INTELIGENCIA ARTIFICIAL Y CIBERSEGURIDAD

Una iniciativa de

isms forum

Con el soporte de

CCN-CERT

ENERO 2024

INTELIGENCIA ARTIFICIAL Y CIBERSEGURIDAD

AUTORES

DIRECTORES	Angel Pérez Paco Lázaro
COORDINADORES	Virginia Rodríguez
PARTICIPANTES	Ana Belén Galán Auxi Ureña Diego Fernandez Diego Ruiz Elisa García Fran Gómez Hector Recio Ignacio Hornes Javier Pinillos Juan Carlos Agüero Luis Ballesteros Manuel Vidal Maria Cumbreras Miguel Angel Cabezas Noel Castillo Oriol Navarro Román Mesa Usama Alanbari
REVISORES	José Ramón Monleón
CON EL SOPORTE DE	CCN
GESTOR DE PROYECTOS	Beatriz García
DISEÑO/MAQUETACIÓN	Rim Souri

CONTENIDOS

CONTENIDOS

CONTENIDOS

INTRODUCCIÓN

> Las soluciones basadas en IA, y más concretamente los modelos que las componen, son elementos relevantes en cada vez más procesos de las organizaciones.

Este trabajo evalúa los principales ámbitos que debe tener en cuenta un profesional de la ciberseguridad para responder adecuadamente al reto de la Inteligencia Artificial (en adelante IA).

Las soluciones basadas en IA, y más concretamente los modelos que las componen, son elementos relevantes en cada vez más procesos de las organizaciones.

Este incremento de uso de la IA conlleva un aumento en la atención de cibercriminales y otros actores maliciosos sobre estas soluciones, que se traduce en el desarrollo de nuevas técnicas de ataque, y potencialmente, de impactos en las organizaciones.

Según el informe de Enisa 2022, las amenazas más frecuentes del pasado año fueron de los siguientes tipos: Ransomware, Malware, Ingeniería social, Amenazas conta los datos, amenazas contra la disponibilidad, cortes de internet, desinformación, ataques a la cadena de suministro. Según se analiza en el capítulo [Principales usos de la IA por parte de los ciberdelincuentes] el uso de la IA por parte de los ciberdelincuentes es intensivo en estos ataques.

Este trabajo evalúa cómo la Inteligencia Artificial afecta a este panorama, y contribuye a desarrollar capacidades y conocimientos técnicos para acompañar a las empresas en este nuevo paradigma.

Se reflexiona por tanto en los diferentes ámbitos para acompañar a los profesionales de ciberseguridad en su objetivo de apoyar a la organización en la protección de:

- **Continuidad de Negocio:** La ciberseguridad desempeña un papel fundamental en la continuidad de Negocio, cualquier tipo de ataque pueden interrumpir las operaciones normales de una empresa comprometiendo su capacidad para atender los compromisos de negocio. Adicionalmente una información incorrecta, sin calidad, puede llevar al paro de las operaciones. Se reflexiona también sobre la necesaria calidad de los datos, que eviten errores en la toma de decisiones e incluso sesgos o desviaciones .

- **Los datos y su confidencialidad:** La filtración, robo o divulgación no autorizada de datos, puede tener graves consecuencias, causando impactos de diversa índole (económicos, operacionales, reputacionales, entre otros) .

- **Las obligaciones legales y éticas:** la ciberseguridad debe ser un soporte del cumplimiento legal, como pueden ser en aquellas cuestiones en las que puede jugar una función dentro de regulaciones tales como la propiedad intelectual o la protección de datos de carácter personal.

- **Su Reputación:** La pérdida de confianza de los clientes y la mala reputación después de un ciberataque puede ser un factor crítico que afecte negativamente a la viabilidad y posibilidades de crecimiento de cualquier empresa

Pero también trataremos a la IA como una aliada con unas muy fuertes expectativas de su contribución para mejorar, optimizar y potenciar la eficacia y la eficiencia de las actividades de la Ciberseguridad, en especial por su aportación al análisis, predicción y apoyo a las decisiones manuales y automáticas de la prevención,

Los modelos que forman parte de soluciones de inteligencia artificial son elementos complejos que, al contrario que otras soluciones de software, en algunos casos no permiten una comprensión o medida directa sobre su funcionamiento o rendimiento. Este hecho puede tener impactos negativos que no sean directamente observables por parte de las organizaciones que las emplean.

El documento se inicia con una breve identificación de riesgos que supone el uso de la IA, enumerando algunos incidentes con repercusión mediática. El capítulo 3 identifica frameworks y guías para la reducción de riesgos de seguridad en la IA así como mecanismos de auditoría. En el capítulo 4 se describe la gestión de proyectos con IA, diferenciando entre el uso de IA como servicio externo versus el desarrollo y entrenamiento de sistemas de IA propios. El capítulo 5 entra en el uso de la IA de forma maliciosa para la realización de ataques y recomendaciones para la prevención del uso malicioso. A su vez el capítulo 6 ofrece el lado amable de la IA como herramienta para proteger la seguridad. Por último, el capítulo 7 ofrece algunas reflexiones sobre motivos por el que puede fallar una implementación de un sistema de IA.

1 RIESGOS ASOCIADOS A LA IA

De cara a poder entender cómo funciona la Inteligencia Artificial y ser capaces de aplicarla de manera efectiva en nuestras organizaciones, así como acompañar su efectivo, es preciso comprender qué son los algoritmos y sus implicaciones de uso, entrenamiento y desarrollo.

A su vez, es relevante prestar atención a los riesgos derivados tanto de la implantación de proyectos de IA dentro de la Organización, como del uso de plataformas de terceros y del uso realizado por atacantes específicamente contra nuestros proyectos de IA o contra cualquier otro servicio que estemos explotando / utilizando.

