, se engloban ICMP, UDP y ataques de inundación de paquetes falsificados. Se miden en bits por segundo (bps)
5) Los ataques *DDoS* de protocolo o de capa de red: aquí, se utilizan muchos paquetes de datos. Estos tipos de ataques comprenden las inundaciones SYN y *Smurf DDoS*, entre otros. Su tamaño se mide en paquetes por segundo (PPS).
6) Los ataques a la capa de aplicaciones: se usan solicitudes que han sido diseñadas con fines malintencionados para inundar las aplicaciones y hacer que no estén disponibles para clientes genuinos. El tamaño se mide en solicitudes por segundo (RPS) (Fruhlinger, 2024).

De igual manera, cabe mencionar las técnicas utilizadas que son comunes a todos los tipos de ataques *DDoS:*

a) Falsificación: los atacantes cambian la información del encabezado del paquete IP y de esta manera ocultan su identidad real.
b) Reflejo: con la dirección IP falsa de la víctima, se envían solicitudes a un tercer sistema. De esta forma, se consigue que este sistema envíe respuestas al usuario perjudicado.

c) Amplificación: se amplía con paquetes mucho más grandes el ataque hacia el objetivo engañando a ciertos servicios en línea (Fruhlinger, 2024).

Seguidamente, se puede ver un gráfico de un ataque distribuido de denegación de servicio (*DDoS*):

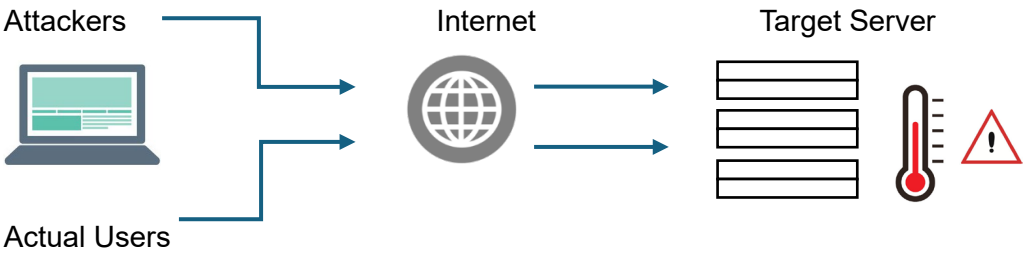

Gráfico 9.Ataque de distribuido de denegación de servicio

Fuente: Radware (2024)

A continuación, una representación de un ataque distribuido de denegación de servicio (*DDoS*).

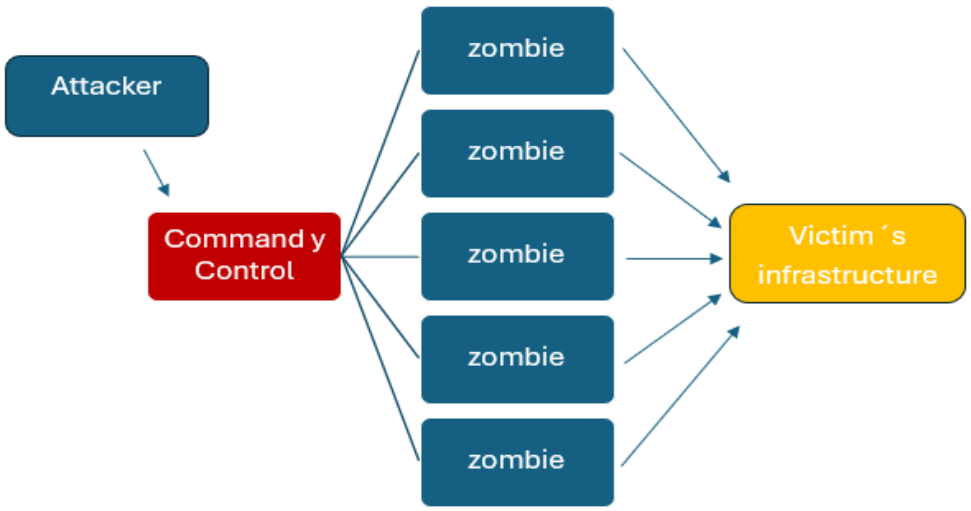

Gráfico 10. Ataque distribuido de denegación de servicio (DDoS)

Fuente: MPDI (2021)

Y, por último, un gráfico de Ataques *DDoS* a la capa de aplicación:

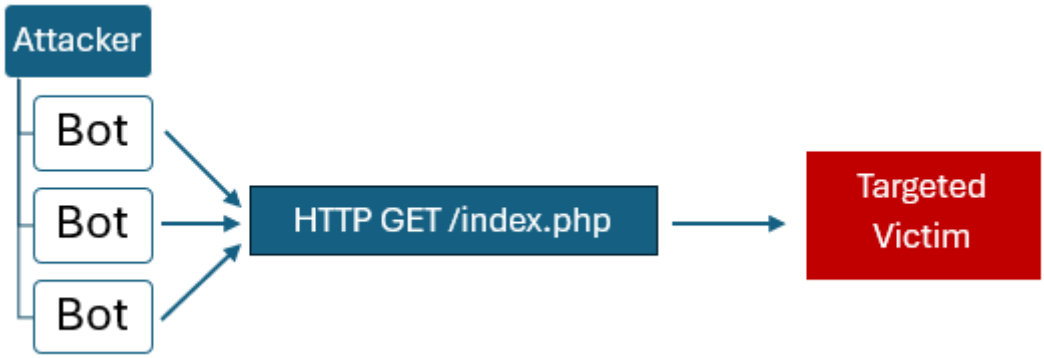

Gráfico 11. Ataque DDoS a la capa de aplicación

Fuente: MPDI (2021)

Por lo que se refiere a la inyección SQL (SQLI), esta es otra forma de robar, exponer, manipular y destruir información protegida introduciendo un código malicioso en SQL en bases de datos o aplicaciones *web*, como en un cuadro de búsqueda donde deberían ir datos normales, para intentar engañar al sitio *web*. Es un tipo de ataque que busca si hay alguna vulnerabilidad en la forma en que una aplicación *web* interactúa con su base de datos. Su objetivo es emplear código malicioso para manipular información de acceso al almacenamiento de la base de datos que se supone que no estaba destinada a ser mostrada (William, Viegas, & Orso, 2006)

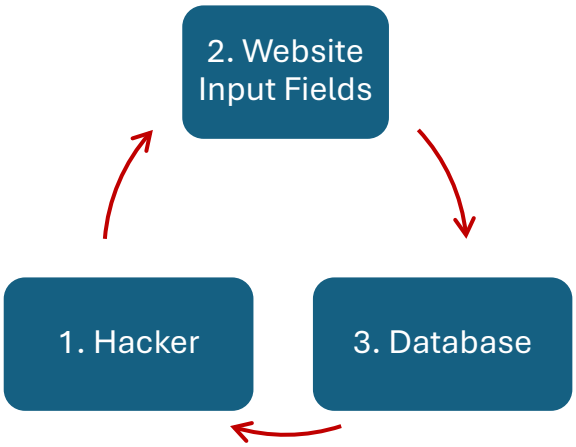

Gráfico 12. Inyección SQL

Fuente: MyIpaddress, 2021.

Por otro lado, se encuentra el ataque día cero (*zero-day exploit attack)*. Tal y como señala este término, aquí, hay cero días para solucionar el error. Esto quiere decir, que en este tipo de ataques se explota cualquier vulnerabilidad que se encuentre antes de que se conozca. Por consiguiente, es una amenaza de una vulnerabilidad de seguridad para la cual aún no se conoce la posible solución ya que el problema no se descubre hasta que ya se esté dando de forma activa. La mejor forma de prevenir este tipo de ataques es seguir una rigurosa y buena rutina en las prácticas de seguridad, como, por ejemplo, mantener el *software* actualizado, tener un excelente *software* de seguridad y limitar los derechos de administrador en los sistemas (Bilge & Dumitraş, 2012).

Otro tipo de ataque que se puede mencionar es el túnel DNS utilizado con fines maliciosos. De esta forma, el atacante puede no ser detectado y enviar información robada desde una red interna hacia el exterior. Se encapsulan datos dentro de mensajes DNS normales y se usa el puerto 53 (que es el utilizado normalmente por DNS y está abierto de forma habitual en los *firewalls*). De este modo, se trasmiten los datos hacia un servidor DNS controlado por el atacante (Ahsan, y otros, 2022).

También, es importante mencionar los ataques a dispositivos IoT (Internet de la Cosas). Son ataques que engloban a varios dispositivos conectados a Internet (por ejemplo: cámaras de seguridad). En la actualidad, la tecnología de IoT es cada vez más necesaria para garantizar que los dispositivos inteligentes, aplicaciones, etc, funcionen con mayor precisión. Es una tecnología en desarrollo que ofrece la ventaja de intercambiar información con otros dispositivos a través de la nube o de las redes inalámbricas (Inayat, Zia, Mahmood, Khalid, & Benbouzid, 2022).

En el año 2020, las estadísticas según los tipos de ataques fueron alarmantes. Los datos que mostraron fueron los siguientes: 41% *exploits*, 33% *malware* y 26% *user practice.*

A continuación, en la siguiente tabla, se muestran las estadísticas de estos ataques.

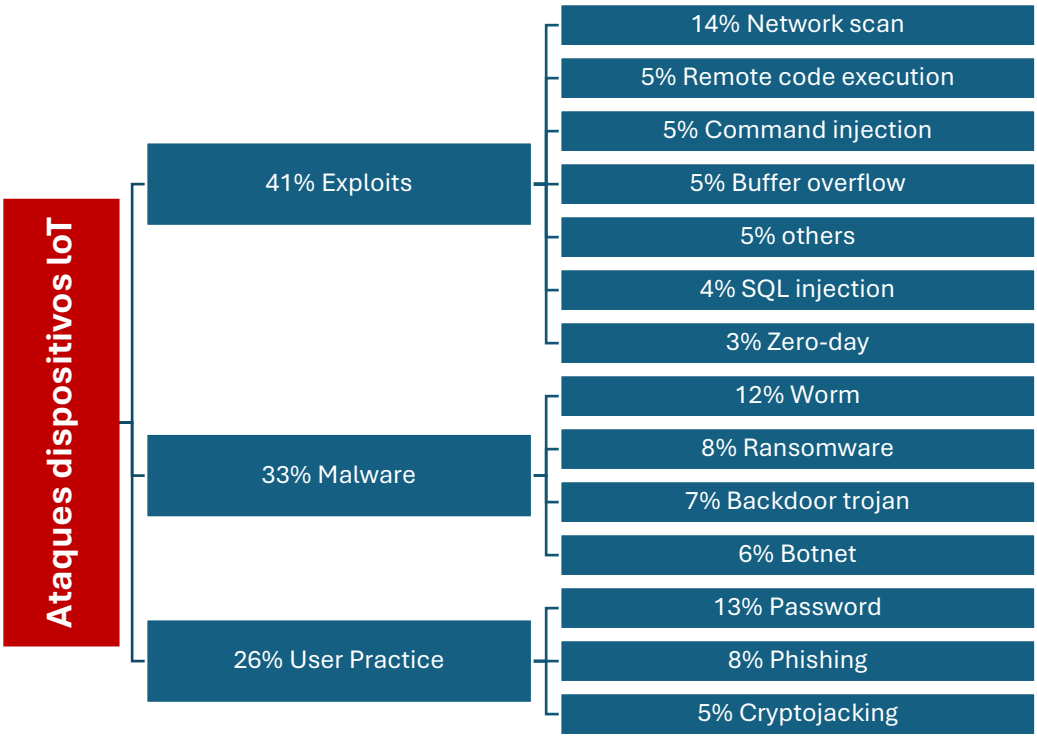

Gráfico 13. Ataques dispositivos IoT

Fuente: Lindsey O'Donnell, 2022.

2.3. Retos en la Ciberseguridad

Como se ha visto anteriormente, el panorama de las amenazas está creciendo y evolucionando continuamente. Según Cisco, en su Índice de Preparación para la ciberseguridad de Cisco 2024 (*2024 Cisco Cybersecurity Readiness Index*): Debido a la creciente complejidad del tráfico, ya que son una infinidad de usuarios, dispositivos y los dispositivos *IoT* los que se conectan a la vez en las redes empresariales, en las aplicaciones en la nube y en las bases de datos, la situación actual con las amenazas cibernéticas cada vez está más difícil. Todo se complica mucho a la hora de proporcionar una óptima seguridad. Según Cisco, el 54% de las organizaciones dice haber tenido en el último año un incidente de ciberseguridad, y el 73% las organizaciones piensan que es probable que tengan un incidente en los próximos de 12 a 24 meses (CISCO, 2024).

En la siguiente tabla, se puede ver un esquema del tipo de ataques que las compañías han experimentado:

Types of attacks experienced by companies	
Malware	76%
Phishing	54%
Credential Stuffing	37%
Supply chain and social engineering attacks	32%
Cryptojacking	27%

Tabla 4. Tipos de ataques experimentados por compañías

Fuente: CISCO (2024)

Y, por otro lado, en el Informe oficial sobre delitos cibernéticos 2023 de ciberseguridad Ventures, se prevé que el ciberdelito le costará al mundo 9,5 billones de dólares en 2024 (Esentire, 2024).

Si se divide ese gasto para calcular a cuánto dinero saldría por segundo, el resultado que se obtiene es de 302.000 dólares por segundo. Y si se midiera el gasto anual para 2024 como si fuera un país, el cibercrimen sería la tercera economía más grande del mundo después de Estados Unidos y China (Morgan, 2023).

Por lo que, hoy en día, están presentes muchos retos en la ciberseguridad. Citaremos algunos ejemplos:

Primero, se encuentra el *malware* sin archivos**,** cuando en un sistema se detecta que tiene un *software* malicioso, lo primero que se hace con la ayuda de un forense experto, es buscar cuál es el programa que no debería estar ahí. Pero el problema que hay con el *malware* sin archivos y que hace que sea muy peligroso es que no está en los archivos ya que no necesita usar ejecutables para llevar a cabo su finalidad maliciosa. Este tipo de *malware* se beneficia de las herramientas confiables. De esta forma, no es detectado por el sistema de detección basado en firmas, pudiendo evadir los antivirus. Si un profesional de seguridad hace un examen con herramientas forenses intentado buscar al atacante, este *malware* utilizará herramientas anti forenses para eliminar sus rastros (Sudhakar, 2020).

Como ya hemos dicho, este tipo de ataques no utilizan descargas de *softwares* maliciosos, ni tampoco intentan comprometer los sistemas guardando información en el disco. Estos ataques son mucho más sofisticados. Se busca la forma de sacar provecho de alguna vulnerabilidad de la aplicación para inyectar código dañino directamente en la memoria RAM. Además, el atacante ejecuta *scripts* e importa de forma inmediata código malicioso en la memoria volátil utilizando aplicaciones confiables como *Microsoft Office* o herramientas de administración nativas del sistema operativo *Windows* como *PowerShell* y *Windows Management Instrumentation* (WMI) (Bitdefender, 2017).

Otro de los desafíos actuales de la ciberseguridad está relacionado con la veloz evolución de las tecnologías digitales junto con el aumento de la conectividad de los dispositivos a través de IoT.

- Nuevo desarrollo tecnológico: tecnologías como la computación cuántica, las redes 5G y la computación en el borde traen nuevos desafíos de ciberseguridad. En la computación cuántica se destacan las debilidades del cifrado ya que dichas computadoras son capaces de romper algoritmos de cifrado.

  Crear y ejecutar algoritmos criptográficos *post* cuánticos capaces de resistir a los ataques cuánticos.

  Hay que destacar que con las redes 5G se aumenta significativamente la superficie de ataque. También, introduce nuevas vulnerabilidades con la división de la red y la virtualización.

  Y, por último, con la computación en el borde, la distribución de la seguridad y su gestión es complicada ya que el procedimiento y el almacenamiento se hace cerca de donde se generan. Pero los dispositivos están separados a grandes distancias formando una arquitectura distribuida. También, con este tipo de computación hay restricciones en la latencia y el ancho de banda, lo que limita los recursos que se pueden utilizar para obtener una óptima seguridad ya que ambas características son fundamentales para garantizar la integridad de los datos y prevenir vulnerabilidades.

- Proteger de forma correcta la seguridad y la privacidad de la información es un constante reto. A medida que la tecnología IoT evoluciona, los riesgos de seguridad, también. Cada vez son mayores los ciberataques de un dispositivo típico de IoT para obtener acceso y comprometer toda la red. Garantizar la seguridad de las redes que dependen de dispositivos IoT es una prioridad (Lim, 2023).

  Es un hecho que el surgimiento de este tipo de tecnología ha abierto muchas posibilidades, pero a la vez, ha expuesto nuevas debilidades y métodos para que los cibercriminales puedan comprometer la confidencialidad, integridad y disponibilidad de los sistemas conectados (Tariq, Ahmed, Bashir, & Shaukat, 2023).

Además, no se pueden obviar los continuos ataques de *ransomware*. Por ejemplo, según la publicación en la revista, La compañía de Seguridad *Sophos*, en su informe anual del estudio "El Estado del *Ransomware* 2024" nos muestra los siguientes datos:

El 59 % de las organizaciones se vieron comprometidas por *ransomware* el año pasado. Viendo el gráfico se puede observar una pequeña reducción, pero el valor medio de los pagos realizados para el rescate ha subido un 500% en el último año. Las víctimas informan de pagos de 2 millones de dólares de media, en comparación con los 400.000 dólares en 2023. Pero, esto solo es una parte del coste. Quitando los rescates, el documento destaca que el coste medio de recuperación fueron los 2,73 millones de dólares. Lo que refleja que ha habido una subida de cerca de un millón de dólares si lo comparamos con el coste medio de 1,82 millones de dólares que reportó la misma empresa en 2023 (Sophos, 2024).

Y las causas fundamentales de los ataques de *Ransomware* fueron las siguientes:

Causas fundamentales	2023	2024
Exploited vulnerability	36%	32%
Compromised credentials	29%	29%
Malicious email	18%	23%
Phishing	13%	11%
Brute forcé attack	3%	3%
Download	1%	1%

Tabla 5. Causas fundamentales de los ataques Ransomware

Fuente: Sophos, 2024.

Pero si se revisan los datos históricos de hace unos años, ya en 2017 Cisco confirmaba en su Informe Anual de Ciberseguridad de Cisco 2017, un crecimiento anual del 350% en los ataques de *ransomware* (Periman, 2017).

De la misma forma, actualmente, en su Informe anual (2023), Cisco Talos destaca que aún hay una persistente amenaza por los ataques de *ransomware* y *pre-ransomware*. Ya que estos incidentes siguen representando el mismo porcentaje (20%), respecto del pasado año, de la totalidad de los ataques que son gestionados por *Talos Incident Response* (CISCO, 2023).

Algo similar ocurre con los ataques *DDoS*, ya que no disminuyen. Según el Informe de análisis de amenazas globales 2024 de *Radware* (Radware's 2024 Global Threat Analysis Report): Este tipo de ataques están progresando y los piratas informáticos configuran sus estrategias para neutralizar las crecientes técnicas de mitigación. Comparando el año 2022 con el 2023, en este último el número de ataques *DDoS* por

cliente subió un 94 %. Y respecto al volumen de ataques por cliente, es decir, la cantidad total de tráfico malicioso generado hacia ese cliente durante los ataques aumento el 48% en el año 2023 comparándolo con el 2022 (Radware´s, 2024).

Según Cisco en el artículo ya mencionado se espera que los gastos totales en ciberseguridad alcancen 1 trillón de dólares para el año 2024 (CISCO, 2024).

Y, para finalizar este apartado, es importante destacar el peligro que representan los troyanos. Según *Talos*, el grupo de inteligencia de amenazas de Cisco, en su informe del año 2022 destacó el reto significativo que es en el ámbito de la ciberseguridad los siguientes troyanos: *Qakbot, Emotet, IcedID*, y *Trickbot*. Cada uno tiene sus específicas características que los hacen peligrosos y muy complicados. Inicialmente se crearon para ser troyanos bancarios. Pero, con el tiempo han evolucionado rápido y han aumentado en sofisticación. Por lo que, en la actualidad, han ampliado sus funciones y se han convertido en un arma muy versátil para los ciberdelicuentes ya que son capaces de realizar un amplio abanico de actividades ilícitas (CISCO TALOS, 2023).

Un año después, *Talos* en su informe del año 2023 destaca la capacidad de troyanos como *Qakbot* y *Trickbot*, de mantenerse activos o de volverse activar. Asimismo, destaca que, a pesar de que se desmantelaran en 2023 redes de dispositivos infectados por *Oakbot*, que estaban controlados por un atacante o un grupo de atacantes*(botnets*), este troyano podría resurgir perfectamente con un nombre diferente porque no se logró detener a los delincuentes que estaban detrás de él. Además, el informe puntualiza afirmando que esto ha ocurrido más veces, por ejemplo, con *Emolet.* También, destaca como los desarrolladores de *Trickbot* continúan activos inclusive después de que su red criminal fuera desmantelada en 2022. Y destaca el potencial de *IcedID* (CISCO TALOS, 2023).

Finalmente, toda la información recopilada y analizada en esta sección indica que, a pesar de todos los esfuerzos para desmantelar estas redes, las amenazas como estas son persistentes, siguen siendo reales y evolucionando a mucha velocidad.

3. IA en Ciberseguridad

3.1. Fundamentos de la Inteligencia Artificial

Así como aprendemos a leer, escribir y pensar, en busca de un resultado, la IA hereda esos rasgos como base a su desarrollo silogístico. La IA es una herramienta que colabora y potencia las capacidades humanas. Y esta perspectiva está aceptada y respaldada en numerosos estudios.

Profesores de *Stanford*, como *Stuart Russell* y *Peter Norvig*, destacan en su libro "*Artificial Intelligence: A Modern Approach*" la importancia de comprender los principios en los que se basa la creación y la elaboración de sistemas que pueden realizar trabajos que habitualmente necesitarían inteligencia humana (Russell & Norvig, 2016).

3.1.1. Definición

Existen numerosas definiciones sobre qué es la IA.
El Grupo de Expertos de Alto Nivel en Inteligencia Artificial (AI HLEG) de la Comisión Europea (CE), en 2019 la definió como sistemas que presentan un comportamiento inteligente evaluando su entorno y operando de manera independiente en ciertas situaciones para obtener unos resultados concretos (High-Level Expert Group on Artificial Intelligence, 2019).

La Agencia de la Unión Europea para la Ciberseguridad (ENISA), en su documento Investigación en Inteligencia Artificial y Ciberseguridad (junio, 2023) señala que no hay una definición estándar de la IA (Samoili, y otros, 2020).

Mientras que, según el Centro Criptológico Nacional (CCN) en su informe de buenas prácticas sobre la aproximación a la Inteligencia Artificial y la Ciberseguridad de octubre del 2023, la IA se puede definir como una subdisciplina de la informática. También, destaca su finalidad de crear sistemas con la facultad de desempeñar actividades que, hasta la actualidad, exigen inteligencia humana.

Estas tareas están enfocadas en cuatro puntos muy importantes que son: el aprendizaje, el razonamiento, la autocorrección y la creatividad.

- El aprendizaje: Se adquieren los datos y se generan reglas para poder transformarlos en un tipo de información que sea procesable. Estas reglas son los algoritmos.

- El razonamiento: Dependiendo del resultado que se esté buscando se selecciona un algoritmo u otro.
- La autocorrección: Para conseguir resultados más precisos los algoritmos se van perfeccionando.
- La creatividad: Se utilizan muchas técnicas de IA para generar nuevas imágenes, textos, ideas y música (CCN, 2023).

3.1.2. Historia de la IA.

El origen de la IA como campo formal de estudio surgió durante una escuela de verano de seis semanas en el MIT en la Conferencia de *Dartmouth* (1956), en el *Dartmouth College de New Hampshire*, Estados Unidos. Allí, un grupo de científicos se reunió para el Proyecto de Investigación de Verano sobre Inteligencia Artificial de *Dartmouth*. Como consecuencia, se acuñó el término de "Inteligencia Artificial" (MIT, 2020).

En el siguiente gráfico, se puede observar la cronología del surgimiento de la IA como disciplina:

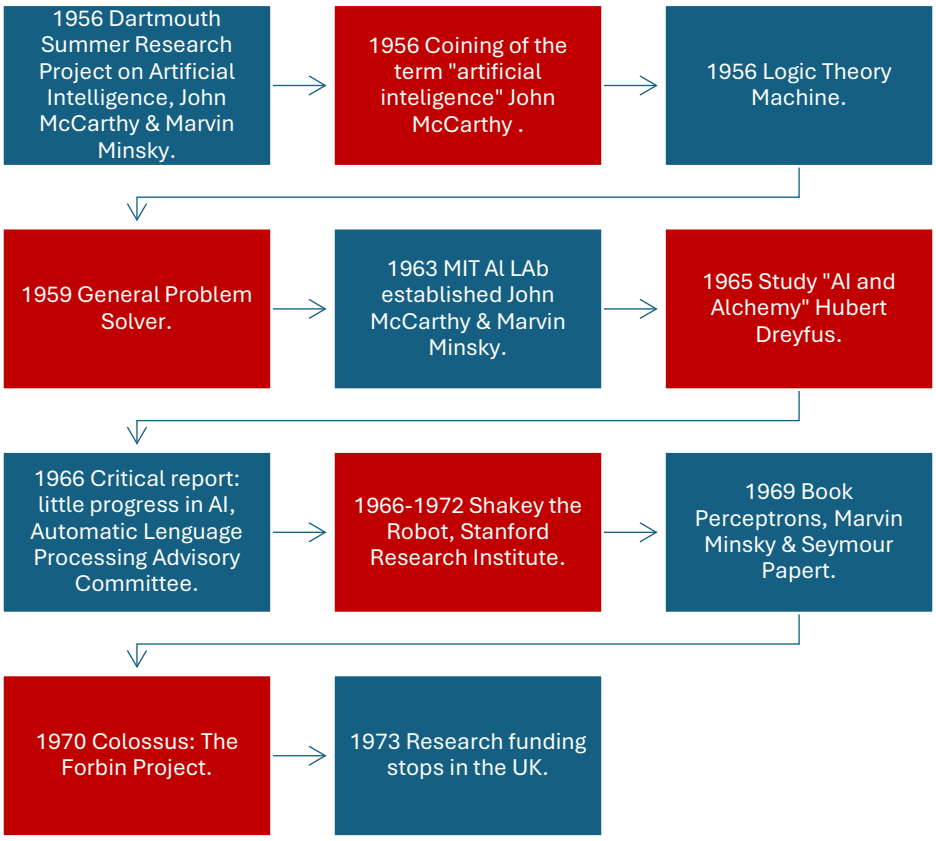

Gráfico 14. Cronología del surgimiento de la IA

Fuente: Springer (2023)

A partir de la segunda mitad del siglo XIX, la idea de la IA empezó a ser vista como una posibilidad real de estudio científico. Y el desarrollo de este concepto coincidió con la construcción de las primeras computadoras (Sheikh, Prins, & Schrijvers, 2023).

En el gráfico 15, se puede observar la cronología de las teorías de la IA:

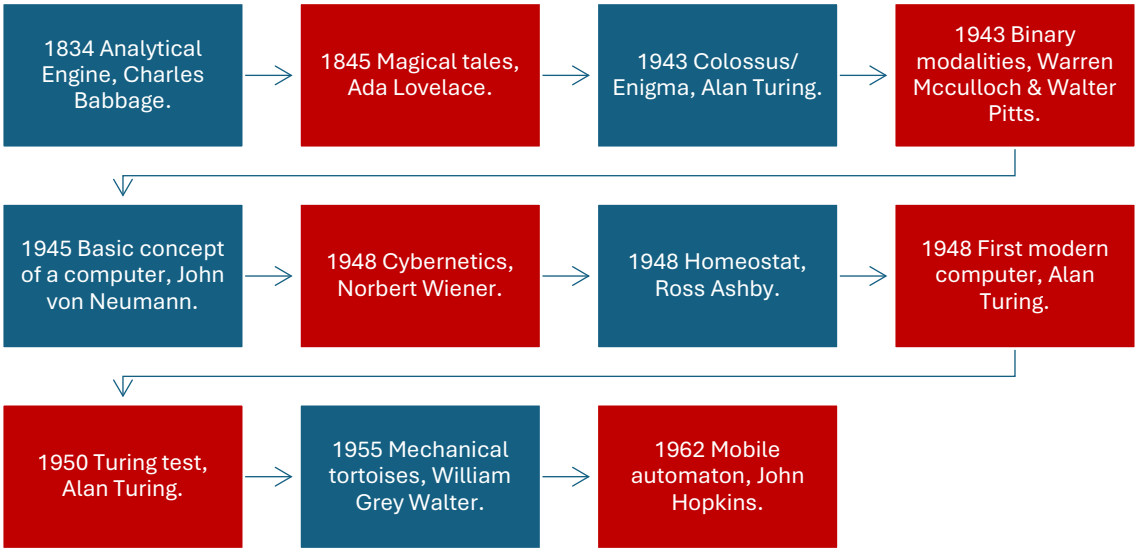

Gráfico 15.Cronología de teorías de la IA

Fuente: Springer (2023)

3.1.3. Conexión de la IA en la Ciberseguridad

La plataforma de investigación y análisis de *Deloitte Insights* (2021) destaca como en el área de la ciberseguridad, la IA es, sin ninguna duda, un multiplicador de fuerzas. Por un lado, capacita a los equipos de seguridad a responder más rápido que los ciberdelincuentes y por otro, puede predecir esos tipos ataques y preparar anticipadamente la forma de actuar.

Las técnicas y tecnologías que se utilizan van desde el aprendizaje automático, pasando por las redes neuronales y la IA generativa.

En lo que se refiere a los modelos de IA, el Centro Criptológico Nacional (2023), para su mejor comprensión, no solo los examina de manera aislada. Si no que los representa teniendo en cuenta el contexto en que se desarrollan y aplican. En primer lugar, está la IA que comprende la base teórica y el crecimiento de sistemas de información que pueden realizar actividades que habitualmente necesitan de inteligencia humana; en segundo lugar, está el aprendizaje automático que facilita a los sistemas la habilidad de

aprender sin una programación previa; y finalmente, está el aprendizaje profundo con algoritmos desarrollados que imitan la manera de proceder del cerebro humano.

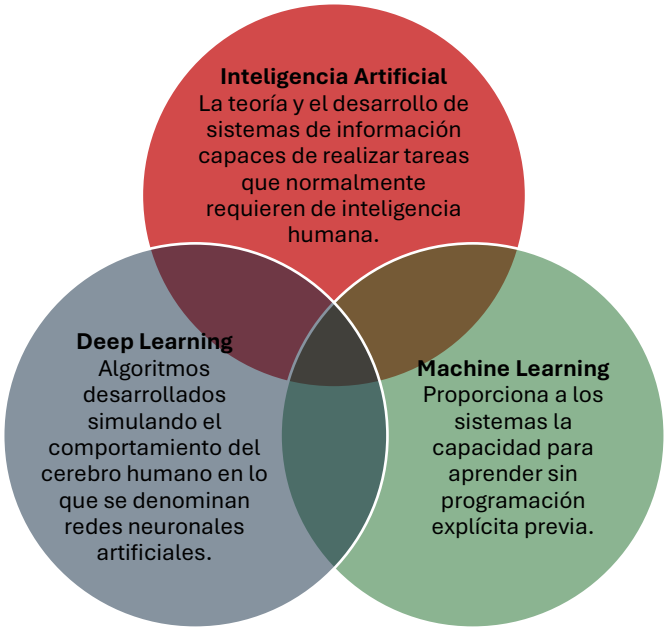

Gráfico 16. Inteligencia artificial, machine learning y deep learning

Fuente: CNN (2023).

A continuación, primero, se realizará un estudio sobre los diferentes enfoques dentro del campo de la IA, para después, poder pasar a un análisis detallado de cómo se han aplicado a ciberseguridad con ejemplos concretos y escenarios de estudio. Se va a exponer una estructura didáctica con una base clara de los conceptos importantes antes de estudiar las aplicaciones particulares.

3.1.4. Técnicas de IA: Modelado de Inteligencia de Seguridad basado en IA

La gestión inteligente de la ciberseguridad está basada en IA. A continuación, en el presente apartado, se explica y detalla las aplicaciones de cada uno de los modelos de IA.

3.1.4.1. Aprendizaje automático *(Machine Learning, ML)*

La IA y el aprendizaje automático (ML) están cambiando la ciberseguridad al perfeccionar y fortalecer la detección, prevención y respuesta a las amenazas cibernéticas. Hay muchos estudios recientes que respaldan esta afirmación y que seguidamente, se van a analizar.

Entre las aplicaciones de IA que están utilizando ML, destacan, entre otras:

1- Detectar amenazas y análisis de comportamiento: Los algoritmos de ML pueden analizar una gran cantidad de información. Por lo que, son perfectos para que sean entrenados con una colección de datos normales de red, del comportamiento del sistema, de la actividad de la base de datos, de la actividad del usuario o de la actividad de la aplicación, etc. Una vez que el entrenamiento ha terminado, serán capaces de detectar los comportamientos anómalos que podrían ser la señal de que hay una amenaza. Destacan como aplicaciones generales en Ciberseguridad:
    - El análisis de código: para identificar *malware*;
    - *Phishing* y detección de fraude: estudia páginas *web* y direcciones de correo electrónico (Sarker, Kayes, Badsha, Hamed Alqahtani, & Alex, 2020).

A continuación, en el siguiente esquema se puede apreciar una clasificación de ML basada en la naturaleza de los datos (etiquetados y no etiquetados) y en la finalidad de los algoritmos (clasificación, regresión, agrupamiento, asociación y toma de decisiones):

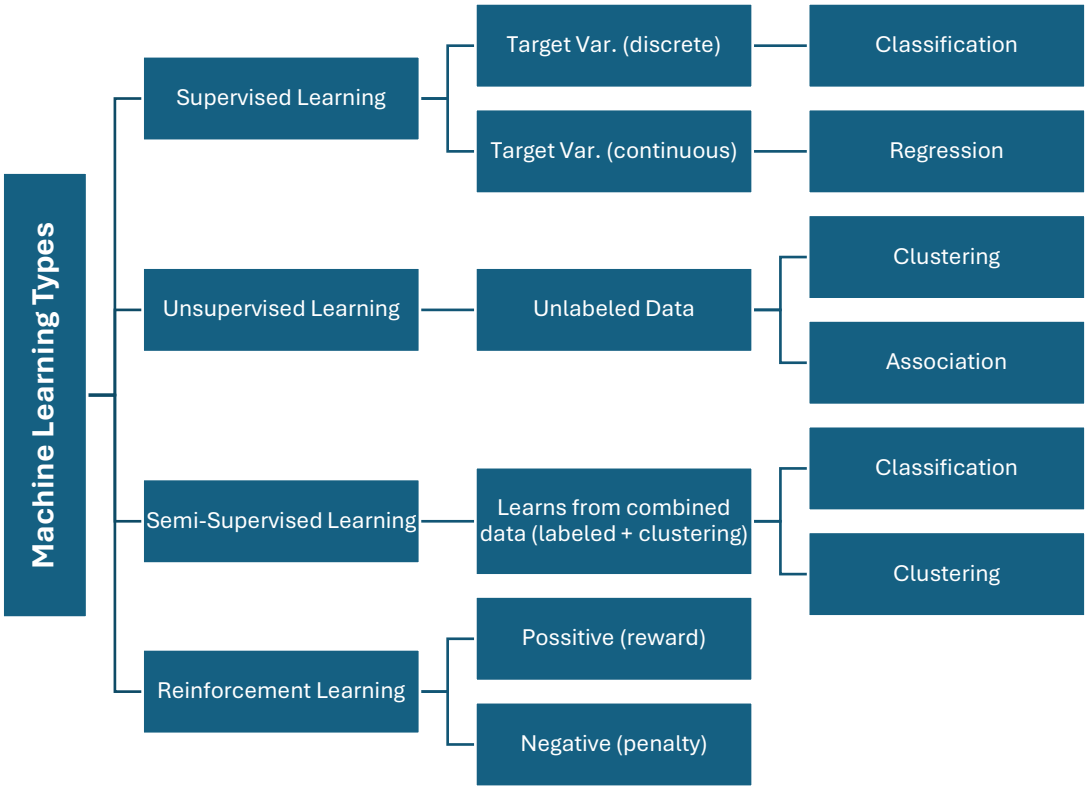

Gráfico 17. Clasificación del aprendizaje automático

Fuente: Springer (2023)

En cuanto al siguiente esquema, un tanto más completo que el anterior, se basa en los diferentes enfoques y técnicas utilizadas por ML. Cada categoría agrupa algoritmos y métodos dependiendo de cómo aprenden de los datos y la forma en la que llevan a cabo ese aprendizaje:

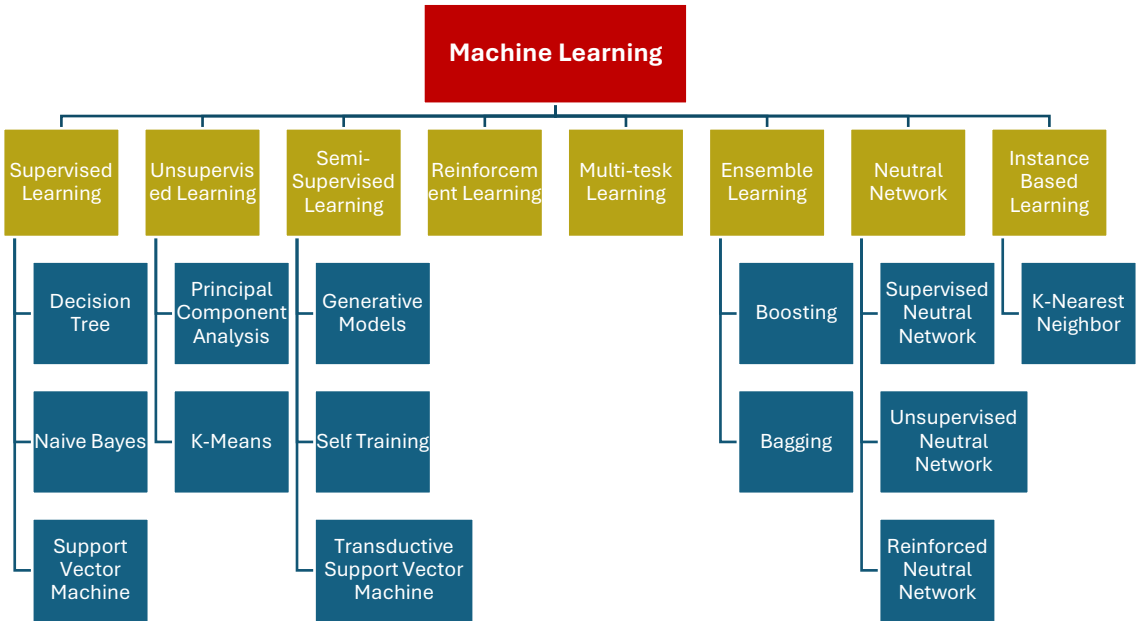

Gráfico 18. Machine Learning

Fuente: ResearchGate (2019)

ML se fundamenta en diferentes algoritmos para resolver problemas de datos. Según los científicos de datos no existe un único algoritmo universal que sea la solución óptima para resolver todos los tipos de problemas. Dependiendo del tipo de problema que se busque resolver, según sea la cantidad de variables y el modelo más idóneo, entre otras cosas, se utilizará un tipo de algoritmo u otro (Mahesh, 2019).

Y, por último, la siguiente clasificación en algoritmos ML:

Primero, según su estilo de aprendizaje.

a) Aprendizaje supervisado (*Supervised Learning*): etiqueta y resultado conocido.