Estos riesgos son especialmente relevantes ante el uso de los modelos grandes de lenguaje (LLM), es decir, de la utilización o incorporación de redes neuronales profundas con arquitecturas complejas que han sido entrenados con grandes volúmenes de contenido no etiquetado. Estos modelos permiten realizar tareas complejas más allá del procesamiento de lenguaje natural tradicional, incluyendo la generación de contenidos o respuesta a preguntas a partir de sus datos de entrenamiento. Algunos de estos modelos permiten una interacción con el usuario mediante un sistema de entrada de datos (prompt) que recoge las preguntas o información aportada.

Con carácter general, los principales riesgos que se puede incurrir por un error o manipulación de un sistema IA son:

RIESGOS PARA LA CONTINUIDAD Y LA CALIDAD DE LAS OPERACIONES, tales como, por ejemplo:

- ***Para la seguridad física:*** Algunos sistemas con componentes de IA manejan aspectos asociados a la seguridad física (safety), no sólo de bienes tangibles como las mercancías y las instalaciones, sino incluso de las personas y otros seres vivos. Un error en este ámbito puede llegar a poner en peligro la vida.
- ***Para las operaciones:*** La mala calidad de la información puede llevar a malas decisiones operativas, así los peligros de una información sesgada o incompleta o inventada (riesgo de alucinación) proporcionada por una IA , puede presentar un grave riesgo a las operaciones e incluso a la continuidad del Negocio Así por ejemplo se habla de riesgo del contexto, el cual es inherente a las IA, actualmente muy buenas trabajando en contextos cerrados, con un gran número pero limitado de variables; esto provoca que no siempre sean capaces de diferenciar el contexto en que suceden las situaciones. Un ejemplo ocurre con la diferenciación de una luna amarillenta y la luz amarilla de un semáforo, es conocido el caso en vehículos autónomos Tesla que no podían diferenciarlo, es decir, tenían un problema de contexto.

» ***Peligro de desinformación:*** Relacionado con el anterior, pero con una identidad propia, debemos tener en cuenta que en el uso de LLM se puede generar información engañosa que provoque tener usuarios mal informados y, por tanto, erosionar la confianza en la información compartida e incluso generar información directamente alineada con lo que un usuario quiere "ver" con independencia de que esta sea verídica o no. Los usuarios tienden a sobreestimar las capacidades de los LLM dado que aparentan un comportamiento similar al de los humanos, esto puede provocar usen estos sistemas de forma insegura.

RIESGOS DE REPUTACIÓN Y CUMPLIMIENTO LEGAL Y ÉTICO, tales como:

» ***Riesgo de discriminación:*** Daños representativos y de asignación que pueden influir en que se perpetúen los estereotipos y prejuicios sociales.

» ***Automatización y daños ambientales:*** La capacitación y operación de LLM requiere mucha capacidad de cómputo, lo que comporta altos costos ambientales derivados del consumo de energía.

» ***Riesgos legales:*** Como los derivados de la utilización de datos de carácter personal fuera de las finalidades, o en número desproporcionado o sin respetar la duración acordada, entre otros.

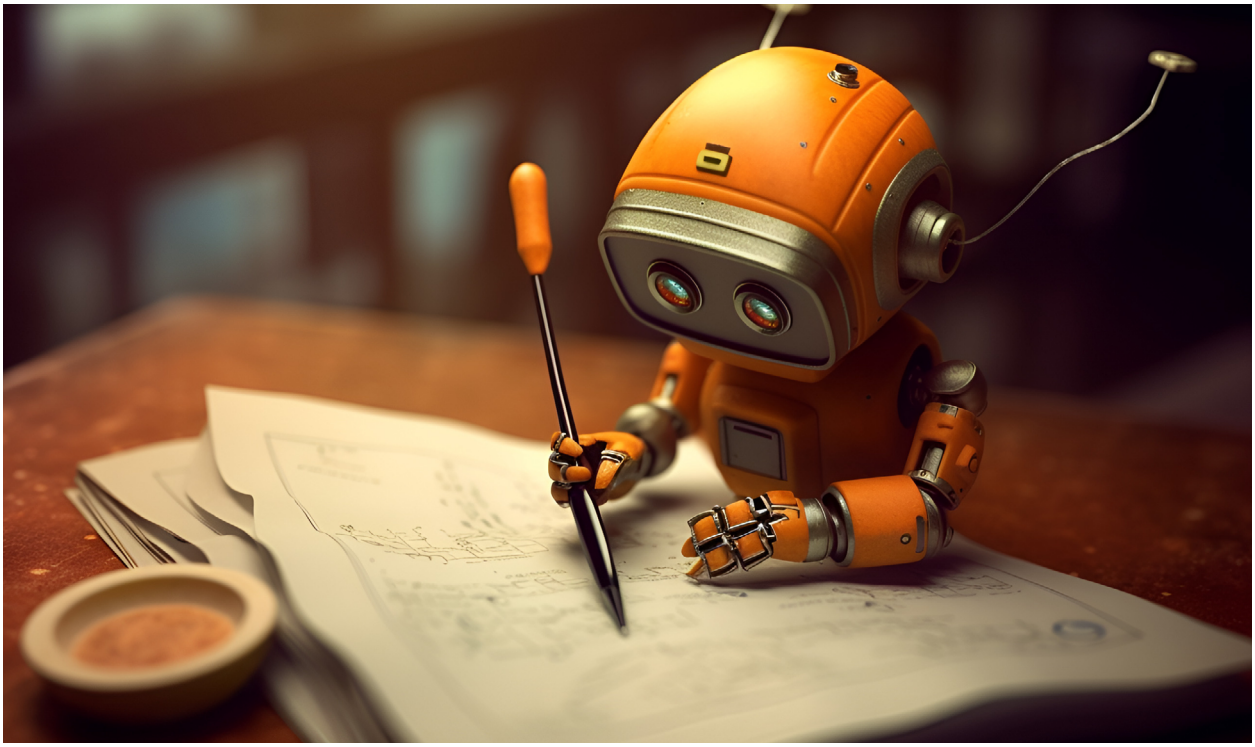
» ***Riesgo de toma de decisiones no éticas:*** Si a la IA generativa se da la posibilidad de tomar decisiones aparece el riesgo de que las mismas sean inadecuadas o no éticas.