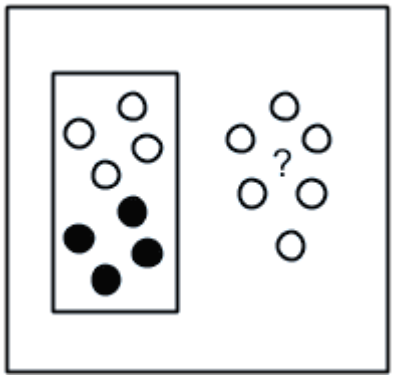

Gráfico 19. Aprendizaje supervisado

Fuente: Brownlee (2020)

b) Aprendizaje no supervisado (*Unsupervised Learning*):no tiene etiqueta y no hay resultado conocido

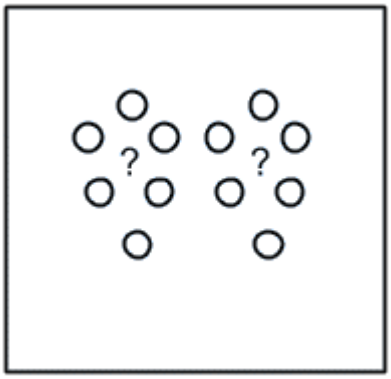

Gráfico 20. Aprendizaje no supervisado

Fuente: Brownlee (2020)

c) Semi-Supervisado (*Semi-Supervised Learning*): mixto

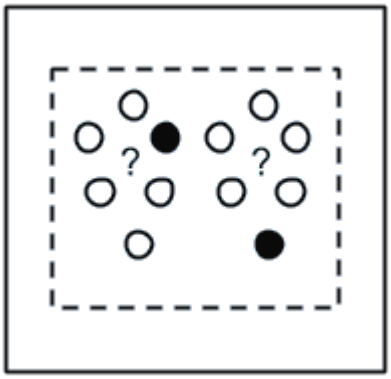

Gráfico 21. Aprendizaje semi-supervisado

Fuente: Brownlee (2020)

Y, después, como una agrupación de algoritmos similares ya sea en forma o por su función (Brownlee, 2023).

1- Algoritmos de regresión *(Regression Algorithms*): modelan relaciones entre las variables.

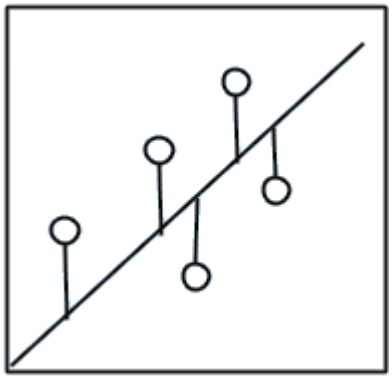

Gráfico 22, Algoritmos de regresión

Fuente: Brownlee (2020)

Ejemplos:

a) Regresión por Mínimos Cuadrados Ordinarios (*Ordinary Least Squares Regression* (OLSR))
b) Regresión Lineal (*Linear Regression*)
c) Regresión Logística (*Logistic Regression*)
d) Regresión por Pasos (*Stepwise Regression*)
e) Splines de Regresión Adaptativa Multivariante (*Multivariate Adaptive Regression Splines* (MARS))
f) Suavizado de Dispersión Estimado Localmente (*Locally Estimated Scatterplot Smoothing* (LOESS)) (Brownlee, 2023).

2-Algorirmos basados en instancias (*Instance-based Algorithms*): basados en ejemplos de datos de entrenamiento.

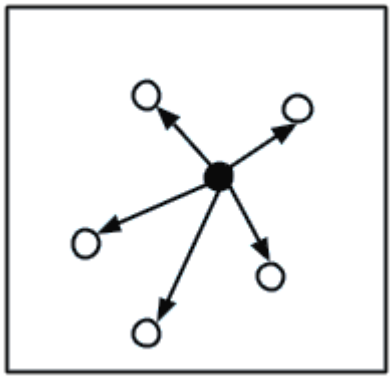

Gráfico 23. Algoritmos basados en instancia

Fuente: Brownlee (2020)

Ejemplos:

- k-Vecinos Más Cercanos (*Nearest Neighbor* (kNN))
- Cuantificación de Vectores de Aprendizaje (*Learning Vector Quantization* (LVQ))
- Mapa Autoorganizado (*Self-Organizing Map* (SOM))
- Aprendizaje Ponderado Localmente (*Locally Weighted Learning* (LWL))
- Máquinas de Vectores de Soporte (*Support Vector Machines* (SVM)) (Brownlee, 2023).

3-Algoritmos de regularización (*Regularization Algorithms*): se basan en la penalización a la función de coste del modelo de regresión según su complejidad. Favorece los modelos más simples.

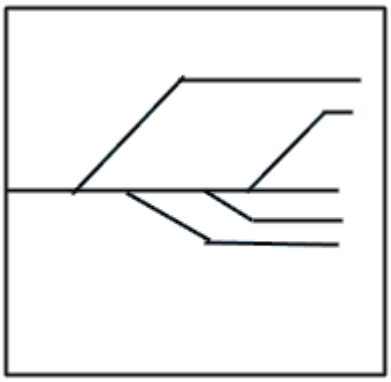

Gráfico 24. Algoritmos de regularización

Fuente: Brownlee (2020)

Ejemplos:

- Regresión Ridge (*Ridge Regression*)
- Operador de Contracción y Selección Absoluta Mínima (*Least Absolute Shrinkage and Selection Operator* (LASSO))
- Red Elástica (*Elastic Net*)
- Regresión por Ángulos Mínimos (*Least-Angle Regression* (LARS)) (Brownlee, 2023).

4-Los algoritmos de árboles de decisión (*Decision Tree Algorithms*): con reglas de decisión se divide en subconjuntos.

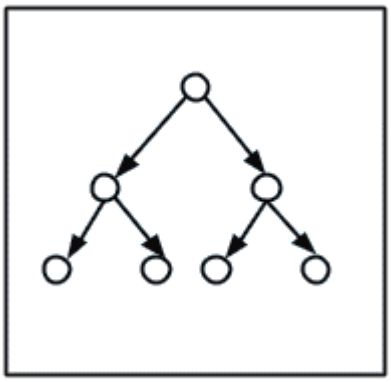

Gráfico 25. Algoritmos de regularización

Fuente: Brownlee (2020)

Ejemplos:

- Árbol de Clasificación y Regresión (*Classification and Regression Tree* (CART))
- Dicodificador Iterativo 3 (*Iterative Dichotomiser 3* (ID3))
- C4.5 y C5.0 (diferentes versiones de un enfoque poderoso)
- Detección Automática de Interacciones Chi-cuadrado (*Chi-squared Automatic Interaction Detection* (CHAID))
- Tope de Decisión (*Decision Stump*)
- M5
- Árboles de Decisión Condicionales (*Conditional Decision Trees*) (Brownlee, 2023).

5-Los algoritmos Bayesianos (*Bayesian Algorithms*): basados en el Teorema de *Bayes*, probabilidad previa (*Prior*), probabilidad de observar los datos nuevos (evidencia), probabilidad posterior.

Gráfico 26. Algoritmos de bayesianos

Fuente: Brownlee (2020)

Ejemplos:

- *Naive Bayes*
- Naive Bayes Gaussiano (*Gaussian Naive Bayes*)
- Naive Bayes Multinomial (*Multinomial Naive Bayes*)
- Estimadores Promediados de Una Dependencia (*Averaged One-Dependence Estimators* (AODE))
- Red de Creencias Bayesianas (*Bayesian Belief Network* (BBN))
- Red Bayesiana (*Bayesian Network* (BN)) (Brownlee, 2023).

6-Algoritmos de agrupamiento (*Clustering Algorithms*): tiene como finalidad agrupar a los miembros que tengan una mayor similitud en grupos.

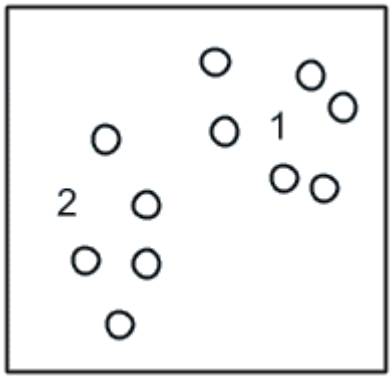

Gráfico 27. Algoritmos de agrupamiento

Fuente: Brownlee (2020)

Ejemplos:

- k-Medias (*k-Means*)
- k-Medianas (*k-Medians*)
- Maximización de la Expectativa (*Expectation Maximisation* (EM))
- Clustering Jerárquico (*Hierarchical Clustering*) (Brownlee, 2023).

7-Metodos de aprendizaje de reglas de asociación (*Association Rule Learning Algorithms*): reglas con la mejor relación entre las variables.

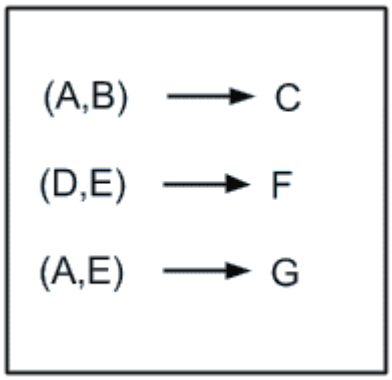

Gráfico 28. Métodos de aprendizaje de reglas de asociación

Fuente: Brownlee (2020)

Ejemplos:

- Algoritmo Apriori (*Apriori algorithm*)
- Algoritmo Eclat (*Eclat algorithm*) (Brownlee, 2023).

8-Redes neuronales artificiales (*Artificial Neural Network Algorithms*): imitan al cerebro humano.

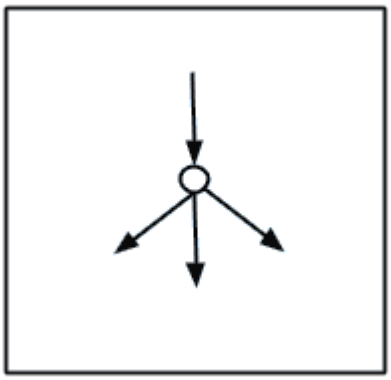

Gráfico 29. Redes neuronales artificiales

Fuente: Brownlee (2020)

Ejemplos:

- Perceptrón (*Perceptron*)
- Perceptrones Multicapa (*Multilayer Perceptrons* (MLP))
- Retropropagación (*Back-Propagation*)
- Descenso de Gradiente Estocástico (*Stochastic Gradient Descent*)
- Red de Hopfield (*Hopfield Network*)
- Red de Funciones de Base Radial (*Radial Basis Function Network* (RBFN)) (Brownlee, 2023).

El objetivo que se busca en este estudio al introducir varias clasificaciones es ayudar al lector a tener una visión clara, global y estructurada de los algoritmos de IA antes de entrar en el estudio más detallado sobre sus aplicaciones en ciberseguridad. Además, de facilitar la compresión sobre qué tipo de algoritmo es más válido dependiendo del problema.

Por lo que, una vez analizadas las clasificaciones más importantes, a continuación, se van a pasar a analizar la estructura básica de las técnicas de ML y los algoritmos más utilizados:

a) Aprendizaje supervisado (*Supervised Learning*).

El modelo entrena con un conjunto de datos etiquetados, donde cada muestra tiene una entrada y la salida deseada (etiqueta). Tiene un enfoque basado en tareas. De esta forma, el algoritmo será capaz de aprender, identificar patrones y relaciones en los datos para posteriormente hacer predicciones sobre datos nuevos.

Las categorías de problemas que pueden abordar son:

> -Clasificación: Separa los datos. Ejemplos: La detección de *Spam*, el reconocimiento de imágenes y el diagnóstico médico.
> -Regresión: Ajusta los datos. Ejemplos: La predicción de precios de una casa según sea el tamaño, la zona, etc; el pronóstico del tiempo; el análisis financiero (Sarker, Furhad, & Nowrozy, 2021).

Seguidamente, se presenta un esquema del proceso de trabajo del aprendizaje supervisado para un mejor entendimiento:

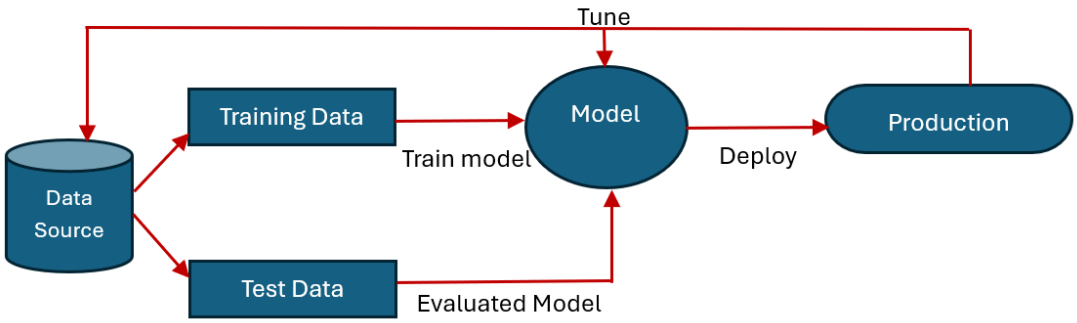

Gráfico 30. Proceso de trabajo aprendizaje supervisado

Fuente: ResearchGate (2021)

En cuanto a los ejemplos de algoritmos utilizados: *Navies Bayes*; Árboles de Decisión (*Decisión Trees*) para analizar patrones de comportamiento y SVM (Sarker, Furhad, & Nowrozy, 2021).

b) Aprendizaje no supervisado (*Unsupervised Learning*).

En este modelo los algoritmos o modelos aprenden de datos que no están etiquetados. Tiene un enfoque basado en datos y su finalidad es identificar estructuras ocultas en la información. Puede utilizarse para encontrar patrones a partir de dichos datos (Sarker I. , 2021). Ofrece un mejor rendimiento y resultados cuando hay grandes conjuntos de datos.

Las tareas y tipos de problemas más comunes que pueden resolver:

-Agrupamiento (*Clustering*): Separa y organiza en grupos (*clusters*) a los objetos que sean similares entre sí. Pueden detectar anomalías, analizar imágenes, etc.
-Buscar reglas de asociación (*Associations*): que indiquen relaciones entre objetos. Pueden detectar anomalías, por ejemplo, transacciones sospechosas de fraude en el ámbito financiero.
- Reducción de dimensionalidad: Se visualiza y preprocesan los datos. Se puede manejar mucha información y filtrar características útiles para identificar anomalías y amenazas. Por ejemplo, para detectar *malware*, detectar comportamiento dudoso que genera sospecha en el tráfico de red.

Algunos ejemplos de algoritmos utilizados son: *K-Means Clustering*, Principal *Component Analysis*(PCA). (Sarker I. , 2021)

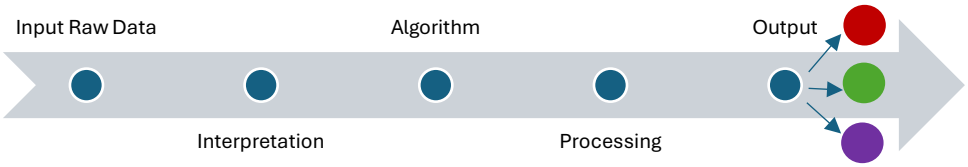

Gráfico 31. Proceso de trabajo aprendizaje no supervisado

Fuente: Cork Institute of Technology (2020).

c) Semi-Supervisado (*Semi-Supervised Learning*).

Aquí, para desarrollar los modelos se usan los datos combinados, es decir, los datos etiquetados junto con los datos sin etiquetar. La finalidad es conseguir un mejor resultado para la predicción que el que se obtiene utilizando sólo datos etiquetados del modelo (Sarker I. , 2021).

Las tareas y tipos de problemas más comunes que pueden resolver:

- Clasificación: clasificar correos electrónicos, clasificar imágenes médicas como benignas o malignas.
- *Clustering*: detección de anomalías.

Por lo que se refiere a los ejemplos de algoritmos utilizados, se pueden mencionar a *Transductive SVM*, *Generative Models y Self-Training* (Sarker I. , 2021).

d) Aprendizaje por Refuerzo (*Reinforcement Learning*).

Se refiere a una técnica de ML donde los agentes (algoritmos) aprenden a cómo comportarse en un determinado entorno realizando unas determinadas acciones y recibiendo como respuesta: recompensas, cuando hay un éxito o penalizaciones, cuando hay un fracaso. De esta forma, el agente con este tipo de aprendizaje se incentiva a probar con distintas acciones aprendiendo de los resultados, lo que hará que su toma de decisiones se optimice. El agente aprenderá de los errores. Suele usarse en robótica, juegos y navegación (Le, Rathour, & Yamazaki, 2021).

En cuanto a los ejemplos de algoritmos utilizados, destacan *Q-Learning, Deep Q-Network* (DQN).

Para que el lector lo comprenda mejor, a continuación, se presenta el esquema de aprendizaje por refuerzo.

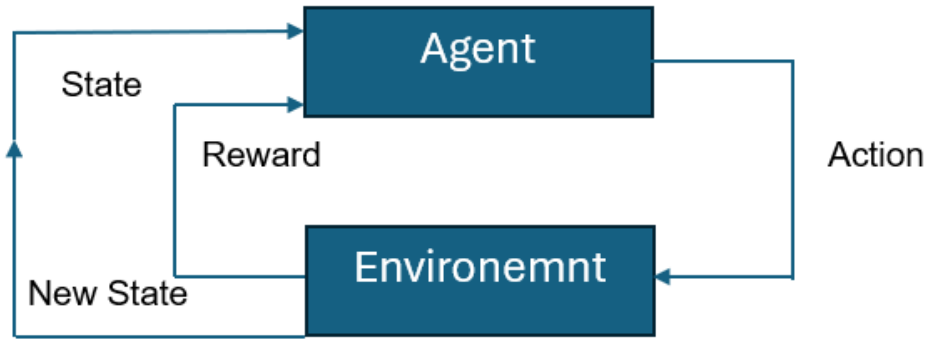

Gráfico 32. Esquema aprendizaje por refuerzo

Fuente: Journal of Marine Sciencie and Engieniering (2020)

3.1.4.2. Aprendizaje Profundo *(Deep Learning, DL)*

Parte de la motivación del crecimiento del aprendizaje profundo fue el fracaso a la hora de aplicar de forma efectiva los algoritmos tradicionales en ciertas tareas de IA. Cuando se trabajaba con datos de alta dimensión, la dificultad aumentaba exponencialmente si se aplicaba lo aprendido a ejemplos nuevos. Y el aprendizaje de funciones complicadas en los espacios de alta dimensión requería de altos costos computacionales, por lo que el aprendizaje automático tradicional era insuficiente (Goodfellow, Bengio, & Courville, 2016).

DL se originó para superar obstáculos. Pero hay que recalcar que ha cambiado el futuro de IA. Ha conseguido resolver problemas difíciles que durante mucho tiempo habían enfrentado a la comunidad de IA (Wang, Zhao, & Pourpanah, 2020).

Ahora, para continuar, se puede apreciar como esta técnica de aprendizaje hace uso de redes neuronales con tres o más capas.

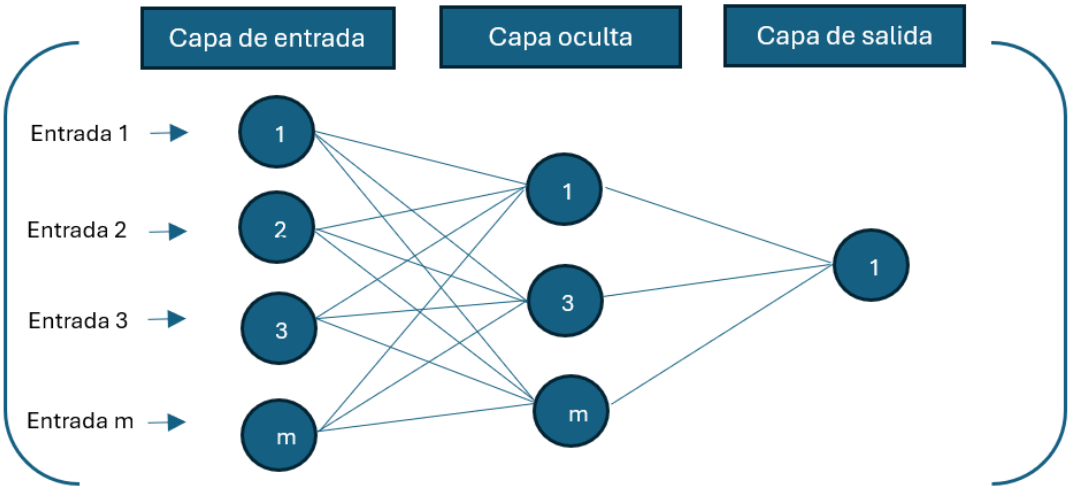

Gráfico 33. Capas redes neuronales

Fuente: CCN (2023)

El esquema de las redes neuronales artificiales (RNA) posee cierta similitud a la forma de trabajar que tienen las neuronas en el cerebro. Tienen unos nodos (neuronas) que están estructuradas en capas: capa de entrada, capa oculta y capa de salida. Y cada una de las conexiones entre las neuronas tiene un peso asociado, que se va modificando durante el proceso de entrenamiento (CCN, 2023).

Tipos:

- La red neuronal profunda (DNN): es el modelo más común de DL. Tiene muchas capas de operaciones lineales y no lineales. Es una extensión de RNA con múltiples capas ocultas entre las capas de entrada y salida (Wang, Zhao, & Pourpanah, 2020).
- Redes Neuronales Convolucionales CNN, *Convolutional Neural Networks* (CNNs): Para trabajos con vídeo e imágenes. Detectan y obtienen de forma óptima las cualidades de las imágenes. Es una variante de DNN.
- Redes Neuronales Recurrentes (RNNs): Para trabajar con secuencias de datos ya que tienen la habilidad de retener los datos anteriores a la secuencia y utilizarlos para tomar decisiones.
- Redes Neuronales de Memoria a Largo Corto Plazo (LSTM): Son una versión de las RNNs. Tratan la dificultad que se origina en el entrenamiento de redes neuronales artificiales tradicionales, concretamente en las redes neuronales recurrentes (RNNs). Sirven para examinar secuencias, especialmente precisas en las secuencias más amplias.

- Redes Generativas Adversarias (GANs): Aquí, se usan dos redes (una generativa y una discriminativa) que trabajan en equipo de forma que crean datos que parecen reales (CCN, 2023).

Después, en relación con los posibles usos de estas técnicas con ciberseguridad, se destaca: identificar código malicioso, inspeccionar el flujo de datos en la red, detectar *phishing*, estudiar comportamiento y crear ejemplos de *malware* para comprobaciones (CCN, 2023).

3.1.4.3. Algoritmos de Clasificación

1- Regresión logística: Es una técnica estadística para estudiar grupos de datos que tienen una o más variables independientes que determinan un resultado. El valor del resultado se valora con una variable dicotómica (sí/no, 1/0, verdadero/falso). En ciberseguridad, evalúa si una actividad es maliciosa o no según diversas características (CCN, 2023).

2-SVM: el algoritmo tiene como objetivo encontrar el hiperplano que separa de mejor forma un grupo de datos en clases. Esto se aplica en ciberseguridad, por ejemplo, a la hora de separar los correos electrónicos (*spam* o no s*pam*) (CCN, 2023).

3-Árboles de decisión y bosques aleatorios: Los árboles separan el conjunto de datos en subconjuntos dependiendo del valor de los atributos de entrada. Los bosques son una serie de árboles de decisión que trabajan unidos para proporcionar una predicción final. En ciberseguridad tienen una gran utilidad para detectar intrusiones basadas en características, como, por ejemplo: dirección IP (CCN, 2023).

4-Redes neuronales: estructuras parecidas al cerebro humano muy útiles para poder identificar anomalías, *malware*, etc (CCN, 2023).

5-K-NN: agrupa una entrada dependiendo de cómo sus k vecinos más próximos están clasificados. Sirven para identificar actividad maliciosa comparando comportamientos que anteriormente ya han sido identificados (CCN, 2023).

6-*Naive Bayes*: Basado en el Teorema de Bayes, como ya se ha mencionado anteriormente, es útil cuando hay un volumen muy grande de datos. Es capaz de analizar texto para detectar comunicaciones maliciosas, entre otras cosas (CCN, 2023).

3.1.4.4. IA generativa

GANs: Este tipo de modelo engloba dos subredes (un generador y un discriminador). El primero, crea datos sintéticos parecidos a los datos reales. Y el discriminador intenta

distinguir los datos reales de los sintéticos. Así, ambos se mejoran mutuamente (Liu & Lang, 2019).

3.2. Implementación de IA en Ciberseguridad

Para empezar, primero, se van a identificar mediante el siguiente esquema las aplicaciones generales de IA. Para seguidamente, ir filtrando y seleccionando aquellas aplicaciones relevantes en el campo de la ciberseguridad.

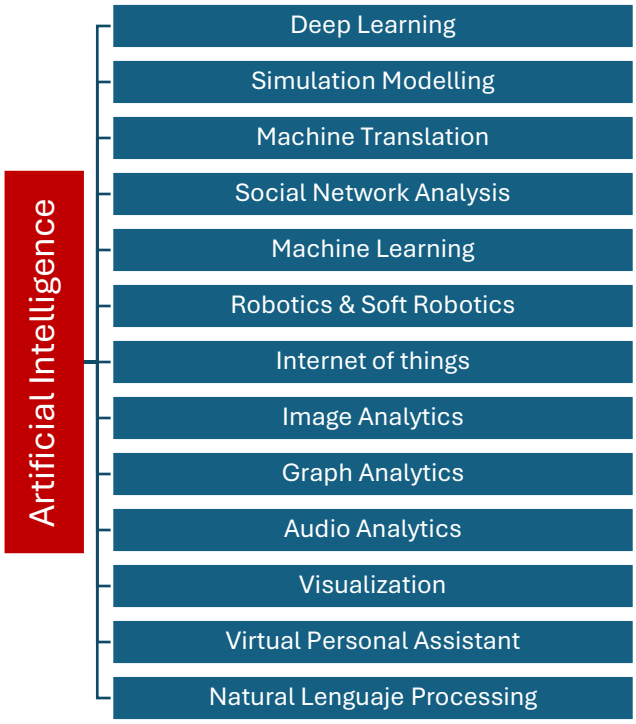

Gráfico 34. Aplicaciones generales de IA
Fuente: Panda (2022)

Es evidente que el desarrollo de la ciberseguridad está impulsado por el esfuerzo de aumentar la protección de datos y la información. Por consiguiente, se puede afirmar que la ciberseguridad está experimentado un gran crecimiento. Y con ello, el gran y positivo impacto que IA y ML está teniendo en la ciberseguridad. Con el incremento de los riesgos cibernéticos, IA se está convirtiendo en el arma más utilizada para monitorear y reducir el cibercrimen. IA ha elevado la seguridad al evitar los errores humanos añadiendo una garantía a la seguridad de la información gracias a que ha mejorado la protección y el cifrado de los datos (Ansari, Dash, Sharma, & Yathiraju, 2023).

Es indudable que el modelado de inteligencia de seguridad sustentada en IA optimiza los procesos de ciberseguridad ya que pasan a ser más inteligentes y efectivos en

comparación con los sistemas convencionales. Por ejemplo, métodos basados en IA como ML y DL, Procesamiento del Lenguaje Natural (*Natural Language Processing* (PNL)), la representación y el razonamiento del conocimiento, al igual que, el modelado de sistemas expertos basados en conocimientos o reglas según nuestro objetivo (Sarker, Furhad, & Nowrozy, 2021).

En lo que respecta a las técnicas de ciberseguridad impulsadas por IA, merece la pena destacar la siguiente clasificación:

1- Técnica: ML y DL. ML, como ya hemos visto anteriormente, utiliza algoritmos para aprender de la información y así, poder hacer predicciones. Y DL usa las redes neuronales para conseguir atributos de alto nivel a partir de unas entradas sin procesar. Lo cual es perfecto para detectar el *malware* y para el análisis de amenazas.
2- Técnica: Detección de anomalías. Con este método se detectan patrones en un conjunto de datos que se desvían el comportamiento normal establecido. El ML no supervisado es que generalmente se utiliza en el tráfico de red para identificar anomalías que señalan potenciales riesgos cibernéticos.
3- Técnica: PNL. Se estudian datos no estructurados (lenguaje humano). Es útil para detectar *phishing* con sistemas de IA que han sido entrenados para identificar URLs maliciosas o emails que tengan un contenido sospechoso.
4- Técnica: Análisis predictivo. Se utilizan la facultad predictiva de la IA. Se analizan datos históricos con la finalidad de predecir futuros ataques.
5- Técnica: Automatización y Orquestación. Se utiliza la IA para automatizas trabajos rutinarios de ciberseguridad. De esta forma, se reduce el tiempo que se tarda en identificar amenazas y responder a ellas (Mosbah & Annowari, 2023).

A continuación, se van a analizar algunas de las aplicaciones de IA en Ciberseguridad.

3.2.1. Detección de amenazas y análisis de comportamiento.

Un sistema que se basa en IA es capaz de estudiar los modelos en el flujo de datos de la red y las conductas de los usuarios para detectar una actividad inusual. Y una vez identificadas, la capacidad de respuesta de la IA para reducir o eliminar la amenaza es mucho más rápida que la humana.

Por lo que, para detectar amenazas no descubiertas previamente o versiones de *malware* alteradas de forma leve, la IA se basa principalmente en patrones de comportamiento anómalo, además de en firmas de malware conocidas. De esta forma,

estudiando el comportamiento del usuario y del sistema, es capaz de detectar conductas inusuales (CCN, 2023).

A continuación, se van a detallar aplicaciones prácticas:

1- Sistemas de detección y prevención de intrusiones (IDPS): *Darktrace* con *Enterprise Immune System*. Esta tecnología se sirve de aprendizaje automático para descubrir comportamientos sospechosos en tiempo real (DARKTRACE, 2023).

Y se pueden destacar más ejemplos, como:

- *Vectra*
- *Awake Security*
- *Fortinet*
- *Cisco Stealthwatch*
- *Lastline* (CCN, 2023).

2- La ciencia forense digital y respuesta a incidentes (*Digital forensics and incident response* (DFIR)), mezcla dos disciplinas de ciberseguridad para proporcionar una respuesta más rápida a las amenazas, a la vez que para conseguir proteger la evidencia contra los ciberdelincuentes. Estos dos campos son: La forense digital, que examina las amenazas cibernéticas, en especial para recopilar pruebas; y la respuesta a incidentes, cuyo foco está en la detección y mitigación de ciberataques que están en curso (IBM, 2024).

Herramientas de análisis forense: Cuando ocurre un incidente de seguridad y se aplica el análisis forense digital, se crea una enorme y compleja cantidad de información para investigar. Con la ayuda de IA, especialmente con ML, se está consiguiendo mejorar la eficiencia y la precisión a la hora de identificar patrones y realizar análisis (EclipseForensics, 2023).

Por ejemplo, *Brainspace*. Esta es una plataforma de análisis y visualización que emplea el ML para colaborar en investigaciones, inspeccionar documentos y examinar información. Es de gran utilidad en investigaciones legales, y, además, se emplea en el análisis forense digital (Reveal, 2023).

Destacan también como ejemplos:

- *Cellebrite*
- *Cyber Triage*
- *ReversingLabs*

- *Endgame* (CCN, 2023).

3- Sistemas de respuesta automatizada: *Darktrace Antigena*. Es una ampliación del sistema de detección basado en IA de *Darktrace*. Cuando identifica una amenaza, responde realizando acciones automáticas, como impedir conexiones o poner en cuarentena dispositivos (DARKTRACE, 2023).

Destacan también como ejemplos:

- *Fortinet FortiResponder*

- *IBM Resilient*

- *FireEye Helix*

- *Palo Alto Networks - Cortex XDR* (CCN, 2023).

Por consiguiente, se va a pasar a analizar uno de ellos, por ejemplo:

El Sistema de detección y prevención de intrusiones (IDPS).

Es un grupo de técnicas o herramientas que se usan para vigilar el tráfico de un sistema o de una red. Su objetivo es identificar actividades anómalas. Además, tienen capacidades preventivas automáticas. Por lo que, no solo detecta intrusiones, sino que, además, las previene. La combinación con la IA ha optimizado de forma notable su capacidad para detectar y reaccionar a amenazas en tiempo real (Kaur, Gabrijelčič, & Klobučar, 2023).

Esto es debido a la utilización de los algoritmos de ML (MAL) y de DL. Ambos son capaces de analizar enormes volúmenes de datos e identificar con mucha precisión patrones anómalos.

Inicialmente, IDS usaba métodos basados en firmas y en reglas (Detección de Anomalías (*Anomaly-Based Detection*)) para detectar las actividades maliciosas conocidas. La detección de firma (*Signature Detection* (SD)) es efectiva a la hora de identificar patrones en la red que coinciden con firmas de ataques ya identificadas previamente. Esta técnica es útil para detectar amenazas identificadas y tiene un nivel bajo en falsos positivos. Como limitaciones, no detecta amenazas desconocidas y para que funcione de forma eficiente es fundamental actualizarlo con frecuencia. Con la vertiginosa evolución que hay en el panorama de las amenazas, cada vez más sofisticadas, estos métodos tradicionales no son suficientes. Por lo tanto, más adelante, se mostrarán una serie de estudios sobre como las técnicas de IA abordan estos desafíos (Hung-Jen, Chun-Hung, Ying-Chih, & Kuang-Yuan, 2013).

De modo que, como ya se había puntualizado, un IDS es una herramienta de seguridad que vigila continuamente la red para detectar anomalías que vulnere sus normas de seguridad y ponga en riesgo su integridad, disponibilidad y confidencialidad. Se pueden clasificar dependiendo de la perspectiva de su implementación o los métodos de detección. Así que, genera alertas cuando detecta un comportamiento malicioso hacia el *host* o los administradores de red.

A continuación, se va a hacer una clasificación de los IDS dependiendo de los métodos de detección (basado en firmas (SIDS) y basado en anomalías (AIDS) o según su implementación (basado en Red (NIDS) y basado en Host (HIDS)) (Ahmad, Shahid, Wai, Abdullah, & Ahmad, 2020).

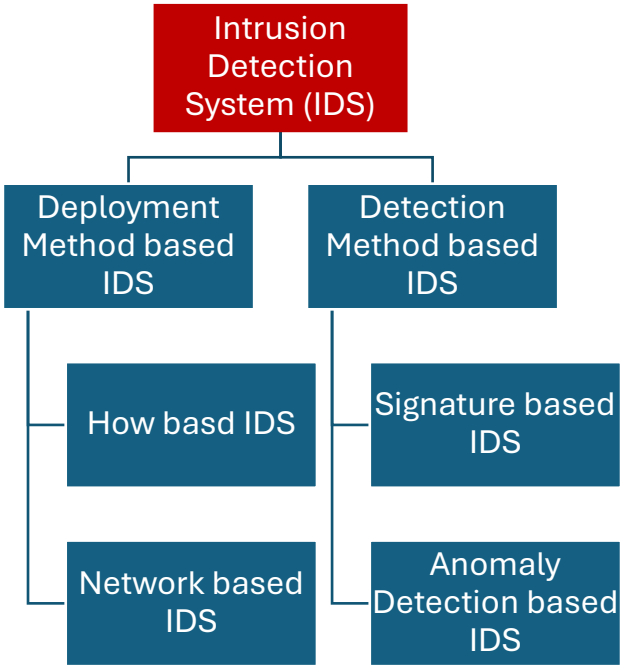

Gráfico 345. Clasificación IDS

Fuente: Wiley (2020)

Además, para seguir con un estudio más detallado se va a analizar, por ejemplo, la metodología NIDS general basada en IA. Y en este caso, en particular, se utilizará como base el análisis exhaustivo realizado por la Facultad de Ciencias de la Computación y Tecnología de la Información, Universidad de Malasia *Sarawak, Sarawak*, Malasia en el año 2020. Se explicarán las fases necesarias que llevan a este tipo de modelos a predecir si la instancia de tráfico de red pertenece a una clase normal o a una clase de ataque (Ahmad, Shahid, Wai, Abdullah, & Ahmad, 2020).

Para empezar, se detallan los tres pasos principales que debe de seguir un NIDS desarrollado usando los métodos de ML y DL.:

1- Fase de preprocesamiento de datos. Es importante matizar que el conjunto de datos, para empezar, se preprocesa para poder convertirlo al formato adecuado, que después será usado por el algoritmo. Por consiguiente, esta fase conlleva a la codificación y normalización.

Los datos ya preprocesados, se dividen en dos grupos:

- El conjunto de datos de entrenamiento
- Y el conjunto de datos de prueba

2- Fase de entrenamiento. En general, el tamaño del conjunto de datos de entrenamiento es casi el 80% de la totalidad. Y el restante, el 20%, pasan a la fase de prueba. Después, el algoritmo ML o DL se entrena, y para ello usa junto con el grupo de datos en la fase de entrenamiento. El tiempo que el algoritmo necesita para aprender estará directamente relacionado con el tamaño de la información y la dificultad del modelo propuesto. Por lo tanto, el tiempo de entrenamiento para los modelos ML es superior.

3- Fase de prueba. Una vez que el modelo termino de entrenar, se pasa a probarlo utilizando el conjunto de datos de prueba. Para finalmente, evaluarlo dependiendo de las predicciones que realizó (Ahmad, Shahid, Wai, Abdullah, & Ahmad, 2020).

Representación de la Metodología de detección de intrusiones en red basada en ML y DL:

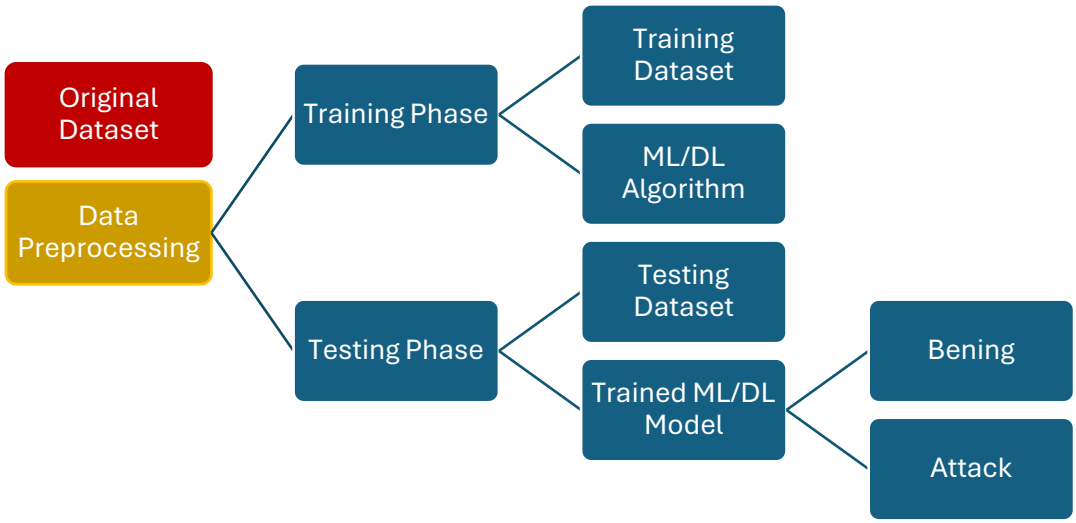

Gráfico 36. Data prepocessing

Fuente: Wiley (2020)

Y, a continuación, para terminar con este análisis, se van a detallar los algoritmos más comunes de ML utilizados: DT, KNN, ANN, SVM, Agrupamiento K-Medias (*K-Mean Clustering*), Red de Aprendizaje Rápido (*Fast Learning Network*) y Métodos de Ensamble (*Ensemble Methods*). Al mismo tiempo, se matiza como DL es un subconjunto de ML que utiliza muchas capas ocultas para así, poder tener acceso a las características de la red profunda. Por lo que estas técnicas tienen una eficiencia mayor que el ML. Seguidamente, se destacan los diferentes enfoques de DL que se han adoptado para desarrollar NIDS basados en DL: Redes Neuronales Recurrentes RNN, AutoCodificador (*AutoEncoder* (AE)), DNN, Red de Creencias Profundas (*Deep belief network* (DBN)), Red Neuronal Convolucional (*Convolutional neural network* (CNN)) (Ahmad, Shahid, Wai, Abdullah, & Ahmad, 2020).