RIESGO DE CONFIDENCIALIDAD: pueden comprometer la confidencialidad al filtrar información clasificada e inferir información confidencial, esto aplica tanto a datos personales como a cualquier otro tipo de información clasificada.

Aunque se explica más en detalle en el capítulo de "La IA como herramienta adversa", cabe citar el uso de la IA Generativa para generar amenazas WormGPT, de funcionamiento similar a ChatGPT pero sin sus limitaciones de control de uso. Lo más peligroso de todo es que cualquiera puede usarla. Se basa en el modelo de lenguaje GPTJ pero sin ninguna restricción ética. De esta manera, con ella cualquiera podría desarrollar ataques y adaptarlo con las indicaciones que le pidamos. La herramienta se promociona en múltiples foros de la Dark Web.

Pero las IAs también pueden ser atacadas con la finalidad de obtener información confidencial o manipular las respuestas que entregarán a otras personas. Los chats de inteligencia artificial están limitados para no ofrecer información confidencial, delictiva y en definitiva para evitar que ofrezcan información peligrosa o éticamente cuestionable, y, sin embargo, enfrentan el reto de ser manipuladas.

Son conocidos los ataques de prompt contra chats de IA generativa que tienen como objetivo realizar un Jailbreak, buscan manipular las limitaciones impuestas por las empresas desarrolladoras en los Chatbots de inteligencia artificial. Al fin y al cabo, hablar con estas IA es como hacerlo con una persona superdotada, una a la que han impuesto normas que no debe saltar, pero no comprende del todo por qué no debe hacerlo. Por ejemplo, es común la trampa de utilizar la psicología inversa: se solicita una lista de algún tipo a la inteligencia artificial y cuando esta responde que no puede proporcionarla por motivos legales, se le pide que indique qué elementos se deben evitar en lugar de lo que se busca en la lista.



En línea con el punto anterior OpenAI tiene un programa de Bug Bounty para que la comunidad de hackers éticos les ayude a detectar y proteger vulnerabilidad y que, el punto relevante viene cuando leyendo las condiciones dice, explícitamente "Los problemas relacionados con el contenido de las indicaciones y respuestas del modelo están estrictamente fuera del alcance, y no serán recompensados" (traducido del inglés), siguiendo en el clausulado indica aspectos específicos fuera del alcance del Bug Bounty:

Ejemplos de problemas de seguridad que están fuera del alcance:

- Jailbreaks/Bypass de seguridad (por ejemplo, DAN y solicitudes relacionadas).
- Hacer que el modelo diga cosas inapropiadas.
- Hacer que el modelo te diga cómo hacer cosas incorrectas.
- Hacer que el modelo escriba código malicioso para ti.

También están fuera del alcance las alucinaciones del modelo:

- Hacer que el modelo finja hacer cosas incorrectas.
- Hacer que el modelo finja darte respuestas a secretos.
- Hacer que el modelo finja ser una computadora y ejecutar código.

Esta limitación de alcance nos debe hacer reflexionar sobre la fiabilidad de los contenidos que generen estos modelos.

La lista de trucos para ChatGPT crece cada día en Jailbreak Chat, página web dónde se recopilan y ponen a prueba votados por los usuarios. En The Prompt Report puedes encontrar más información al respecto, así como una newsletter donde estar al tanto de este tipo de problemas.

A continuación, a modo ilustrativo y no extenso se enumeran algunos incidentes de alto impacto mediático derivados de una mala implementación de algoritmos de IA.

1.1. IDENTIFICACIÓN SESGADA - REKOGNITION

La API de reconocimiento facial de Amazon "Rekognition" presentó una problemática significativa debido a los riesgos asociados con el mal funcionamiento de los algoritmos de Inteligencia Artificial. El sistema demostró errores en la identificación precisa de personas, lo que llevó a consecuencias potencialmente perjudiciales. Estos errores generaron situaciones de discriminación racial y de género al identificar incorrectamente a individuos de ciertas etnias y géneros, lo que puso en peligro la equidad y la justicia en diversos contextos, como la vigilancia y el cumplimiento de la ley.

El mal funcionamiento de los algoritmos de Inteligencia Artificial en "Rekognition" planteó preocupaciones significativas en cuanto a la privacidad y la seguridad de los datos. El sistema tenía la capacidad de reconocer y seguir a personas sin su consentimiento, lo que generaba la posibilidad de una vigilancia invasiva y un uso indebido de la información personal. Estos errores destacaron la necesidad de abordar de manera rigurosa y ética el desarrollo y despliegue de algoritmos de Inteligencia Artificial para evitar impactos negativos en la sociedad y proteger los derechos fundamentales de los individuos.

1.2. ATROPELLOS EN COCHES AUTÓNOMOS - ELAINE HERZBERG

Es conocido que el desarrollo de prestaciones autónomas en los vehículos tendrá un impacto muy positivo en la reducción de siniestrabilidad, según informe de la Unión Europea de 2019 el error humano está involucrado en aproximadamente el 95% de todos los accidentes de tránsito, cuando se alcancen niveles óptimos globales de conducción autónoma dicha siniestralidad descenderá en su práctica totalidad.

Pero también es verdad que, actualmente, la conducción autónoma plantea riesgos para la seguridad de las personas, como se evidencia en el caso de Elaine Herzberg. La trágica muerte de Herzberg, atropellada por un vehículo autónomo de Uber en 2018, ilustra la vulnerabilidad asociada a la tecnología de conducción sin conductor. Este incidente puso de manifiesto las limitaciones y desafíos que enfrentan los algoritmos de Inteligencia Artificial en situaciones de tráfico complejas y la falta de capacidad para reconocer y reaccionar adecuadamente ante peatones y obstáculos inesperados.

Adicionalmente hay que tener en cuenta que el riesgo inherente a la conducción autónoma, y generado por la introducción de la IA, radica en la dificultad para asignar responsabilidades y tomar decisiones éticas en situaciones de emergencia. ¿a quién protege el vehículo en caso de accidente? El desarrollo de algoritmos que deben decidir entre salvar la vida del conductor o la de terceros presenta un dilema moral complejo.