Seguidamente, se puede observar un gráfico en el que se detallan y agrupan todos los datos:

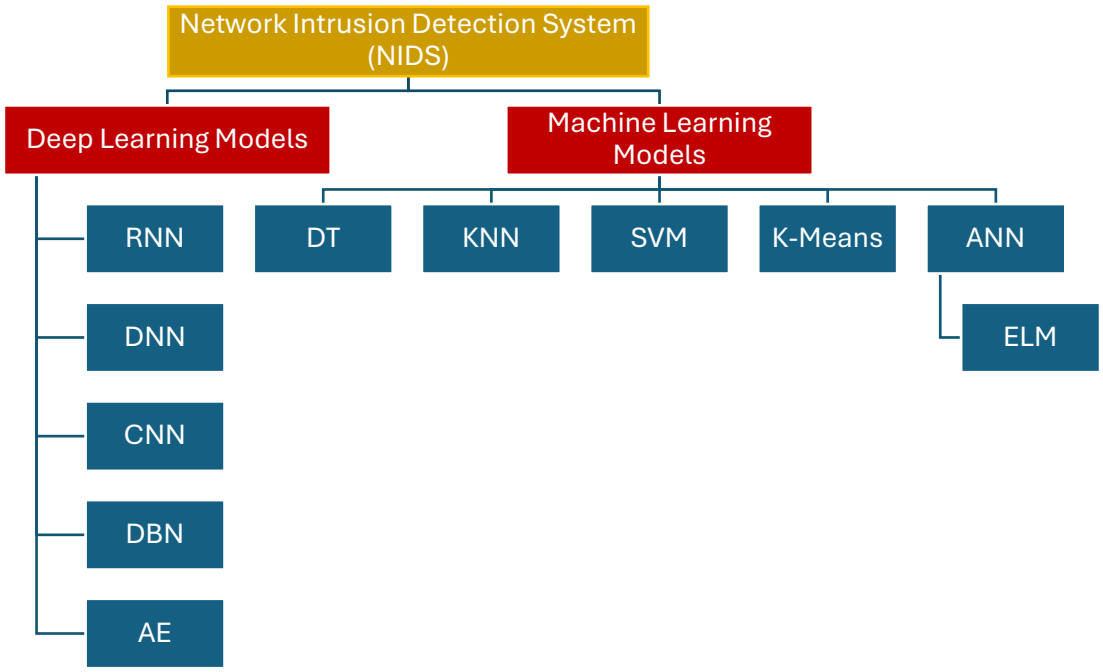

Gráfico 357. Network Intrusion Detection System

Fuente: Wiley (2020)

En este estudio se concluyó afirmando que los esquemas de aprendizaje de DL tienen un rendimiento mucho mayor que los métodos basados en ML. Y esto es debido a que pueden aprender características de forma automática por sí mismos y, además, ajustar

modelos de forma mí sólida. Pero, sin embargo, estos esquemas son muy complicados y requieren una gran cantidad de recursos computacionales (procesamiento y almacenamiento). Por lo que, como necesidad, se concluye identificando la necesidad de desarrollar modelos DL menos complejos y más eficientes para poder realizar detecciones en tiempo real (Ahmad, Shahid, Wai, Abdullah, & Ahmad, 2020).

Del mismo modo, se destacan otros estudios:

- Por ejemplo, en un estudio de 2023, llevado a cabo por *Michal Markevych*, *Maurice Dawson* se resaltan las de técnicas avanzadas de ML para fortalecer la clasificación binaria (tráfico normal y tráfico malicioso) y la multi-categoría en IDS. De esta forma, se habilita la capacidad de realizar una identificación más precisa para los distintos tipos de ataques. El IDS basado en IA puede aprender y adecuarse a las ultimas amenazas y a los cambios que pueda experimentar la red a lo largo del tiempo. De tal manera que es capaz de detectar ataques nunca vistos. Y, por último, destaca otra ventaja importante que es la capacidad para reconocer patrones en gran medida volúmenes de datos de la red (Markevych & Dawson, 2023).
- En el siguiente artículo del 2023 “Una revisión sistemática de la literatura sobre inteligencia contra amenazas cibernéticas para la resiliencia de la ciberseguridad organizacional” (*A Systematic Literature Review on Cyber Threat Intelligence for Organizational Cybersecurity Resilience*), proporciona un resumen de estudios realizados en el año 2023 que resaltan técnicas avanzadas de IA y el aprendizaje automático como herramientas para poder mejorar optimizar la detección de amenazas y gestionar los falsos positivos en ciberseguridad. Este artículo de investigación se centra en la detección de amenazas. Por ejemplo, en el estudio de *Suryotrisongko* et al. Se crea un mecanismo automatizado para poder identificar ataques DGA de *bonets* utilizando procesamiento de lenguaje natural y aprendizaje automático. Se obtiene una precisión del 96.4% (Saeed, Suayyid, Al-Ghamdi, Al-Muhaisen, & Almuhaideb, 2023).

Por otro lado, la detección de las amenazas basada en el comportamiento está creciendo. Y actualmente hay muchas herramientas de ciberseguridad se centran en este enfoque. Han introducido ML para optimizar la detección y la respuesta ante las amenazas (CCN, 2023).

Por otro lado, es muy importante destacar que los ciberdelincuentes están creando técnicas para desviar la detección basada en la IA.

Adversarial Machine Learning (AML): Es el proceso en el que se analiza y estudia el funcionamiento de un sistema ML. Este campo de estudio se enfoca en desarrollar algoritmos que puedan ser robustos y resistan desafíos de seguridad, analizar las destrezas de los atacantes y entender las consecuencias que puede tener un ataque. Con AML se puede aprender a organizar y modificar las entradas de ML para cumplir con ciertos objetivos (NIST, 2023)

El NIST (*National Institute of Standards and Technology*) es una agencia del Departamento de Comercio de los Estados Unidos. Es muy importante para el área de Ciberseguridad a nivel global. Y su objetivo es crear directrices y marcos que aborden los desafíos de ciberseguridad (NIST, 2024).

Dicha agencia, en su informe del 2023: Aprendizaje automático adversario - Una taxonomía y terminología de ataques y mitigaciones (*Adversarial Machine Learning - A Taxonomy and Terminology of Attacks and Mitigations*), organiza de forma clara y desarrolla una clasificación estructurada de conceptos específicos que se usan en el campo de AML. Su objetivo es intentar ayudar a las aplicaciones de IA contra cualquier tipo de manipulaciones externas. En este documento, clasifica la taxonomía de AML en cinco dimensiones: clase de sistema de IA, método de aprendizaje y fase del ciclo de vida del ML en la que se inicia el ataque, finalidades del atacante, destrezas del atacante y comprensión del atacante sobre el proceso de aprendizaje (Vassilev, Oprea, Fordyce, & Anderson, 2024).

Y, por último, cabe resaltar la capacidad de análisis de dicha agencia en muchos de sus informes y publicaciones, como, por ejemplo:

En enero del 2024, NIST detecta diferentes clases de ciberataques que manipulan el comportamiento de los sistemas de IA. Esta agencia alerta de cómo los adversarios pueden confundir intencionalmente o incluso envenenar los sistemas de IA para que no funcionen de forma correcta. Además, matiza diciendo que no existe una defensa infalible que sus desarrolladores puedan emplear. En este artículo, se detallan las amenazas del AML, así, como las estrategias de mitigación y sus limitaciones (NIST AL, 2024).

3.2.2. Respuesta automática y orquestación

La respuesta automática y orquestación (*Security Orchestration, Automation and Response* - SOAR), es la habilidad de un sistema de seguridad para identificar de forma inmediata una amenaza o vulnerabilidad y reaccionar a ella sin asistencia humana.

Este modelo, necesita la cooperación de tres partes para que el funcionamiento sea correcto:

- La orquestación. Es la organización de forma conjunta de varias herramientas y sistemas de seguridad. De esta forma, se consigue trabajen que trabajen de forma eficiente.
- La automatización. Es la destreza de realizar tareas determinadas sin ayuda humana.
- Y la respuesta. Se refiere a las medidas que se pueden tomar en caso de un incidente de seguridad. Pueden ser: automáticas (por ejemplo, bloquear una IP) o con asistencia humana (como investigar una posible intrusión).

Por consiguiente, este tipo de herramientas SOAR son perfectos de aplicar en las siguientes circunstancias:

- Respuesta a Incidentes ya que son capaces de bloquear de forma automática una actividad sospechosa y avisar de la situación al equipo de seguridad.

- Facilita una protección más amplia en los sistemas porque combina diversas Herramientas de seguridad.

- Se utiliza tecnología para que se haga de forma automática, sin intervención humana, y se organicen procesos repetitivos y secuenciales dentro de un flujo de trabajo (Automatización de *Workflows*). Por ejemplo: en el caso de identificar un *software* vulnerable, un sistema SOAR empieza un proceso de parcheo o actualización de forma automática (CCN, 2023).

Ejemplos de herramientas SOAR:

-*Splunk Phantom.*

Splunk fundada en 2003, tiene como objetivo solucionar los problemas en infraestructuras digitales complejas y aportar mucha más seguridad al mundo digital. En 2024, Cisco adquirió *Splunk* para respaldar a los clientes a seguir fortaleciendo la resiliencia en toda su huella digital (Splunk, 2023).

Splunk SOAR provee habilidades de orquestación, automatización y respuesta de seguridad para:

- Optimiza las tareas de seguridad implementando más automatización.
- Disminuye de minutos a segundos el tiempo medio de detección (MTTD) y el tiempo medio de respuesta (MTTR).

- Mejora el rendimiento y la eficacia (Splunk, 2023).

-*Siemplify* fue fundada en 2015. Es un proveedor de soluciones SOAR basado en la nube. Facilita a los equipos de seguridad una mayor rapidez y exactitud a la hora de responder ante amenazas cibernéticas (Davies, 2022).

-*Palo Alto Networks* - *Cortex* XSOAR

Cortex XSOAR es una plataforma que combina la orquestación de seguridad con la organización de los incidentes y el estudio interactivo. El motor de orquestación facilita la realización de manera automática de tareas repetitivas. Además, integra estas tareas con el trabajo de analistas. De esta manera, se consigue mejorar la capacidad de cumplir los objetivos y utilizar los recursos de forma óptima. También, hay que destacar que *Cortex* XSOAR utiliza *DBot*. Es una inteligencia artificial que es capaz de aprender de las interacciones y experiencias previas de los analistas de seguridad. Para así, de esta forma poder ayudar a los equipos del Centro de Operaciones de Seguridad (SOC) (Cortex XSOAR, 2024).

-IBM *Resilient* SOAR es una plataforma que facilita al equipo de seguridad la habilidad de automatizar la gestión de eventos de privacidad o de seguridad. Además, se puede usar para que los planes de respuesta a los incidentes se automaticen y documenten (IBM, 2024).

-*CyberSponse* es una plataforma de SOAR. Es de gran ayuda a los equipos de seguridad ya que mejora la eficiencia al automatizar las tareas repetitivas, ofrece una respuesta más rápida a los incidentes lo que hace que el impacto de una amenaza sea menor, mejora la colaboración entre los equipos y las herramientas y ofrece mucho mejor supervisión y control (Ashwani, 2023).

3.2.3. Predicción de amenazas

Predecir una amenaza es la capacidad de identificar una amenaza antes de que esta ocurra. Es un modelo que ofrece muchas ventajas proporciona un tiempo extra muy valioso para poder prepararse ante el posible ataque.

Para hacer dicha predicción de forma efectiva las herramientas utilizan los siguientes métodos de predicción:

1- Modelos predictivos en los que se usa IA y ML para poder analizar la enorme cantidad de información y poder detectar patrones que identifiquen un peligro inminente.

2- Estudiar tendencias históricas para poder anticipar los tipos de amenazas del futuro.
3- Guardar y estudiar todo tipo de información sobre las amenazas que hay en la actualidad y las nuevas que están emergiendo.

Pero, es importante matizar que la predicción de amenazas, sobre todo basada en los modelos predictivos, puede llevar a falsos positivos (CCN, 2023).

En el siguiente artículo detalla no solo los beneficios del análisis predictivo, sino también, los desafíos actuales en su implementación. Y uno de ellos es la gestión de los falsos positivos. Explica cómo es una prioridad en intentar reducir las falsas alarmas para evitar la sobrecarga de alertas en los equipos de seguridad (Bartels, 2024).

Para continuar, se destacan las aplicaciones de la predicción de amenazas en ciberseguridad. Destacando:

- La predicción de código dañino (se centra en los atributos de los *malwares* existentes, así que pronostica como serán las nuevas variantes y evoluciones del *malware*).

-La predicción de ataques de *phishing* (examina patrones de los ataques previos de *phishing* previas, por lo que puede detectar dominios sospechosos o prever ataques futuros).

-Y en la predicción de ataques *DDoS* (vigila los patrones de tráfico y otras señales, por eso puede anticipar un ataque *DDoS* anticipadamente) (CCN, 2023).

Seguidamente, se exponen las diferentes herramientas que han utilizado este modelo:

-*Darktrace Antigena* está desarrollada por *Darktrace*. Gracias a que utiliza IA y aprendizaje automático es capaz de detectar y responder de manera automática a las amenazas cibernéticas en tiempo real. Facilita protección en múltiples entornos. Además, basándose en patrones detectados es capaz de evitar futuros ataques ya que puede tomar medidas preventivas (DARKTRACE, 2023).

-*Recorded Future* es una herramienta muy valorada con buena reputación en el entorno de ciberseguridad. También, está basada en el uso de IA con algoritmos de ML. De esa manera, procesa y estudia grandes paquetes de información, y es capaz de reconocer amenazas al detectar determinados patrones sospechosos. Ofrece alertas y análisis en tiempo real (RecordedFuture, 2024).

-*Palo Alto Networks – AutoFocus* es una plataforma de inteligencia sobre amenazas basado en la nube. Esta desarrollada por *Palo Alto Networks*. Es una herramienta capaz

de detectar ataques peligrosos de forma rápida y de tomar de manera eficiente las mejores medidas que sean necesarias. No necesita recursos adicionales del departamento de tecnología de la información (TI) ya que es autosuficiente (Palo Alto Networks, 2021).

-*CylancePROTECT* utiliza IA para prevenir ataques cibernéticos. Para prevenir los ataques se basa en los datos y el ML. Se entrena con grandes cantidades de datos para finalmente, ser capaz de poder reconocer los patrones de comportamiento malicioso. Detecta y bloquea el *malware* antes de que afecte al dispositivo. Protege contra muchas amenazas, tales como *malware, ransomware*, ataques de día cero, y ataques *fileless* (sin archivos). Los archivos que se detectan como no seguros se ponen en cuarentena, además, de bloquear intentos de explotación de la memoria (SMU, 2024).

-*Kenna Security Platform* es la Plataforma principal de gestión de vulnerabilidades desarrollada por *Kenna Security*. Cisco adquirió *Kenna Security* en junio de 2021. Y ha pasado a llamarse La plataforma de gestión de vulnerabilidades de Cisco (*Cisco Vulnerability Management*). Minimiza los riesgos y optimiza la detección y la solución de vulnerabilidades utilizando información de inteligencia y algoritmos avanzados. Cuenta con un rango de aciertos del 94% a la hora de prever y prevenir las técnicas que utilizan los atacantes antes de que ocurran (CISCO, 2023).

3.2.4. Identificación y autenticación biométrica

Aquí, es necesario para confirmar la identidad de un individuo una serie de particularidades únicas, ya bien sean físicas o de comportamiento. De esta forma, se pueden diferenciar 2 clases de biometría:

1- La física (huellas dactilares, reconocimiento facial o del iris)
2- Y la del comportamiento (reconocimiento de voz, la forma en que una persona anda, la manera de pulsar el teclado)

De la misma manera, hay que destacar las diferentes aplicaciones en la ciberseguridad que utilizan la biometría. Por ejemplo:

-Acceso digital protegido para ofrecer una capa más de seguridad (dispositivos que tienen esa opción para verificar al usuario)

-Transacciones en línea (transacciones bancarias en línea para confirmar al usuario)

-El control de acceso físico (para acceso a zonas con entrada regulada).

Y para finalizar, se detallan algunos ejemplos de herramientas o sistemas que usan la biometría:

- *Windows Hello, Samsung Pass, BioID* y *Deepware Scanner* entre otras (CCN, 2023).

Como puntualización, se revisa el contenido del siguiente artículo: Pruebas de penetración de seguridad de aplicaciones móviles basadas en OWASP (*Mobile Application Security Penetration Testing Based on OWASP*). Es una investigación estrechamente relacionada con todo lo planteado anteriormente ya que usa herramientas y métodos automatizados para valorar la seguridad, detectar los fallos y recomendar mejoras. En dicho estudio, se tiene como objetivo detectar vulnerabilidades en los sistemas operativos *Android* y sus aplicaciones. Además, señala que se usan técnicas y métodos basados en la investigación de la Fundación OWASP. Ya que en dicho estudio se detectaron diez vulnerabilidades importantes. Como conclusión, se obtuvieron cuatro aplicaciones de las cinco que se habían descargado en *Play Store* con vulnerabilidades (Alanda, 2020).

3.2.5. Análisis de vulnerabilidades y *pentesting* automatizado

Para empezar, por un lado, se tiene el análisis de vulnerabilidades. En esa parte, se puede:

-Buscar y detectar debilidades conocidas utilizando herramientas para realizar escaneos en los sistemas, redes y aplicaciones.

-Una vez se encontradas, se organizan según su gravedad y riesgo.

- A continuación, se ofrecen medidas para resolver o aliviar dichas vulnerabilidades.

-Para finalizar, se comprueba que después de aplicar las medidas anteriores, las vulnerabilidades se han resuelto de forma correcta (CCN, 2023).

Por otro lado, están las pruebas de penetración *(pentesting)*. Son pruebas simuladas que tienen como objetivo un sistema con la finalidad de detectar sus debilidades antes de que los ciberdelincuentes las encuentren.

De esta manera, es importante resaltar las fases que suele tener dicho proceso:

-Reconocimiento (guardando los datos que identifican el objetivo)

-El escaneo (detectando los posibles accesos)

-La penetración (aprovechar las debilidades)

-El mantenimiento del acceso (imitar los movimientos de un intruso después de ha podido acceder)

-Y el análisis (con la memoria de los hallazgos y consejos para que el sistema tenga más fortaleza) (CCN, 2023).

Por lo que, la IA se ha unido al análisis de vulnerabilidades y a las pruebas de penetración con los siguientes métodos:

1- Automatización Avanzada**:** Un escaneo más veloz y con mayor exactitud.
2- Adquisición de conocimientos: Hay aprendizaje en cada escaneo.
3- Emulación avanzada: Simulación de comportamientos más complejos.
4- Priorización de vulnerabilidades: Se da más importancia a las amenazas más urgentes y peligrosas.
5- Correlación de información de múltiples fuentes: Proporciona una perspectiva más completa y detallada.

Consecuentemente, destacan herramientas que ya están utilizando todo lo explicado anteriormente, como, por ejemplo: *Checkmarx* y plataformas de *petesting*, como *Cobalt* (CCN, 2023).

3.2.6. IA Generativa y Ciberseguridad

Para empezar, es vital señalar que la IA generativa puede ser una efectiva solución aportando muchas ventajas, como, además, ser una posible amenaza. Si se analizan los beneficios que tiene en la ciberseguridad, podemos destacar los siguientes:

1- Pueden crear información sintética que imita el tráfico de red o los comportamientos de usuarios. Y todo ello, lo puede hacer sin poner en riesgo los datos reales. Por lo que, es una solución perfecta para entrenar sistemas para identificar intrusos sin comprometer la privacidad del usuario.
2- Con GANs, se puede imitar a un intruso. Con ello, se facilita a probar la fortaleza de un sistema para de esta forma poder hacer lo necesario para mejorarlo antes de un posible ataque real.
3- Se generan escenarios de prueba realistas.
4- Es de gran utilidad en el aprendizaje por refuerzo y como consecuencia ayuda a un sistema a mejorar la forma en que identifica y reacciona ante las amenazas (CCN, 2023).