Otro aspecto complejo para resolver con la evolución de la conducción autónoma será la convivencia entre vehículos conducidos por personas y vehículos autónomos; pongamos por ejemplo rotondas con gran congestión en horas punta y las acciones temerarias que realizan muchos conductores ansiosos.

1.3. SABOTAJE A VEHÍCULOS AUTÓNOMOS DESDE EL ENTORNO FÍSICO

Con la introducción progresiva de la conducción autónoma se han conocido varias posibles acciones de sabotaje cuando menos curiosas:

- Alteración de señales de tráfico mediante pequeñas marcas en las mismas: es conocido el estudio de 2018 de diversas universidades que, mediante el concepto de algoritmo adversario, llegaron a demostrar cómo se alteraba la interpretación de una señal de "Stop" por "límite 60 millas por hora" mediante la introducción de distintos puntos de color en la señal de Stop.

- Detención de vehículos autónomos mediante la inclusión de una señal Stop en la parte trasera del vehículo precedente.

- Detención de vehículos autónomos mediante el posicionamiento de un cono en el capó de este. El sistema detecta un obstáculo demasiado cerca. Este acto de sabotaje recientemente se ha puesto de moda en la ciudad de San Francisco donde se están utilizando los primeros vehículos autónomos en ambiente urbano.

1.4. RECURSOS HUMANOS DISCRIMINATORIOS – CASO AMAZON

En 2018, la utilización del sistema de inteligencia artificial por parte de Amazon para la contratación generó una problemática significativa. La intención de automatizar el proceso de contratación para agilizar la selección de candidatos para los miles de vacantes disponibles terminó siendo un desastre de relaciones públicas, ya que el sistema resultó ser sexista, favoreciendo a hombres blancos. Es probable que los datos de entrenamiento utilizados para crear el modelo estuvieran desequilibrados, lo que resultó en un sesgo en la selección de candidatos. Este incidente representa otro ejemplo de fracaso de la inteligencia artificial.

2 GESTIÓN DE RIESGOS EN IA

En este capítulo se explican diferentes frameworks y estudios sobre tipos de ataque a sistemas de IA.

2.1. MITRE ATLAS MATRIX – TÁCTICAS Y TÉCNICAS CONTRA SISTEMAS DE MACHINE LEARNING

MITRE es una organización sin ánimo de lucro cuya misión principal es la mejora de la seguridad de los sistemas y redes. Entre otros contenidos es una entidad de gran prestigio en la comunidad cíber por publicar y mantener el framework "MITRE ATT&CK", que proporciona una descripción detallada del conjunto de tácticas, técnicas y procedimientos (TTP) que utilizan los atacantes.

El TTP se ha constituido en los últimos años, junto con los IoC (Indicadores de Compromiso), en una de las fuentes de información de inteligencia más valiosas para la prevención y respuesta a ciberataques.

Ya en 2020 MITRE publicó, y actualiza periódicamente, el framework MITRE ATLAS , que brinda el conjunto de Tácticas y Técnicas (y subtécnicas) específico contra sistemas de Machine Learning.

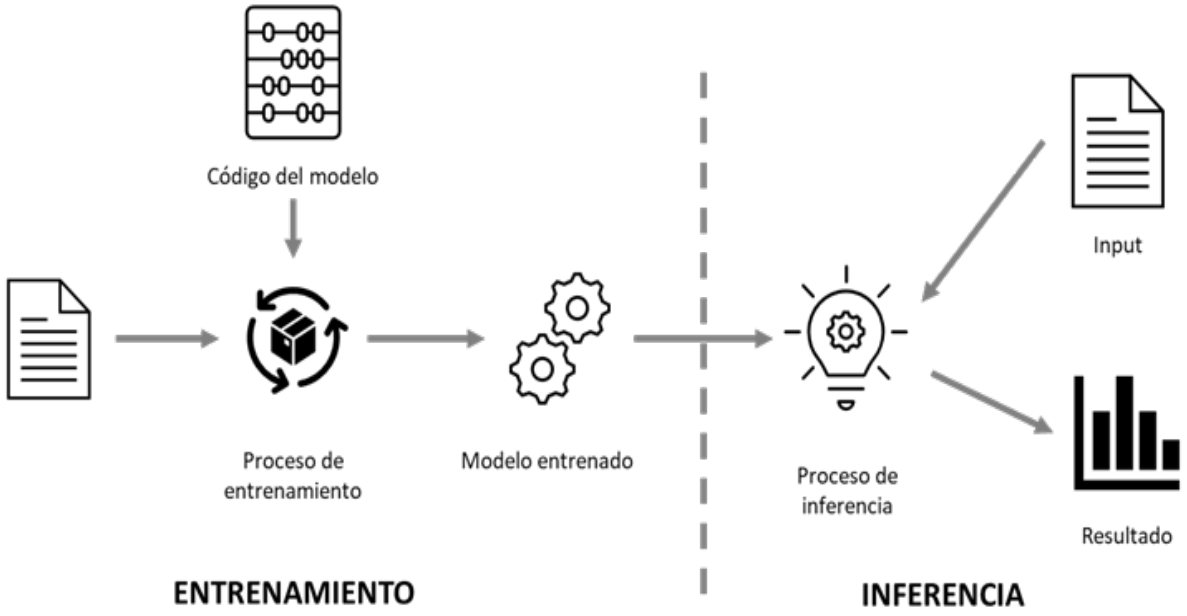
Se enumeran las Tácticas descritas en el framework en el momento de redacción de este estudio:

Táctica	Nº Técnicas	Qué intenta el adversario
Reconnaissance	5	Recopilar información sobre el sistema de ML para planificar operaciones futuras (p.ej. buscando todo tipo de información pública de la víctima y lanzando escaneos externos)
Resource Development	7	Establecer recursos que pueda utilizar en sus operaciones (p.ej.: buscar información pública sobre artefactos ML, infraestructura, credenciales, …)
Initial Access	4	Ganar acceso al sistema ML (p.ej: comprometiendo porciones del sistema ML de la cadena de suministro, abusando credenciales, …)
ML Model Access	4	Ganar algún tipo de acceso a un modelo ML (p.ej. accediendo a la API de inferencia, influyendo en el modelo mediante el acceso a la ubicación física donde se recopilan los datos y alterándolos en origen, …)
Execution	2	Ejecutar código malicioso embebido en sistema ML (p.ej. provocando que usuario autorizado lo ejecute sin darse cuenta)
Persistence	2	Lograr mantener la intrusión de forma continuada en el tiempo (p.ej. desplegando un "backdoor" en el modelo)
Defense Evasion	1	Evitar ser detectado por las defensas del sistema (p.ej. creando datos adversarios)
DiscoveryW	3	Obtener conocimiento sobre el sistema y la red interna (p.ej.: descubriendo la ontología del modelo, descubriendo los artefactos ML y obteniendo información sobre ellos, …)
Collection	3	Como continuación del paso anterior recopilar toda la información posible de contexto del entorno ML. (p.ej. exfiltrando información almacenada en los repositorios)W
ML Attack Staging	4	Adaptar el ataque al contexto del target (p.ej. creando un ML Proxy, creando datos adversarios, …)
Exfiltration	2	Robar los modelos ML o cualquier otra información relevante (p.ej. mediante cualquier mecanismo descrito en MITRE ATT&ck)
Impact	7	Afectar directamente a la víctima mediante la interrupción, manipulación o destrucción del sistema ML y sus datos

En el propio site de MITRE ATLAS se describen, en el momento de redacción de este documento, un conjunto de 18 acciones de mitigación de ejecución de técnicas y subtécnicas descritas.

A través de este framework podemos comprender el proceso de despliegue de sistemas que requieren de un entrenamiento previo, así como los dos flujos técnicos (pipelines): Entrenamiento e inferencia:

- En el flujo de entrenamiento los responsables del sistema introducen un conjunto de datos a partir de los cuales se prepara el modelo para realizar sus funciones.
- En el flujo de inferencia, un modelo ya entrenado recibe datos de los usuarios sobre los que ejecutará la tarea para el que ha sido diseñado y devuelve los resultados.



En la cadena de ataques sobre un sistema de inteligencia artificial la parte relacionada con el propio modelo puede ser considerado como el punto más disruptivo respecto del resto de ciberataques habituales.

2.2. OWASP - GUÍA SOBRE SEGURIDAD Y PRIVACIDAD

La fundación OWASP (Open Web Application Security Project), formada por un conjunto muy extenso de voluntarios y expertos en ciberseguridad, cuya misión es mejorar la seguridad del software.

OWASP ha publicado su Guía sobre la Seguridad y Privacidad en entornos de IA que también realiza descripciones de amenazas.

La siguiente tabla resume, según esta organización, las principales amenazas que pueden impactar el funcionamiento de los modelos de inteligencia artificial:

Amenaza	Descripción
Envenenamiento de datos de entrenamiento	Manipulación de datos de entrenamiento para comprometer el rendimiento del modelo. Puede tener diversas variantes: » **Modificación de etiquetas (Label modification)**: Consiste en la modificación de las etiquetas asociadas a los datos de entrenamiento de los modelos. » **Inyección de datos (Data Injection):** El atacante no tiene acceso a los datos de entrenamiento ni al algoritmo de aprendizaje, pero tiene la capacidad de agregar nuevos datos al conjunto de entrenamiento. Es posible corromper el modelo objetivo mediante la inserción de muestras contaminadas en el conjunto de datos de entrenamiento. » **Modificación de datos (Data Modification):** El atacante no tiene acceso al algoritmo de aprendizaje, pero tiene acceso completo a los datos de entrenamiento. Los datos de entrenamiento pueden ser envenenados directamente mediante la modificación de los datos antes de utilizarlos para entrenar el modelo. » **Corrupción lógica (Logic Corruption):** El adversario tiene la capacidad de interferir con el algoritmo de aprendizaje.

Evasión del modelo (Input manipulation attack)	» Técnicas de engaño al modelo para que realice clasificaciones incorrectas o predicciones erróneas. » La introducción de datos con objetivos maliciosos en la inferencia puede, según la tipología de estos y de los parámetros con los que funcione el modelo, en resultados no esperados o incluso que puedan alterar otros procesos que se apoyan en ellos. » El atacante puede disponer de conocimientos previos que le ayuden a preparar estas técnicas. Un ejemplo de ello puede ser pintar una señal de tráfico de color rojo, que podría confundir un reconocedor de señales y hacer crear al sistema que se trata de una señal de stop. Los sistemas de IA generativa suelen disponer de un sistema de entrada para peticiones o preguntas del usuario (prompt). Esta entrada puede ser manipulada por el atacante para conseguir saltarse restricciones del sistema generativo y hacer que genere información o actividades maliciosas.
Inversión del modelo	La interacción con la inferencia de un modelo puede permitir a un atacante recuperar o deducir características de los datos de entrenamiento han sido usados. Esta técnica puede permitir la preparación de otros ataques con más impacto sobre el modelo, o bien revelar datos sensibles que han sido usados en el entrenamiento del modelo.
Ataque de inferencia de pertenencia	Similar a la inversión del modelo, un ataque de inferencia de pertenencia puede deducir si un registro de datos pertenece al conjunto de entrenamiento usado en el modelo.
Extracción de información sobre los modelos	Mediante la interacción con el modelo, un actor malicioso puede deducir qué características tiene, incluso llegando a poder replicar su funcionamiento completo.
Ataques sobre la cadena de suministro o a la plataforma operativa del modelo	Los ciberataques tradicionales sobre varios componentes de los que depende la cadena de suministro de un modelo pueden crear varios niveles de impacto sobre sistemas productivos. Por ejemplo, comprometer una base de datos usada para el entrenamiento de modelos, o bien la plataforma operativa en la que se ejecuta el motor de inferencia de un modelo desplegado en producción.
Extracción de modelos	intento de replicar el modelo de ML a partir de sus respuestas a consultas.
Robo de propiedad intelectual	extracción de información confidencial o propietaria del modelo de ML.