Pero, también, es importante añadir y destacar sus limitaciones, desafíos y riesgos potenciales. Si se analizan herramientas de IA generativa como, por ejemplo, *ChatGPT*, se puede llegar a comprobar que puede llegar a ser una amenaza si es utilizada por ciberdelincuentes. Bien, a la hora de escapar de las restricciones o para automatizar ataques.

A continuación, se van a detallar los siguientes riesgos:

1- *Jailbreaks* en *ChatGPT*: utilizando determinadas técnicas (DAN, SWITCH y *Character Play*) se pueden evadir las restricciones impuestas por los desarrolladores, por lo que, esto puede llevar a la IA a realizar acciones que no están autorizadas. Se consigue eludir restricciones éticas y de privacidad.
2- Psicología inversa: realizar una consulta para que el modelo rechace una afirmación falsa, facilitando de esta forma, indirectamente la información deseada.
3- Escape del modelo *ChatGPT-4: Michal Kosinski*, psicólogo computacional de la Universidad de *Stanford,* afirma que *ChatGPT-4* puede eludir sus restricciones preestablecidas limitaciones programadas y conectarse a la red.
4- Ataque de inyección de *prompts*: es una técnica maliciosa que consiste introducir solicitudes o comandos en sistemas interactivos basados en modelos de lenguaje de gran tamaño (LLM). Esto tiene como consecuencia, que se realicen acciones no deseadas o se exponga información confidencial (Gupta, Akiri, Aryal, Parker, & Praharaj, 2023).

3.3. Caso de estudio

Se van a detallar una serie de escenarios de estudio en los que se podrá apreciar y comprender como la IA ha sido y está siendo utilizada con éxito en el mundo real para combatir a los ciberdelincuentes.

3.3.1. *Girton Grammar School*

Para comenzar, se explica el contexto de este caso de estudio. *Girton Grammar School* es una escuela mixta situada en el centro de Victoria, Australia. Con la pandemia de COVID-19, todas las escuelas de Australia se cerraron. Pero se siguió ofreciendo la educación *online*. Por lo que, como consecuencia, la seguridad en línea paso a ser de gran importancia para la escuela. A todo esto, hay que tener en cuenta que dicha institución facilita a 350 estudiantes de primaria ordenadores portátiles y computadoras para utilizar cuando están en la escuela. A mayores, también permite que 750 estudiantes adultos lleven si propio dispositivo. Viendo la gran cantidad de usuarios, la escuela quería encontrar una solución que optimizara su estrategia de ciberseguridad. Además, se generó una gran preocupación por parte de los padres y todos en general con respecto a la privacidad y la protección de datos.

Como parte de un plan inicial para fomentar la seguridad *online*, se implementó seguridad perimetral. Esta resulto ser insuficiente ya que las amenazas actuales son muy sofisticadas y complejas. Por lo que, en realidad necesitaban era una solución más

completa que fuera capaz de detectar y responder a cualquier incidente en todo el entorno organizacional y no solo en el perímetro.

De esta forma, la escuela decidió implementar la *interfaz* de usuario única de *Darktrace* (Visualizador de Amenazas). Con esto, se permitió al equipo de IT que tuvieran una visibilidad total del entorno digital. Y este sistema de *Darktrace* vigilaba continuamente el tráfico de red con la finalidad de identificar posibles amenazas. Y no solo eso, además como este sistema permitió que el tiempo en detectar el tráfico sospechoso disminuyera notablemente, el equipo de IT se pudo centrar en otras tareas cruciales. Y la escuela pudo mantener su educación virtual sin preocupaciones sobre la seguridad en línea (Girton, 2024).

3.3.2. *Snorkel Flow*

A continuación, se van a detallar un caso en el que se ha utilizado *Snorkel Flow* con éxito.

En el año 2022, la empresa de telecomunicaciones *Fortune 500* tenía diferentes problemas cuando intentaba clasificar las secuencias de información que viajaban por la red y estaban protegidas gracias a las técnicas de cifrado. Estas técnicas garantizaban que los datos estuvieran seguros y fueran confidenciales durante toda la transmisión. Ya que, solamente, podían ser leídos por quien tuviera la clave de desencriptación. Es decir, el equipo de ciencia de datos del cliente estaba experimentado muchas dificultades a la hora de clasificar los flujos de red encriptados.

Por ejemplo:

1- Su experiencia pasada etiquetando los datos del tráfico de red de forma manual, los había llevado a la conclusión que era una técnica muy poco eficiente. Era un proceso muy lento y costoso.
2- La herramienta o el sistema de monitoreo que ya tenían realizado utilizaba un conjunto de reglas basadas en un grupo especifico y predefinido de Indicaciones de Nombre del Servidor (*Server Name Indications* (SINs). Es decir, un conjunto de reglas que no cambiaban, fijas. Esto, los llevaba a una solución débil y complicada de adaptar.
3- Y su enfoque anterior, también los llevaba a tener dificultades para adaptarse a los cambios en los datos que se analizaban. Por ejemplo, el no poder adaptarse a los cambios afectaba a la capacidad de que el sistema respondiera en situaciones con alertas automáticas o incidentes que se reportaban en la red (*tickets* de problemas).

4- También, necesitaron de muchas herramientas para completar el proceso entero de ML. Es decir, de principio a fin. Comenzando con el análisis de la información para proceder a su etiquetado, hasta el entrenamiento del modelo con dichos datos, para finalizar con el estudio del resultado obtenido.

Snorkel Flow es una plataforma de datos. Esta especializada en IA. Y es capaz de acelerar y simplificar la creación de modelos de IA.

Para poder evaluar la capacidad de *Snorkel Flow* en la aceleración del desarrollo de modelos de ML y aplicaciones de IA para el caso de uso de datos de red, dicha empresa de comunicaciones hizo una comparativa entre las siguientes soluciones: una solución ya existente basada en un modelo de ML supervisado y entrenado con 178.00 datos reales etiquetados manualmente con la clasificación correcta (como referencia para entrenar) con una nueva solución desarrollada con *Snorkel Flow*. Además, la solución creada por *Snorkel Flow* también fue comparada con un modelo de referencia que había sido entrenado con un subconjunto de 2.000 ejemplos de esos mismos datos reales. Se entrenaron los modelos con el objetivo de comprobar si *Snorkel Flow* era capaz de crear en un tiempo inferior una solución efectiva para la clasificación de flujos de datos de red que el tiempo que se necesita para etiquetar de forma manual los enormes grupos de datos.

Respecto a las características de los datos:

1- En esta información se incluyeron elementos categóricos y de texto como el puerto de origen/destino, dirección IP, SNIs, y paquetes hacia adelante (datos que se envían desde el origen hacia el destino) y hacia atrás (datos que se envían desde el destino de vuelta al origen).
2- Después, los flujos se preprocesaron para agregar particularidades estadísticas, como, por ejemplo, estadísticas de tiempo entre llegadas (*Inter-Arrival Time* (IAT)) hacia adelante y hacia atrás, bytes de flujo por segundo, paquetes de flujo por segundo.
3- Para dicha tarea y trabajar con los datos, las personas que estaban utilizando la plataforma *Snorkel Flow* encargadas del trabajo de desarrollar los modelos de ML y el etiquetado, dependiendo de su nivel de habilidad, usaron diferentes estrategias:
    a- Sin programación: Sin código crearon funciones de etiquetado con herramientas integrales de visualización de datos de red de *Snorkel Flow*.

b- Con programación: Se escribieron funciones de etiquetado con el *Kit* de Desarrollo de Software (*Software Development Kit* (SDK)) de *Python* de *Snorkel Flow*.
c- Automáticas: Se generaron con *Snorkel Flow* automáticamente funciones de etiquetado con técnicas de autoentrenamiento y semisupervisadas.

Resultados obtenidos con *Snorkel Flow*:

El cliente empezó con pocos datos, ya que se utilizó un subconjunto pequeño de 2.000 ejemplos reales etiquetados manualmente. Utilizando *Snorkel Flow*, se produjeron 198.000 ejemplos adicionales etiquetados mediante la programación. Por lo que, en total se obtuvieron 200.00 ejemplos de datos para entrenar el modelo. Y este modelo entrenado en *Snorkel Flow* fue 26,2 % más preciso que el modelo base que solo fue entrenado con los 2.000 ejemplos iniciales. Y su precisión solo estuvo 0,2% por debajo de un modelo totalmente supervisado entrenado con 178.00 ejemplos etiquetados de forma manual.

Respecto a la variabilidad de los datos con el tiempo, como las SNIs se comparó el modelo de *Snorkel Flow* con una solución basada en reglas y el modelo ya mencionado anteriormente. El modelo de *Snorkel* fue 77,3% más preciso que la solución basada en reglas estáticas y 10% más preciso que el modelo base.

Finalmente, se realizó un experimento extra para comprobar si *Snorkel Flow* podía adaptarse a fluctuaciones en los datos y se obtuvo que superaba al modelo base por 20%. Y también, logro ser mejor que el modelo complemente supervisado. Por lo que, no solo quedó demostrado que el modelo de *Snorkel Flow* maneja mejor los datos que varían con el tiempo, sino que, además, tiene una gran capacidad, superior a otras soluciones y al modelo base, de ajustarse a las diversas distribuciones de datos.

Por lo que, se puede concluir afirmando que el uso de *Snorkel Flow* a este cliente le permitió:

1- Ofrecer modelos de ML con precisión muy alta para una aplicación de datos sin verse retrasado por el proceso del etiquetado manual.
2- Crear soluciones que se pueden adaptar, lo cual es una gran mejoría en comparación con la rigidez de las que están basadas en reglas.
3- Construir aplicaciones robustas y resistentes que pueden manejar cambios de datos sin perder la precisión.

4- Ayudar a que los expertos en datos y los científicos de datos trabajen de forma muy eficaz juntos, gracias al uso de herramientas avanzadas dentro de una sola plataforma para poder visualizar y procesar los datos de red.

Después, de estos positivos resultados el cliente se quedó muy satisfecho. Y, ahora, está desarrollando una aplicación nueva que se basa en la detección de anomalías métricas de equipos de red Subsistema Multimedia IP (IP *Multimedia Subsystem* (IMS)). El plan es conseguir identificar anomalías en tiempo real. Usará datos de series de tiempo. Y empezará con la vigilancia del número de intentos de llamada por segundo (Acton, 2022).

3.4. Limitaciones, desafíos y el futuro de la IA aplicada a Ciberseguridad

3.4.1. Limitaciones y desafíos

Para empezar, se detallarán los desafíos que conlleva la utilización de técnicas de IA en Ciberseguridad.

Según esta clasificación, destacan:

1- Calidad y disponibilidad de datos. La eficacia de las técnicas de la IA se apoya en la calidad y cantidad de la información que esté disponible. Si los datos son pobres o insuficientes puede llevar a un rendimiento menos eficiente de los modelos de IA.
2- Ataques adversarios. Son una amenaza seria para los sistemas de IA ya que pueden engañarlos y hacer que generen resultados incorrectos.
3- Dependencia excesiva de la IA. No es positivo depender exageradamente de la IA para la ciberseguridad porque llevaría a la falta de supervisión humana. Y el juicio humano es vital para interpretar los hallazgos que se encuentren, así como la toma de decisiones.
4- Interpretabilidad. Muchos modelos de IA, especialmente los modelos de aprendizaje profundo son muy complejos y difícil de entender el porqué de ciertos resultados (Mosbah & Annowari, 2023).

De la misma manera, en esta otra clasificación se destacan las limitaciones y desafíos de una forma más amplia:

1- Falsos positivos y falsos negativos: a pesar de que la IA puede potenciar la precisión, aún existe la posibilidad de que aparezcan resultados falsos. Los falsos positivos pueden llevar a detener de forma innecesaria una aplicación

legitima, por ejemplo. Además, genera un agotamiento del equipo de seguridad ya que deben comprobar cada alerta. Y finalmente, puede llevar a que se omitan las verdaderas alertas, debido a la alta frecuencia de las falsas. Respecto a los falsos negativos puede generar una falsa confianza en el sistema y que las amenazas reales pasen desapercibidas (CCN, 2023).

2- Los ataques adversarios contra modelos de IA: estos ataques pueden ser de caja blanca (el atacante tiene conocimiento completo del modelo) y de caja negra (el atacante no tiene el conocimiento del modelo). Los ciberdelincuentes tienen acceso a herramientas avanzadas por lo que pueden llegar a ser un verdadero problema (CCN, 2023).
3- Depender de forma excesiva de las soluciones automatizadas: lo que puede llevar a percepción errónea de la seguridad, dificultad a la hora de interpretar decisiones, las soluciones de IA deben estar adaptadas a la evolución de las amenazas, que se produzcan errores en la automatización, que se ignore la valoración humana, así, como el alto coste que supone mantener y actualizar sistemas de IA (CCN, 2023).
4- Privacidad y la ética en la aplicación de la IA: en relación con la recolección de datos, el problema surge que se guarden más de los necesarios llegando a invadir la privacidad de los usuarios; además, frecuentemente, se guardan sin la autorización o conformidad del usuario, lo que lleva a una inquietud ética y legal. Otro problema puede surgir a la hora de guardar la información ya que de no hacerlo de forma debida puede ser un fácil objetivo para los ciberdelincuentes. También, otra cuestión importante está relacionada con la vigilancia, ya que al igual que las soluciones de ciberseguridad basadas en IA vigilan redes y sistemas para identificar amenazas, también pueden utilizarse para observar con finalidad maliciosa el comportamiento de los usuarios (CCN, 2023).

Llegados a este punto, es importante añadir que el Parlamento Europeo aprobó al pasado diciembre del 2023 la Ley de Inteligencia Artificial (IA) para garantizar la seguridad y el cumplimiento de los derechos fundamentales, promoviendo la innovación. En dicha ley, entre otras cosas, se prohíben aplicaciones determinadas aplicaciones de IA que vulneren los derechos de los ciudadanos, tales como por ejemplo sistemas de categorización biométrica que se fundamenten en atributos personales que se consideren sensibles. A parte, permite a las fuerzas del orden en casos excepcionales el uso de sistemas de identificación biométrica (European Parliament, 2024).

Para continuar, no se puede olvidar detallar los puntos débiles que hay en el grupo de directrices, prácticas y procedimientos (ENISA, NIST y MITRE ATTACK) que se

establecen con la finalizad de detectar, evaluar y reducir los peligros vinculados a implementar y utilizar IA. Ya que no se consideran los factores humanos (perfiles psicológicos y comportamientos de los atacantes), además, está la falta de integración de Sistemas de IA explicables (XAI), entre otras cosas. La incorporación de XAI y los componentes humanos en la ciberseguridad son fundamentales para asegurar que las aplicaciones de Modelos de Lenguaje de Gran Escala (LLMs) sean seguras y confiables. Dichos modelos están diseñados para procesar y crear texto de forma parecida a como lo harían los humanos (ejemplos como: *GPT-3* y *GPT-4* de *OpenAI*) (Polemi, Praça, Kioskli, & Bécue, 2024).

Y, por último, se debe incluir la matización de que la IA no es infalible. El código que genera no es seguro y tiene vulnerabilidades de seguridad. Ya sea cuando genera desde cero, como si genera una reparación del código o incluso hace sugerencias. En todos esos casos, se han detectado fallas.

Además, hay investigaciones que afirman haber encontrado errores simples con una solución sencilla, pero, a la vez, también, ha aparecido código malicioso oculto entre líneas benignas (Negri-Ribalta, Geraud-Stewart, Sergeeva, & Lenzini, 2024).

3.4.2. Futuro de la IA aplicada a Ciberseguridad

A continuación, se explorarán las tendencias que existen y que marcarán el futuro de la IA aplicada a la ciberseguridad.

1- Priorización de amenazas de alta precisión. El futuro de la ASM con pruebas de superficie de ataque. Gracias a la IA y los algoritmos de ML, la ASM mejorará de forma que permitirá una detección más rápida a vulnerabilidades y amenazas nuevas. Por ejemplo: La introducción de *Attack Surface Testing* (AST) por parte de *Cortex Xpanse* es un avance que guarda una relación muy cercana con la gestión de la superficie de ataque. Una herramienta como AST proporciona automatización y una respuesta rápida (Heon, 2024).

2- También, como tendencias para el futuro, se destacan, entre otras:

   - IA junto con ML ya que cada vez está creciendo más dentro del sector de ciberseguridad. Pero esto conlleva riesgos, también.

   -La ciberresiliencia: la facultad de una organización para estar preparada, reaccionar y reponerse de incidentes de ciberseguridad.

   -Dar mayor importancia a la protección de la información.

   -Implementar arquitectura de confianza cero: la totalidad de las solicitudes de acceso deberán ser verificadas y autenticadas (Sealpath, 2024).

3- La IA en la informática forense continuará siendo de ayuda ya que se conseguirá una optimización en la eficiencia y la precisión a la hora de estudiar la información, así como de identificar amenazas (EclipseForensics, 2023).
4- El aprendizaje federado: tiene un gran potencial para ser explotado en el futuro. Se basa en la idea de entrenar un modelo de forma local en la fuente de datos. Es decir, entrena un modelo de aprendizaje utilizando la información que tiene localmente. Después, cada dispositivo comparte los resultados con un servidor central. Este servidor será quien combine los resultados de todos los dispositivos en un modelo global. Finalmente, este modelo global actualizado se envía de vuelta a todos los dispositivos para que siga entrenado o use el modelo mejorado. Es un mecanismo diseñado para proteger la privacidad y mejorar la seguridad de la información (Driss, Sabir, Elbiaze, & Saad, 2023).
5- XAI: quiere decir Inteligencia Artificial Explicable. Son técnicas que consiguen ayudar a los humanos porque permiten que los resultados de los algoritmos se puedan entender mejor. Y, por consiguiente, aporta beneficios tales como elevar la confianza del usuario ya que puede entender que hay detrás de algunas decisiones de la IA; además, ayuda a que se cumplan las regulaciones lo que garantiza que los modelos de IA sean equitativos y confiables. Y, por último, mejora de forma significativa la eficacia y la precisión del modelo (Moja, 2024).
6- Adopción de *Blockchain* para seguridad: ya que puede se puede utilizar para comprobar y almacenar de forma segura las identidades; las auditorias son más sencillas ya que tiene un registro de todas las transacciones; es complicado de cerra porque una red *blockchain* está repartida entre muchos nodos; al estar distribuido de esa forma no depende de un registro central, por lo que es difícil de *hackear*; además, almacena los datos de forma que no se pueden alterar; finalmente, garantiza la comunicación segura con autenticación entre los dispositivos IoT y tienen una automatización segura (CCN, 2023).

Pero, cabe destacar que a pesar de dichas técnicas son reconocidas por su potencial para identificar y reducir sesgos; vigilar los ataques adversariales y garantizar la integridad de los sistemas IA en su ciclo de vida, aún no están introducidas en los procesos estándar de evaluación de riesgos y detección de ataques adversarios. Por lo que, no se están aprovechando las capacidades de XAI de forma completa (Polemi, Praça, Kioskli, & Bécue, 2024).

7- IA cuántica: esto supondrá una revolución en lo referente al cifrado y la seguridad de la información. Las computadoras cuánticas funcionan con *qubits* lo que hace que se procesen los datos a velocidades extremadamente altas, lo que puede presentar un gran desafío. También, para enfrentar desafíos se están desarrollando algoritmos resistentes a la computación cuántica (algoritmos cuántico-resistentes) (Smalakys, 2024).
8- IA en el borde (*Edge AI*): permite que los dispositivos periféricos examinen y contesten a la información de forma inmediata, sin que haya que enviar los datos a un servidor central para su procesamiento; disminuye el tiempo en que un dato va desde el origen al destino (latencia), lo que origina más velocidad en las respuestas; hay una mejor eficiencia en el ancho de banda ya que procesa los datos de forma local; además, hay un aumento en la seguridad y la privacidad de la información porque no es necesario transmitir datos sensibles a través de la red (Wevolver, 2024).