Ataques de privacidad	obtención de información sensible sobre los datos de entrenamiento mediante consultas al modelo.
Ataques de reidentificación	Vinculación de registros anonimizados del modelo de ML con personas reales.
Ataques de adversario universal	Identificación de patrones de entrada que causan malas predicciones en un amplio conjunto de modelos.
Ataques de oráculo	Explotación del modelo de ML como un «oráculo» para resolver problemas que no fueron diseñados para abordar.
Ataques de suplantación	Manipulación del sistema de IA para que confíe en un adversario, lo que permite el acceso no autorizado a datos y recursos.
Ataques de denegación de servicio (DoS)	Inundar el sistema de IA con solicitudes para agotar sus recursos y hacerlo inaccesible.

2.3. OWASP TOP 10 FOR LARGE LANGUAGE MODELS (LLM)

Los modelos grandes de lenguaje (LLM) se componen de arquitecturas complejas basadas en redes neuronales que han sido entrenadas con grandes conjuntos de datos. Algunos de estos modelos (como el conocido ChatGPT, Google Bard, Bing, etc.) disponen de un sistema de interacción a partir de una entrada de texto para el usuario (prompt). En algunos sistemas más evolucionados se dispone incluso de una interacción multimodal (cuya entrada pueden ser textos, imágenes, sonidos, etc.).

El desarrollo de estas soluciones, y su uso en procesos de negocio de organizaciones en todo el mundo, ha incrementado rápidamente su presencia. Se espera que esta ubicuidad siga como tendencia.

La organización OWASP ha desarrollado un Top 10 de vulnerabilidades potenciales para tener en cuenta en estos sistemas:

Vulnerabilidad	Descripción
LLM: Prompt Injection	Algunas implementaciones de modelos LLM suelen disponer de un sistema de entrada para peticiones o preguntas del usuario (prompt). Esta entrada puede ser manipulada por el atacante para conseguir saltarse restricciones del sistema generativo y hacer que genere información o actividades maliciosas.
LLM02: Insecure Output Handling	Esta vulnerabilidad podría ser presente si se usan sin revisión los datos de salida de un modelo LLM. Por ejemplo, en un entorno de aplicación web, si alguno de los textos que se muestran en partes interactivas de la aplicación han sido generados por un modelo LLM, el atacante podría conseguir que estos textos incluyeran código JavaScript, incluyendo posibles cadenas de ataque de XSS CSRF o SSRF.
LLM03: Training Data Poisoning	El envenenamiento de datos de entrenamiento ocurre en sistemas LLM de manera similar a la descrita en este documento sobre el resto de los modelos. En el caso particular de los modelos grandes de lenguaje, al ser en muchos casos entrenados a partir de información pública que incluyen datos generados por los usuarios (como Wikipedia, Reddit, etc…) la presencia de información que pudiera generar corromper la funcionalidad, efectividad o sesgos es más difícilmente controlable.
LLM04: Model Denial of Service	Los ataques de denegación de servicio (DoS) ocurren en modelos LLM como en los tradicionales. Para aquellos LLMs expuestos públicamente que aceptan un prompt como entrada este vector de ataque aumenta en su probabilidad, y podría permitir a un atacante crear peticiones que hicieran un consumo de recursos elevado que comprometiera la plataforma operativa en la que funciona el modelo.
LLM05: Supply Chain Vulnerabilities	El ciclo de desarrollo y producción de un modelo LLM podría verse afectado por componentes de terceros vulnerables, incluidos conjuntos de datos o modelos pre-entrenados.
LLM06: Sensitive Information Disclosure	Estos ataques, como la inversión de modelos, afectan también a sistemas LLM. Un atacante podría revelar datos confidenciales sobre el propio modelo, sus datos de entrenamiento o datos sensibles de la plataforma operativa a partir de peticiones maliciosas.
LLM07: Insecure Plugin Design	Los plugins que puedan integrarse en los sistemas donde funcionan los modelos LLM pueden desactivar otros controles sobre los parámetros de entrada o control de acceso de los sistemas.
LLM08: Excessive Agency	El exceso confianza sobre el funcionamiento de los modelos LLM puede llevar a acciones no deseadas cuando éstos tienen otorgadas capacidades o privilegios excesivos en la organización.

LLM09: Overreliance	El exceso de dependencia de procesos de negocio sin una supervisión adecuada podría generar contenido malicioso, desinformación e incluso vulnerabilidades en los sistemas y aplicaciones de la organización.
LLM10: Model Theft	El robo de modelos puede ser posible mediante la extracción de funcionamiento del modelo o bien mediante el aprovechamiento de fallos operativos de la plataforma. Podría tener implicaciones sobre la propiedad intelectual de la organización que desarrolla el modelo.

2.4. NIST AI RISK MANAGEMENT FRAMEWORK

NIST (National Institute of Standards and Technology) es una agencia pública del gobierno de Estados Unidos cuya misión es promover la innovación y competencia industrial.

En la perspectiva de ciberseguridad NIST es ampliamente conocido por la comunidad por su Cybersecurity Framework (CSF), adoptado como un estándar de referencia a nivel internacional

Esta organización ha lanzado toda una línea de trabajo, recogida en su Centro de recursos para una AI Confiable y Responsable (disponible en la sección Trustworthy and Responsible AI Resource Center).

En su site se pueden encontrar recursos para la comprensión y gestión del riesgo de la inteligencia artificial que permiten ayudar en la aplicación real del Framework de Riesgo propuesto por el NIST, entre las que cabe destacar la guía para la implantación (NIST AI RMF Playbook), su vídeo explicativo (AI RMF Explainer Video), así como interesantes perspectivas sobre el futuro de la IA (Perspectives).