4. Conclusiones

Llegados a este punto, se va a desarrollar una conclusión sobre el trabajo realizado, así como las áreas para futura investigación.

4.1. Conclusión

Con la presente investigación se ha demostrado la estrecha relación que hay entre la IA y la ciberseguridad. El diseño de mi trabajo cumple con los objetivos que me planteé. Buscaba realizar una exposición clara pero profunda sobre ambos campos, así como, su conexión. Por lo que, inicialmente, estoy satisfecha con el resultado final. Pero, es verdad, que se quedan muchos temas sin un amplio desarrollo.

En el trabajo se ha detallado la gran importancia de utilizar IA como un arma para fortalecer la ciberseguridad. Pero de la misma manera, se han explicado todos los desafíos que ello conlleva. A parte, de los beneficios que aporta el avance de la tecnología, no se pueden olvidar los peligros que trae de la mano. Ya que, no solo son aprovechados para luchar contra los ciberdelincuentes, sino que ellos están aprendiendo, cada vez más, la forma de realizar sus ataques de forma más peligrosa, difícil de detectar y sofisticada, lo que hace que se entre en un círculo vicioso en el que la defense avanza a la par que el ataque.

También, merece destacar las cuestiones éticas y legales. Cada vez hay una sensibilidad mayor sobre la protección de la privacidad de los usuarios. A la vez, existe un temor legítimo, y está en crecimiento, sobre el abuso de la IA en ciberseguridad. Es decir, un uso indebido que lleve a una vigilancia desmedida, invadiendo la privacidad de los usuarios.

Por otro lado, no hay que olvidar, como ya hemos dicho anteriormente en el análisis de esta investigación, que tiene que haber un equilibrio entre la IA y la automatización con la experiencia humana. De esta manera y siguiendo la normativa legal al respecto se consigue la mejor defensa contra la evolución de cualquier incidente.

Por lo que, para finalizar, yo opino que la mejor forma de estar preparados para el futuro es tener y fomentar una continua educación a todos los niveles.

4.2. Áreas para futura investigación

La IA es una herramienta muy valiosa en ciberseguridad y está creciendo exponencialmente.

Esta investigación deja al descubierto un amplio abanico de argumentos que podrían abrir nuevas investigaciones o seguir ampliando las actuales.

Áreas como la IA cuántica se presentan prometedoras.

E inclusive, algo más básico, pero no menos importante, como, por ejemplo, desarrollar una investigación de como introducir y fomentar la educación en IA y ciberseguridad en todos los niveles. Los avances de estos dos campos son fundamentales para proteger nuestros sistemas y nuestros datos. Aunque, carecen de verdadero impacto si no son capaces de transmitirse de forma efectiva a través de la educación.

5. Glosario

-Algoritmo: Conjunto de instrucciones paso a paso que sigue una computadora para hacer una tarea o solucionar un problema.

-Algoritmos basados en instancias: Modelos que almacenan y hacen uso de instancias de entrenamiento para hacer predicciones, como k-NN.

-Algoritmos de agrupamiento: Métodos que agrupan datos en función de similitudes, como *k-means*.

-Algoritmos de regresión: Modelos que predicen un valor continúo usando datos de entrada.

-Algoritmos de regularización: Técnicas que evitan el sobreajuste penalizando la complejidad del modelo.

-Amenaza cibernética: Es cualquier posible amenaza que pueda comprometer la seguridad de sistemas y datos informáticos.

-Amenazas de alta precisión: Ciberataques que apuntan a objetivos particulares con gran exactitud.

-Antivirus: *Software* diseñado para detectar, prevenir y eliminar virus y otros tipos de *malware* de los ordenadores.

-Aprendizaje Automático: es un subcampo de la inteligencia artificial (IA) que permite a las computadoras aprender y mejorar automáticamente en función de la experiencia sin estar programadas explícitamente para cada tarea.

-Aprendizaje automático adversario: Técnica utilizada para entrenar modelos de IA para resistir ataques diseñados para engañar al modelo.

-Aprendizaje no supervisado: Un tipo de aprendizaje automático que detecta patrones a partir de datos sin etiquetar.

-Aprendizaje profundo: Una rama del aprendizaje automático que utiliza redes neuronales profundas para modelar datos complejos.

-Aprendizaje supervisado: Tipo de aprendizaje automático en donde se entrena el modelo usando un conjunto de datos etiquetados.

-Ataque cibernético: Cualquier ataque que pueda poner en peligro la seguridad de sistemas informáticos y de datos.

- Ataques basados en *Host (Host-Based DoS)*: Ataques solo dirigidos a un sistema o servidor.

- Ataques basados en Red *(Network-Based DoS)***:** Se basan en la infraestructura de red.

-Ataques de denegación de servicio *DDoS: Se refieren a un tipo de a*taques que sobrecargan un servidor o red con tráfico excesivo, haciéndolos inaccesibles para los usuarios legítimos.

- Ataques Distribuidos (*DDoS)*: Incluyen a varios sistemas comprometidos

-Automatización y Orquestación: Utilizar un software para automatizar tareas y coordinar procesos complejos en sistemas informáticos.

-*Bot*: Robot según la dirección que se le dé puede realizar diversas funciones y no necesita la intervención de una persona.

-*BotnetMaster*: El robot que controla la red.

-*Botnes*: Redes zombis.

-*Blockchain para seguridad:* Uso de la tecnología *blockchain* para proteger datos y transacciones a través de un registro descentralizado y seguro.

-Capa de aplicación: Nivel del modelo OSI, que interactúa directamente con el software de aplicación y proporciona servicios de red.

-Centro Criptológico Nacional: Organización española que garantiza la seguridad de las tecnologías de la información en el sector público.

-Ciber criminología: Campo que examina lo referente a los delitos del ciberespacio y el comportamiento delictivo en línea.

-Ciberataque: Se refiere a un intento malicioso de acceder, dañar, robar o cambiar información en sistemas informáticos, redes o dispositivos digitales.

-Cibercrimen: Referente a ciberdelincuencia.

-Cibercriminalidad: Comportamiento delictivo en el área de informática.

-Ciberdelincuencia: Actividades delictivas que se realizan por medios informáticos y redes digitales.

-Ciberespacio: Espacio virtual donde se llevan a cabo las comunicaciones y actividades en línea a través de redes informáticas e internet.

-Ciberseguridad: Sinónimo de seguridad informática.

-Cisco: Empresa del sector tecnológico que fabrica y vende equipos de redes y telecomunicaciones.

-Computación Cuántica: Realización de cálculos a velocidades mayores que ordenadores tradicionales mediante principios la mecánica cuántica.

-C*ryptojacking*: Uso autorizado del ordenador de una persona para minar criptomonedas.

-*Darktrace:* Empresa de ciberseguridad que hace uso de IA con el fin de detectar y responder a amenazas en tiempo real.

-Dispositivos IoT: Objetos conectados a internet que pueden recopilar y transmitir datos.

-*Drive-day attack:* Ataque en el que el usuario queda infectado por visitar una página web comprometida.

-El *Kit* de Desarrollo de Software: Un conjunto de herramientas y bibliotecas utilizadas por desarrolladores para crear aplicaciones.

-Entropía: En teoría de la información, mide la incertidumbre o imprevisibilidad de la información contenida en un mensaje o conjunto de datos.

-*Firewall: Software* para controlar el tráfico y proteger un sistema contra accesos no autorizados.

-GANs (*Generative Adversarial Networks*): Redes neuronales que generan datos nuevos, compitiendo entre los modelos generador y discriminador.

-Gusanos: Tipo de *malware* autorreplicable, se propaga automáticamente a través de redes informáticas sin la intervención humana.

-Herramientas de análisis forense: Software y técnicas utilizadas para investigar incidentes referentes a seguridad cibernética.

-*Host:* Dispositivo conectado a una red que puede enviar o recibir datos.

-*IA generative:* Inteligencia artificial que crea nuevo contenido a partir de patrones aprendidos en datos existentes.

-ICMP: Protocolo de red utilizado para enviar mensajes de error y operacionales.

-Indicaciones de Nombre del Servidor: Extensión del protocolo TLS que permite conocer el nombre del servidor al que se está conectando.

-Instituto Nacional de Estándares y Tecnología (NIST): Agencia del gobierno de Estados Unidos dedicada a la seguridad de los sistemas tecnológicos.

-Inteligencia Artificial: Simulación de procesos de IA por sistemas informáticos.

-Inteligencia Artificial Explicable: IA diseñada para que sus decisiones y procesos sean transparentes y comprensibles para los humanos.

-Inteligencia de Seguridad de *Microsoft:* Servicio de Microsoft para proteger los datos, sistemas y redes de amenazas cibernéticas.

-Inteligencia de Seguridad de *Microsoft:* Servicio de Microsoft para proteger los datos, sistemas y redes de amenazas cibernéticas.

-Inundaciones SYN: Tipo de ataque que envía solicitudes de conexión a un servidor con la finalidad de sobrecargarlo y hacerlo inaccesible.

-La ciencia forense digital y respuesta a incidentes: Métodos para investigar y mitigar ciberataques mediante el análisis de evidencia digital.

-La inyección SQL: Ataque para manipular bases de datos mediante la inserción de un código SQL malicioso.

-Los Algoritmos Bayesianos: Modelos que utilizan el Teorema de Bayes para hacer predicciones probabilísticas.

-Los algoritmos de árboles de decisión: Modelos que dividen los datos en ramas basadas en características para tomar decisiones.

-*Malware*: Es un software malicioso.

-*Man-in-the-Middle attack:* Ataque donde el atacante intercepta y posiblemente altera la comunicación entre dos partes sin que ellas lo sepan.

-Métodos de aprendizaje de reglas de asociación: Técnicas que identifican relaciones significativas entre variables en grandes bases de datos.

-Paquetes por segundo (PPS): Cantidad de paquetes de datos transmitidos en un segundo.

-*Phishing*: Se trata de una técnica de fraude en línea donde los atacantes hacen el intento de engañar a las personas para que revelen información personal y sensible.

-*Plugins:* Programas adicionales para ampliar la funcionalidad de un software.

-Procesamiento del Lenguaje Natural: Área de la IA que permite a las máquinas entender y procesar el lenguaje humano.

-Pruebas de penetración: Evaluación de seguridad que simula ataques cibernéticos para identificar vulnerabilidades en un sistema.

-*Ransomware*: Tipo de *malware*, que cifra los archivos de una víctima, impidiendo el acceso a su información hasta que se pague un rescate.

-Redes 5G: Quinta generación de tecnología en redes móviles tiene velocidades más rápidas y mayor capacidad.

-Redes neuronales: Modelos de computación basados en el cerebro humano, aprenden a través de capas de nodos interconectados.

-Revisión Sistemática de Literatura: Método de investigación que recopila, evalúa y sintetiza toda la evidencia relevante sobre una pregunta de investigación específica de manera sistemática y estructurada.

-Seguridad Informática*:* o ciberseguridad, conjunto de medidas y prácticas con la finalidad de proteger los sistemas informáticos, las redes y los datos, en contra del acceso no autorizado, los ataques cibernéticos, los daños o la destrucción.

-Semi-Supervisado: Tipo de aprendizaje automático, utiliza una combinación de datos etiquetados y no etiquetados para entrenar modelos.

-Sistemas de respuesta automatizada: Tecnologías que detectan y responden automáticamente a incidentes de seguridad cibernética.

-Sistemas Informáticos: Conjunto de componentes interrelacionados, que trabajan juntos para procesar, almacenar, y transmitir información.

-*Smurf DDoS*: Ataque medio de mensajes ICMP para sobrecargar una red al amplificar el tráfico enviado.

-*Spam*: Correo electrónico no deseado.

-*Spyware: Malware que* monitorea y registra información del usuario sin su permiso para luego enviarla sin autorización a un tercero.

-Talos Cisco: Grupo de inteligencia de seguridad de Cisco encargado de la investigación y combate de amenazas cibernéticas.

-Tecnología *blockchain:* Sistema de registro descentralizado y distribuido que se utiliza para almacenar datos de forma segura y transparente. Funciona a través de una cadena

de bloques (*blocks*) vinculados, donde cada bloque contiene una colección de transacciones o datos.

-Teorema de Bayes: Fórmula matemática que describe la probabilidad de un evento basado en conocimientos previos.

-Teoría de la Computabilidad: Es la rama de la informática encargada de estudiar los problemas que pueden ser resueltos por una computadora y una solución eficiente.

-Teoría de la Información: Estudio matemático de la transmisión, procesamiento, separación y almacenamiento de información.

-Teoría de la Recursión: Una rama de la teoría de la computabilidad, centrada en funciones y procesos que se definen en términos a sí mismos.

-Teoría del Empujón (*Nudge Theory*): Concepto de las ciencias del comportamiento, economía y política que recomienda formas de influir en las decisiones y comportamiento sutilmente, sin prohibir ninguna opción ni cambiar significativamente los incentivos económicos de una persona.

-Teoría General de Sistemas: Enfoque interdisciplinario para el estudio de sistemas en general, intentando encontrar principios que puedan aplicarse a todos los tipos de sistemas en diferentes campos.

-Teoría Matemática de la Comunicación: Rama de la teoría de la información, enfocada en cómo se transmite la información mediante canales de comunicación su codificación para minimizar errores y maximizar su eficiencia.

-Troyano: Programa que se disfraza para conseguir una apariencia legítima con funciones maliciosas.

-Tunel DNS: Técnica que encapsula infromación en consultas y respuestas DNS para evitar la detección.

- UDP: Protocolo de comunicación que permite enviar datos sin conexión previa, rápido, pero no fiable.

- Virus: Son programas maliciosos que reproducen y propagan de un sistema a otro sin control

6. Lista de abreviaturas

-AE: AutoCodificador (*AutoEncoder)*

-AI HLEG: El Grupo de Expertos de Alto Nivel en Inteligencia Artificial

-AIDS: IDS con el método de detección basado en anomalías

-AML: Adversarial Machine Learning

-AODE: Estimadores Promediados de Una Dependencia (*Averaged One-Dependence Estimators)*

-ASM: Amenazas de alta precisión

-AST: *Attack Surface Testing*

-AV-TEST: AV-TEST GmbH

-BBN: Red de Creencias Bayesianas (*Bayesian Belief Network)*

-BN: Red Bayesiana (*Bayesian Network)*

-BPS: Bits por Segundo

-CART: Árbol de Clasificación y Regresión (*Classification and Regression Tree)*

-*CE:* Comisión Europea

-CHAID: Detección Automática de Interacciones Chi-cuadrado (*Chi-squared Automatic Interaction Detection)*

-CIA: Principios de Confidencialidad, Integridad y Disponibilidad

-CNN: Redes Neuronales Convolucionales (*Convolutional Neural Networks)*

-CNN*: Centro Criptológico Nacional*

-DDoS: Ataque denegación de servicio

-DFIR: La ciencia forense digital y respuesta a incidentes (*Digital forensics and incident response)*

-*DL: Aprendizaje Profundo* (Deep Learning)

-DNN: Red neuronal profunda

-DNS: Sistema de Nombres de Dominio (Domain Name System)

-DQN: *Deep Q-Network*

-EM: Maximización de la Expectativa (*Expectation Maximisation)*

-ENISA: Agencia de la Unión Europea para la Ciberseguridad

-EY: EY Global

-GANs: Redes Generativas Adversarias

-HIDS: IDS con el método de implementación basado en Host

-IA: Inteligencia Artificial

-ICMP: Protocolo de Mensajes de Control de Internet (*Internet Control Message Protocol*)

-IAT: Tiempo entre llegadas (*Inter-Arrival Time)*

-ID3: Dicodificador Iterativo 3 (*Iterative Dichotomiser 3)*

-IDPS: Sistema de detección y prevención de intrusiones

-IMS: Subsistema Multimedia IP (IP *Multimedia Subsystem)*

-IoT: Internet de las cosas

-kNN: k-Vecinos Más Cercanos (*Nearest Neighbor)*

-LARS: Regresión por Ángulos Mínimos (*Least-Angle Regression)*

-LASSO: Operador de Contracción y Selección Absoluta Mínima (*Least Absolute Shrinkage and Selection Operator)*

-LLM: Modelos de lenguaje de gran tamaño

-LOESS: Suavizado de Dispersión Estimado Localmente (*Locally Estimated Scatterplot Smoothing)*

-LSTM: Redes Neuronales de Memoria a Largo Corto Plazo

-LVQ: Cuantificación de Vectores de Aprendizaje (*Learning Vector Quantization)*

-LWL: Aprendizaje Ponderado Localmente (*Locally Weighted Learning*

-MAL: algoritmos de ML

-MARS: Splines de Regresión Adaptativa Multivariante (*Multivariate Adaptive Regression Splines)*

-MITM: Ataque de intermediario (*Man-in-the-Middle attack)*

-ML: *Aprendizaje automático* (Machine Learning)

-MLP: Perceptrones Multicapa (*Multilayer Perceptrons)*

-MTTD: Tiempo medio de detección

-MTTR: Tiempo medio de respuesta

-NIDS: IDS con el método de implementación basado en Red

-NIST: Instituto Nacional de Estándares y Tecnología (*National Institute of Standards and Technology)*

-OLSR: Regresión por Mínimos Cuadrados Ordinarios (*Ordinary Least Squares Regression)*

-PNL: Procesamiento del Lenguaje Natural (*Natural Language Processing)*

-PPS: Paquetes por Segundo

-RBFN: Red de Funciones de Base Radial (*Radial Basis Function Network)*

-RNNs: Redes Neuronales Recurrentes

-RNA: Redes neuronales artificiales

-SDK: el *Kit* de Desarrollo de Software (*Software Development Kit)*

-SD: La detección de firma (*Signature Detection)*

-SIDS: IDS con el método de detección basado en firmas

-SINs: Indicaciones de Nombre del Servidor (*Server Name Indications)*

-SLR: Revisión Sistemática de Literatura

-SOC: Centro de Operaciones de Seguridad

-SOM: Mapa Autoorganizado (*Self-Organizing Map*

-SOAR: La respuesta automática y orquestación (*Security Orchestration, Automation and Response)*

-SQLI: Inyección SQL

-SVM: Máquinas de Vectores de Soporte (*Support Vector Machines)*

-SYN: Ataque de inundación SYN (SYN Flood Attack)

-TGS: Teoría General de Sistemas

-UDP: Protocolo de Datagrama de Usuario (User Datagram Protocol)

-WMI: Instrumental de Administración de Windows *(Windows Management Instrumentation)*

Bibliografía

Aakanksha, T., Ankit, k., & Dharma, A. (2017). Fighting against phishing attacks: state of the art and future challenges. *28*. Neural Computing and Applications. Obtenido de https://www.researchgate.net/publication/298908229_Fighting_against_phishing_attacks_state_of_the_art_and_future_challenges

Acton, N. (2022). AI in cybersecurity an introduction and case studies. Snorkel. Obtenido de https://snorkel.ai/ai-in-cybersecurity/

Ahmad, Z., Shahid, A., Wai, C., Abdullah, J., & Ahmad, F. (2020). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Wiley. Obtenido de https://onlinelibrary.wiley.com/doi/pdf/10.1002/ett.4150

Ahsan, M., Nygard, K., Gomes, R., Chowdhury, M., Rifat, N., & Connolly, J. (2022). Cybersecurity Threats and Their Mitigation Approaches Using Machine Learning. 527-555. Journal of Cybersecurity and Privacy. Obtenido de https://doi.org/10.3390/jcp2030027

Alanda, A. (2020). Mobile Application Security Penetration Testing. IOP Conference Series: Materials Science and Engineering.

Alandro, E. (2011). La Teoría de la Información ante las nuevas tecnologías de la comunicación. (U. C. Madrid, Ed.) España : CIC. Cuadernos de Información y Comunicación. Obtenido de https://www.redalyc.org/pdf/935/93521629005.pdf

Alenezi, M., Alabdulrazzaq, H., Alshaher, A., & Alkharang, M. (diciembre de 2020). Evolution of Malware Threats and Techniques: A Review. *12(3)*. International Journal of Communication Networks and Information Security (IJCNIS).