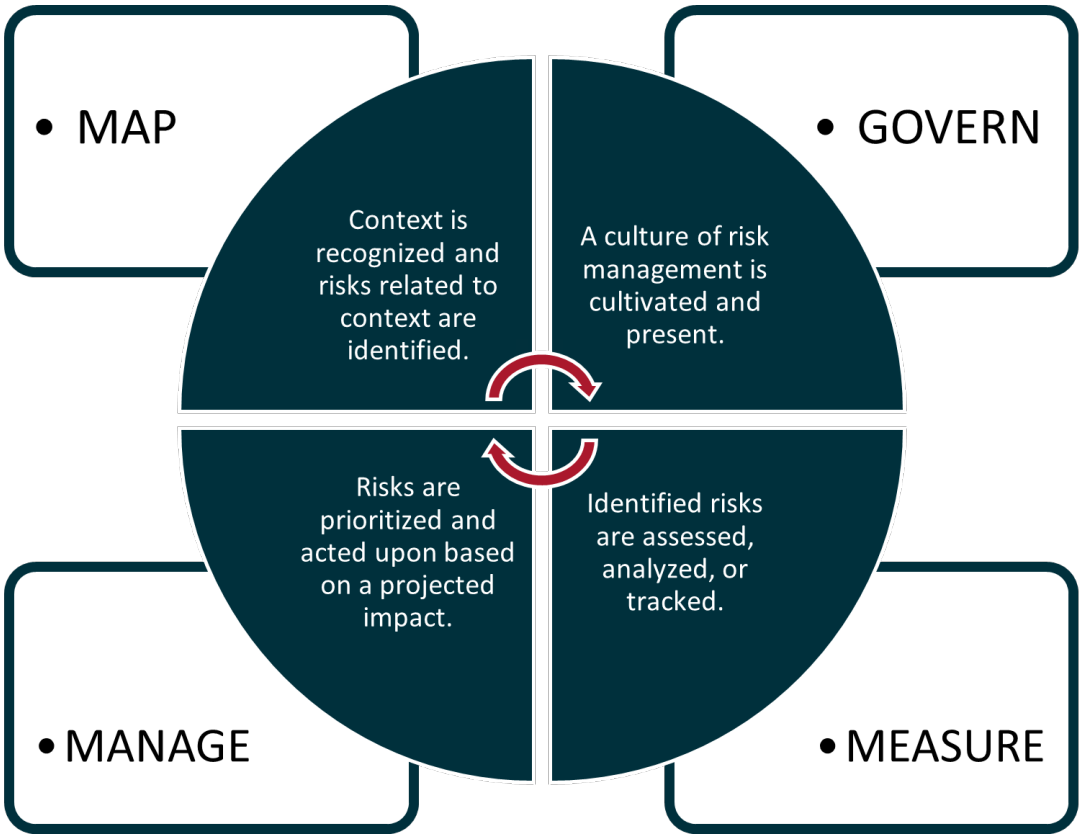
El framework, titulado "AI Risk Management Framework " (AI RMF 1.0) fue publicado a principios de 2023, y resultado de un largo trabajo en régimen de Colaboración Público-Privada que involucró a más de 240 organizaciones.

El objetivo de esta iniciativa es permitir a las organizaciones de cualquier tipo diseñar, desarrollar, desplegar y utilizar los sistemas de AI de una manera confiable y responsable.

La lectura de este framework proporciona herramientas para:

- Comprender las principales amenazas (a las personas, las organizaciones, la sociedad y el medio ambiente).
- Entender las principales características de los sistemas de IA para poder identificar los riesgos inherentes al uso de estas tecnologías.
- Identificar en qué varían los riesgos asociados a la IA respecto al riesgo tradicional que estamos acostumbrados a gestionar.
- Gestionar los riesgos, incluyendo la medida del riesgo, la determinación de los umbrales de tolerancia, la priorización de los riesgos
- Tener enfoques para la integración de la gestión del IA en la actual gestión de riesgos de las compañías.

Para conseguir estos objetivos, se proporciona una guía para poder desarrollar las 4 funciones que establece como clave: Gobernar, mapear, medir y gestionar los riesgos asociados a la IA.



2.5. PUNTOS RELEVANTES EN AUDITORÍAS DE MODELOS DE INTELIGENCIA ARTIFICIAL

Según se detalla en el capítulo de Proyectos con algoritmos de IA, auditar los modelos de Machine Learning/ Inteligencia Artificial (ML/IA) es fundamental para mitigar amenazas de ciberseguridad y operativas. A continuación, se presentan los puntos relevantes a tratar desde una perspectiva de auditoría en torno a estos modelos:

- **Gobierno:** Es conveniente establecer un marco de gobierno sólido para los modelos de ML/IA. Esto implica definir políticas y procedimientos claros para su desarrollo, implementación y mantenimiento.

  Auditorías posteriores deben evaluar si se siguen las mejores prácticas de gobernanza, como la asignación adecuada de responsabilidades, la gestión de riesgos y la transparencia en la toma de decisiones. Este proceso de auditoria debe estar presente a lo largo de todo el ciclo de vida del sistema de IA.

- **Privacidad:** Dentro del ciclo de vida hay ciertas fases en las que la privacidad tiene una relevancia especialmente importante.

  Se deben evaluar las medidas de seguridad implementadas para evitar el acceso no autorizado durante el diseño de los sistemas, a el uso indebido o la divulgación de información sensible. En la parte que se refiere al algoritmo esto afecta:

  - En la fase del diseño en lo que se refiere a la cadena de proveedores, es decir, todas las dependencias que sean usadas en el proceso de creación del modelo y su entrenamiento.
  - En la fase de despliegue asegurando que los datos sensibles o que estén dispuestos en diferentes niveles de acceso en base a unos criterios de autorización no sea expuestos.
  - Y por último en la fase de monitorización para detectar posibles ataques capaces de extraer información haciendo un mal uso del modelo.

- **Funcionamiento/Precisión:** La auditoría debe evaluar la precisión y el rendimiento de los modelos de ML/IA. Esto implica verificar la calidad de los conjuntos de datos utilizados, la validez de las técnicas de entrenamiento y los criterios de evaluación utilizados. También se deben analizar las métricas de rendimiento, como la tasa de aciertos y los falsos positivos/negativos, para asegurar que los modelos estén funcionando de manera óptima y detectar posibles ataques que se hayan realizado sobre los algoritmos y que puedan afectar al funcionamiento de estos. Principalmente estos controles deben de estar enfocados a las fases de despliegue y monitorización.