Ansari, M. F., Dash, B., Sharma, P., & Yathiraju, N. (2023). The Impact and Limitations of Artificial Intelligence in Cybersecurity: A Literature Review. The Impact and Limitations of Artificial Intelligence in Cybersecurity: A Literature Review.

Aqib, A., Samreen, N., Sania, A., & Munawar, A. (junio de 2023). Shannon Entropy in Artificial Intelligence and Its Applications Based on Information Theory. *13*, 9-17. Journal of Applied and Emerging Sciences.

Ashwani, K. (2023). What is CyberSponse and use cases of CyberSponse? CyberSponse.

AV ATLAS. (2023). Total amount of malware and pua. The Independent IT-Security Institute. Obtenido de https://portal.av-atlas.org/malware

AVTEST. (febrero de 2024). 2023 Cyber-Incidents in Numbers. The Independent TI Security Institute. Obtenido de https://www.av-test.org/fileadmin/pdf/security_report/AV-TEST_Cyber_Incidents_Report_2023_en.pdf

AVTEST. (2024). Cyber-Incidents in Numbers: Year 2023. The Independent IT-Security Institute. Obtenido de https://www.av-test.org/en/news/cyber-incidents-in-numbers-year-2023

Ayerbe, A. (10 de noviembre de 2020). La ciberseguridad y su relación con la inteligencia. España: Real Instituto El Cano. Obtenido de https://media.realinstitutoelcano.org/wp-content/uploads/2021/10/ari128-2020-ayerbe-ciberseguridad-y-su-relacion-con-inteligencia-artificial.pdf

Bartels, J. (2024). Harnessing Predictive Analytics In Cybersecurity. BIIA. Obtenido de https://www.biia.com/harnessing-predictive-analytics-in-cybersecurity/

Bilge, L., & Dumitraş, T. (2012). Before we knew it: an empirical study of zero-day attacks in the real world. 833–844. ACM conference on Computer and communications security. Obtenido de https://doi.org/10.1145/2382196.2382284

Bitdefender. (2017). Stop Fileless Attacks at Pre-execution.

Brownlee, J. (2023). A Tour of Machine Learning Algorithms. *Machine Learning Algorithms*. Obtenido de https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

CCN. (octubre de 2023). Approach to Artificial Intelligence and Cybersecurity. Centro Criptológico Nacional. Obtenido de https://www.ccn-cert.cni.es/es/informes/informes-de-buenas-practicas-bp/7192-ccn-cert-bp-30-approach-to-artificial-intelligence-and-cybersecurity/file.html

Chandramouli, R., Raj, R., & Bhagyaveni, M. (2020). A comprehensive review on cybersecurity. *7*, 365-372. Journal of Critical Reviews.

CISA. (2021). Understanding Denial-of-Service Attacks. American´s Cyber Defense Agency. Obtenido de https://www.cisa.gov/news-events/news/understanding-denial-service-attacks

CISCO. (2023). Revisión del año. Obtenido de https://www.cisco.com/c/dam/global/es_mx/products/pdfs/talos-2023-reporte.pdf

CISCO. (2024). Underprepared and Overconfident Companies Tackle an Evolving Landscape. Cisco Cybersecurity Readiness Index. Obtenido de https://newsroom.cisco.com/c/dam/r/newsroom/en/us/interactive/cybersecurity-readiness-index/documents/Cisco_Cybersecurity_Readiness_Index_FINAL.pdf

Cisco Systems. (2021). Protección contra ransomware. Obtenido de https://www.cisco.com/c/dam/global/es_mx/solutions/pdf/protecting-against-ransomware-spa.pdf

CISCO TALOS. (2023). Revisión del año 2022 Panorama general de amenazas. Obtenido de https://www.cisco.com/c/dam/global/es_mx/products/pdfs/onepager-threat.pdf

CISCO TALOS. (2023). Year in review. Obtenido de https://blog.talosintelligence.com/content/files/2023/12/2023_Talos_Year_In_Review.pdf

Clarke, J. (2012). SQL Injection Attacks and Defense. *2*. USA: Elsevier.

Conti, M., Dragoni, N., & Lesyk, V. (2016). A Survey of Man In The Middle Attacks. *18(3)*, 2027-2051. IEEE Communications Surveys & Tutorials. Obtenido de https://ieeexplore.ieee.org/abstract/document/7442758/authors#authors

Cortex XSOAR. (2024). Overview.

Daniel, G., & Arce, M. (2018). Malware and Market Share. The University Texas of Dallas. Obtenido de https://utd-ir.tdl.org/server/api/core/bitstreams/abe77742-3c77-47b9-97e8-cf080ce14007/content

DARKTRACE. (2023). Introducing the Darktrace ActiveAI Security Platform. Obtenido de https://darktrace.com/

Dasgupta, D., Akhtar, Z., & Sen, S. (2020). Machine learning in cybersecurity: a comprehensive survey. *19*. Journal of Defense Modeling and Simulation (JDMS). Obtenido de https://doi.org/10.1177/15485129209512

Davies, V. (enero de 2022). Company profile: Who are Siemplify? Cyber Megazine.

Deloitte Insights. (diciembre de 2021). Cyber AI: Real defense. *Augmenting security teams with data and machine intelligence*. Obtenido de https://www2.deloitte.com/us/en/insights/focus/tech-trends/2022/future-of-cybersecurity-and-ai.html

Driss, M. B., Sabir, E., Elbiaze, H., & Saad, W. (2023). Federated Learning for 6G: Paradigms, Taxonomy, Recent Advances and Insights. University of Quebec at Montreal (UQAM).

EclipseForensics. (febrero de 2023). How Will AI Transform Digital Forensics in 2023 and Beyond? Obtenido de https://eclipseforensics.com/how-will-ai-transform-digital-forensics-in-2023-and-beyond/

Ernst & Young Global. (24 de abril de 2023). La ciberdelincuencia sigue en aumento: los ciberataques se multiplican. España. Obtenido de https://www.ey.com/es_es/cybersecurity/la-ciberdelicuencia-sigue-aumento-los-ciberataques-se-multiplican

Erquiaga, M. (2011). Botnets: Mecanismos de Control y de propagación. XVII Congreso Argentino de Ciencias de la Computación. Obtenido de https://sedici.unlp.edu.ar/bitstream/handle/10915/18764/Documento_completo.pdf?sequence=1&isAllowed=y

Esentire. (2024). 2023 Official Cybercrime Report. Obtenido de https://www.esentire.com/resources/library/2023-official-cybercrime-report

European Parliament. (2024). Artificial Intelligence Act: MEPs adopt landmark law. Press Releases.

Fernández, D., & Martínez, G. (2018). Ciberseguridad, Ciberespacio y Ciberdelincuencia. Universidad a Distancia de Madrid . Obtenido de https://udimundus.udima.es/handle/20.500.12226/84

Fruhlinger, J. (mayo de 2024). DDoS attacks: Definition, examples, and techniques. CSO.

Gallardo, D., Lesta, I., & Arques, P. (2003). Introducción a la teoría de la computabilidad. España: Universidad de Alicante. Obtenido de https://observatorio-cientifico.ua.es/documentos/5f0500ba2999524666430775

Girton. (2024). DarkTrace. Girton Grammar School.

Gonzales, S., Dormido, S., & Sánchez, J. (2019). Un modelo de aproximación sistémica como herramienta de investigación y solución ante la ciberseguridad en sistemas de automatización industrial. *23*, 16-25. Revista Internacional de Sistemas. Obtenido de https://ojs.uv.es/index.php/ris/article/view/14105/14656https://ojs.uv.es/index.php/ris/article/view/14105/14656

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT. Obtenido de https://www.deeplearningbook.org/

Guaña, J., Sánchez, A., Chérrez, P., Chulde, L., Jaramillo, P., & Pillajo, C. (2022). Ataques informáticos más comunes en el mundo digitalizado. 87-100. Revista Ibérica de Sistemas e Tecnologias de Informação. Obtenido de https://dspace.itsjapon.edu.ec/jspui/bitstream/123456789/3445/1/ATAQUES%20INFORMATICOS.pdf

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. *11*. IEEE.

Heon, G. (2024). Confirm Attack Surface Vulnerabilities with Cortex Xpanse Attack Surface Testing. Palo Alto Networks.

High-Level Expert Group on Artificial Intelligence. (2019). A definition of AI: Main capabilities and scientific disciplines. European Commission.

Hung-Jen, L., Chun-Hung, R., Ying-Chih, L., & Kuang-Yuan, T. (2013). Intrusion detection system: A comprehensive review. *36*, 16-24. Journal of Network and Computer Applications.

IBM. (2024). What is digital forensics and incident response (DFIR)? Obtenido de https://www.ibm.com/topics/dfir

Inayat, U., Zia, M. F., Mahmood, S., Khalid, H. M., & Benbouzid, M. (2022). Learning-Based Methods for Cyber Attacks Detection in IoT Systems: A Survey on Methods, Analysis, and Future Prospects. Electronics. Obtenido de https://doi.org/10.3390/electronics11091502

INCIBE. (7 de julio de 2023). Machine learning, el fórmula 1 inteligente. Obtenido de https://www.incibe.es/empresas/blog/machine-learning-el-formula-1-inteligente

Javaheri, D., Hosseinzadeh, M., & Rahmani, A. (2018). Detection and Elimination of Spyware and Ransomware by Intercepting Kernel-Level System Routines. *6*, 78321-78332. IEEE Access. Obtenido de https://ieeexplore.ieee.org/abstract/document/8566151

kaspersky. (2023). ¿Qué es el cibercrimen? Cómo protegerse del cibercrimen. Obtenido de https://latam.kaspersky.com/resource-center/threats/what-is-cybercrime

Kaur, R., Gabrijelčič, D., & Klobučar, T. (abril de 2023). Artificial Intelligence for Cybersecurity: Literature Review and Future Research Directions. Elsevier. Obtenido de 10.1016/j.inffus.2023.101804

Kitchenham, B., & Charters, S. (enero de 2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. EBSE Technical Report.

Le, N., Rathour, V., & Yamazaki, K. (2021). Deep reinforcement learning in computer vision: a comprehensive survey. Artif Intell Rev. Obtenido de https://doi.org/10.1007/s10462-021-10061-9

Lim, A. (agosto de 2023). An Executive View of Key Cybersecurity Trends and Challenges in 2023. ISACA. Obtenido de https://www.isaca.org/resources/news-and-trends/industry-news/2023/an-executive-view-of-key-cybersecurity-trends-and-challenges-in-2023

Liu, H., & Lang, B. (2019). Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey. China: Beihang University.

Liu, X., Fayaz, S., Khalid, M., Ke, J., Irshad, M., Ul-Haq, J., & Abbas, S. (2022). Cyber security threats: A never-ending challenge for e-commerce. Frontiers in psychology.

Maad, M., Omega, J., Youssef, F., Indu, B., & Humam, A.-S. (2023). Exploring the Top Five Evolving Threats in Cybersecurity: An In-Depth Overview. 57-53. Mesopotamian journal of Cybersecurity. Obtenido de https://mesopotamian.press/journals/index.php/CyberSecurity/article/view/44/80

Mahesh, B. (2019). Machine Learning Algorithms -A Review. International Journal of Science and Research. Obtenido de https://www.researchgate.net/publication/344717762_Machine_Learning_Algorithms_-A_Review

Markevych, M., & Dawson, M. (2023). A Review of Enhancing Intrusion Detection Systems for Cybersecurity Using Artificial Intelligence (AI). *29*, 30-37. Sciendo. Obtenido de https://doi.org/10.2478/kbo-2023-0072

Ministerio del Interior. (2022). España. Obtenido de https://www.interior.gob.es/opencms/es/detalle/articulo/Espana-registro-374.737-ciberdelitos-en-2022/#:~:text=El%20n%C3%BAmero%20de%20detenidos%20e,con%20respecto%20al%20a%C3%B1o%202021.

MIT. (1956). Artificial General Intelligence Robots AI Agi Deepmind Google Openai. MIT Technology Review: The al summer.

MIT. (2020). Artificial general intelligence: Are we close, and does it even make sense to try? MIT Technology Review.

Moja, M. (junio de 2024). Inteligencia Artificial Explicable: El Futuro de la Transparencia en la IA. LANUEVAIA.

Morgan, S. (octubre de 2023). Cybercrime To Cost The World $9.5 Trillion USD Annually In 2024. Cybercrime Magazine. Obtenido de https://cybersecurityventures.com/cybercrime-to-cost-the-world-9-trillion-annually-in-2024/

Mosbah, A., & Annowari, N. B. (2023). Artificial Intelligence in Cybersecurity: Opportunities and Challenges. 7, 789-794. International Journal of Business Society.

Mostofa, A. (noviembre de 2021). Increasing the Predictive Potential of Machine Learning Models for Enhancing Cybersecurity. North Dakota State University. Obtenido de https://www.proquest.com/docview/2540461101?pq-origsite=gscholar&fromopenview=true&sourcetype=Dissertations%20&%20Theses

Muñoz, M. (05 de agosto de 2023). Qué es la 'teoría del empujón' y por qué puede ayudarte a evitar ciberriesgos. España. Obtenido de https://www.20minutos.es/tecnologia/ciberseguridad/que-es-teoria-empujon-por-que-puede-ayudarte-evitar-ciberriesgos-5161517/

National Security Agency USA. (23 de julio de 2021). Artificial Intelligence: Next Frontier is Cybersecurity. United States of America. Obtenido de https://www.nsa.gov/Press-Room/News-Highlights/Article/Article/2702241/artificial-intelligence-next-frontier-is-cybersecurity/

Negri-Ribalta, C., Geraud-Stewart, R., Sergeeva, A., & Lenzini, G. (mayo de 2024). A systematic literature review on the impact of AI models on the security of code generation. Sec. Cybersecurity and Privacy.

NIST. (2023). Artificial Intelligence: Adversarial Machine Learning. National Cybersecurity Center of Excellence. Obtenido de https://www.nccoe.nist.gov/ai/adversarial-machine-learning

NIST. (febrero de 2024). The NIST Cybersecurity Framework (CSF) 2.0. National Institute of Standards and Technology. Obtenido de https://doi.org/10.6028/NIST.CSWP.29

NIST AL. (2024). NIST Identifies Types of Cyberattacks That Manipulate Behavior of AI Systems. National Institute of Standards and Technology.

Palmer, D. (7 de febrero de 2019). Trojan malware: The hidden cyber threat to your PC. ZDNET. Obtenido de https://www.zdnet.com/article/trojan-malware-the-hidden-cyber-threat-to-your-pc/#google_vignette

Palo Alto Networks. (2021). AutoFocus Administrator’s Guide.

Peña, D. (2023). Ciberdelitos y Criminalidad Informática. *Rol de la prevención en la expansión de la ciberdelincuencia, 13*, 57-72. Revista Iberoamericana de Derecho Informático. Obtenido de file:///C:/Users/Adriana/Downloads/CIBERDELITOS+Y+CRIMINALIDAD+INFORM%C3%81TICA%20(3).pdf

Periman, K. (2017). Ransomware Lessons for the Financial Services Industry. CISCO. Obtenido de https://blogs.cisco.com/financialservices/ransomware-lessons-for-the-financial-services-industry

Polemi, N., Praça, I., Kioskli, K., & Bécue, A. (2024). Challenges and efforts in managing AI trustworthiness risks: a state of knowledge. *7*. Sec. Cybersecurity and Privacy. Obtenido de https://www.frontiersin.org/articles/10.3389/fdata.2024.1381163/full

Radware´s. (2024). Global Threat Analysis Report. Obtenido de https://es.radware.com/cyberpedia/ddospedia/ddos-meaning-what-is-ddos-attack/

RecordedFuture. (2024). Strengthen Your Defenses with Threat Intelligence. Obtenido de https://www.recordedfuture.com/

Reveal. (2023). Accelerate speed to insights with augmented intelligence.

Russell, S., & Norvig, P. (2016). Artificial Intelligence a Modern Approach. *3*. Stuart Russell and Peter Norvig, Editors.

Saeed, S., Suayyid, S. A., Al-Ghamdi, M. S., Al-Muhaisen, H., & Almuhaideb, A. M. (2023). A Systematic Literature Review on Cyber Threat Intelligence for Organizational Cybersecurity Resilience. Obtenido de https://doi.org/10.3390/s23167273

Samoili, S., Cobo, M. L., Gómez, E., Prato, G. D., Martínez-Plumed, F., & Delipetrev, B. (2020). Defining artificial intelligence : towards an operational definition and taxonomy of artificial intelligence. Joint Research Centre (European Commission).

Sarker, I. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science.

Sarker, I., Furhad, M., & Nowrozy, R. (2021). AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions. *2*. SN COMPUT. Obtenido de https://doi.org/10.1007/s42979-021-00557-0

Sarker, I., Kayes, A., Badsha, S., Hamed Alqahtani, P. W., & Alex, N. (2020). Cybersecurity data science: an overview from machine learning perspective. *(41)*. Journal of Big Data. Obtenido de https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00318-5

Sealpath. (2024). Tendencias en Ciberseguridad para 2024 según los Expertos.

Sharma, A., & Sahay, S. (2019). Evolution of Malware and Its Detection Techniques. 139-150. nformation and Communication Technology for Sustainable Development. Obtenido de

https://www.researchgate.net/publication/334037102_Evolution_of_Malware_and_Its_Detection_Techniques

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Misión AI Investigación para políticas. *Inteligencia artificial: definición y antecedentes*, 15 - 41. Springer.

Sheldon, L. (2016). Implementing Information Security Architecture and Governance: A Big Framework for Small Business. Syracuse University. Obtenido de https://www.researchgate.net/publication/308522451_Implementing_Information_Security_Architecture_and_Governance_A_Big_Framework_for_Small_Business

Smalakys, T. (febrero de 2024). Cybersecurity Trends and Threats in 2024. NordPass.

SMU. (2024). CylancePROTECT. Office of Information Technology.

Sophos. (mayo de 2024). El índice de ataques de ransomware desciende ligeramente, pero los costes de recuperación alcanzan los 2,73 millones de dólares. Obtenido de https://www.revistaciberseguridad.com/2024/05/el-indice-de-ataques-de-ransomware-desciende-ligeramente-pero-los-costes-de-recuperacion-alcanzan-los-273-millones-de-dolares/

Splunk. (2023). Let's build a safer and more resilient digital world. CISCO Company. Obtenido de https://www.splunk.com/en_us/about-splunk.html

Stewart, L. (2024). ¿Qué es la investigación descriptiva y cómo se utiliza? Obtenido de https://atlasti.com/es/research-hub/investigacion-descriptiva

Sudhakar, K. (2020). An emerging threat Fileless malware: a survey and research challenges. Cybersecur. Obtenido de https://doi.org/10.1186/s42400-019-0043-x

Talukder, S., & Talukder, Z. (2020). A Survey on Malware Detection and Analysis Tools. *12*. nternational Journal of Network Security & Its Applications. Obtenido de https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3901568#paper-citations-widget

Tariq, U., Ahmed, I., Bashir, A., & Shaukat, K. (2023). A Critical Cybersecurity Analysis and Future Research Directions for the Internet of Things: A Comprehensive Review. Sensors. Obtenido de https://doi.org/10.3390/s23084117

U.S. Department of States. (2024). United States International Cyberspace & Digital Policy Strategy. *Towards an Innovative, Secure, and Rights-Respecting Digital Future*. Obtenido de https://www.state.gov/united-states-international-cyberspace-and-digital-policy-strategy/

Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2024). Adversarial Machine Learning, A Taxonomy and Terminology of Attacks and Mitigations. National Institute of Standards and Technology.

Wang, X., Zhao, Y., & Pourpanah, F. (febrero de 2020). Recent advances in deep learning. *11*, 747-750. International Journal of Machine Learning and Cybernetics.

Wevolver. (2024). 2024 State of Edge AI Report. *Exploring the Dynamic World of Edge AI Applications Across Industries*. Obtenido de https://www.wevolver.com/article/2024-state-of-edge-ai-report/introduction

William, G., Viegas, J., & Orso, A. (2006). A Classification of SQL Injection Attacks and Countermeasures. College of Computing Georgia Institute of Technology.