- **Ética y sesgos:** Tanto en la fase de diseño como en la fase de monitorización es necesario realizar auditorías enfocadas a examinar si los modelos de ML/IA están sujetos a sesgos injustos o discriminación. Se deben analizar los datos de entrenamiento en busca de sesgos y se deben implementar medidas para mitigarlos. Además, es importante evaluar si se han establecido salvaguardias éticas para evitar decisiones o resultados injustos y discriminatorios.

- **Monitorización:** Es importante concienciar a las empresas que la participación de ciberseguridad no se debe de realizar únicamente en la fase de proyecto ya que es necesario monitorizar de manera continua su buen funcionamiento posterior, así como colaborar en la detección de posibles anomalías en su funcionamiento.

3 PROYECTOS CON ALGORITMOS DE IA

Antes pasar a analizar los riesgos queremos hacer una reflexión previa sobre la implantación de sistemas de IA. Si bien los sistemas de IA son considerados como una tecnología disruptiva los pasos previos que hay que dar no difieren mucho de los deberíamos realizar antes de incorporar algo nuevo en nuestra organización, por ello deberemos:

- Identificar las necesidades de nuestra organización o los problemas que deseamos resolver.
- Evaluar si existen otras alternativas a un sistema IA, hay que tener en cuenta que su implantación y gestión no es trivial.
- Identificar los resultados que deseamos obtener
- Establecer un presupuesto y ver la disponibilidad de recursos que podríamos necesitar y si están alineados con la dimensión organización.

Los modelos que forman partes de soluciones de inteligencia artificial son elementos complejos que, al contrario que otras soluciones de software, en algunos casos no permiten una comprensión o medida directa sobre su funcionamiento o rendimiento. Este hecho puede tener impactos negativos que no sean directamente observables por parte de las organizaciones que las emplean.

Existen dos elementos clave que será necesario realizar y para los que recomendamos la lectura en detalle del apartado de Existen dos elementos clave que será necesario realizar:

- **Modelado de amenazas:** Tanto el entorno tecnológico operativo como los propios algoritmos y modelos requieren de una observación detallada de los riesgos que presentan estas soluciones. Es conveniente no sólo modelar, sino también monitorizar, aquellas amenazas que puedan conseguir alterar el funcionamiento o precisión de los modelos.
- **Evaluación y auditoría:** Las medidas de precisión de algunos modelos, como en los casos de IA generativa, no puede realizarse como en modelos de aprendizaje automático supervisados. Se requiere de frameworks de referencia. Asimismo, es imprescindible realizar estas tareas de forma continua en el tiempo para evitar el impacto de amenazas sobre estos sistemas.

Entidades públicas y privadas están desarrollando marcos de revisión para soluciones de inteligencia artificial. La aparición de nuevas amenazas y técnicas de ataque hará que estos marcos modifiquen sus requisitos para acomodar nuevos puntos de revisión. Los requisitos de transparencia y explicabilidad son en sí mismos un reto en modelos que pueden tener billones de hiperparámetros.

Si bien este documento se refiere al uso de IA y ciberseguridad, es importante destacar también la importancia de colaborar en el desarrollo de soluciones que respeten el medio ambiente ya que el entrenamiento de modelos de IA complejos es una tarea que requiere capacidades de cómputo muy elevadas.

Las infraestructuras tecnológicas que pueden ejecutar estos modelos en tiempo razonable tienen un impacto ambiental muy elevado. Existen iniciativas para introducir principios de sostenibilidad en los sistemas de IA, tanto a partir de la optimización de los algoritmos actuales como de generar infraestructuras con menor impacto ambiental.

3.1. USO DE IA COMO CONSUMIDOR

Este es el enfoque en el que la organización adquiere e integra un sistema de IA comercializado por un tercero.

Es la opción que menos recursos consume y que, probablemente, se despliegue más rápidamente, o sea la más viable de implantar en cualquier organización, pero esto no implica que esté exenta de riesgos, precisamente esto es lo que se ve a lo largo del capítulo: los riesgos a los que estamos expuestos y recomendaciones sobre cómo afrontarlos

Nos podemos encontrar con dos situaciones de productos:

- Aquellos que claramente son IAs o incorporan IAs de forma explícita.
- Aquellos otros productos que, incorporando elementos de IA, o bien no lo especifican o bien son un componente más del producto. Por ejemplo, en Microsoft 365 hay más de 100 componentes de IA (reconocimiento de voz, recomendaciones, corrección automática, traducción, análisis de datos, creación de contenido, DLP, IRM, …) y sin embargo no identificamos a Microsoft 365 como un producto IA.

La elección de una tecnología, y proveedor, no debe realizarse de forma precipitada, es importante investigar y comparar entre varios antes de tomar la decisión final, ya que esta puede marcar el éxito o el fracaso de la implantación de una IA en la organización. Por ello, conviene tener en cuenta los siguientes aspectos:

- **Experiencia y conocimiento técnico:** Es importante que disponga de un equipo sólido de científicos e ingenieros con experiencia.
- **Reputación:** Los casos de éxito, las referencias o las reseñas son fundamentales para establecer una confianza en el mismo.

- **Productos y servicios:** Es importante que estos se adapten al alcance y a los resultados esperados. Unos servicios personalizados nos permitirán solucionar de forma más adecuada nuestras necesidades.
- **Calidad y rendimiento:** La precisión, escalabilidad, integración y capacidad de adaptación a las necesidades definidas son criterios fundamentales.
- **Ética:** El uso de las IA requiere de un enfoque ético en su desarrollo y uso, la adhesión a protocolos o pactos nos puede ayudar en este caso. Esto es especialmente importante cuando se puedan producir situaciones de discriminación o sesgo.
- **Transparencia:** Es fundamental que nos puedan identificar la forma en que los modelos de IA toman las decisiones. Algunos sistemas de IA, como los deep learning, que pueden llegar a ser difíciles de entender. Es fundamental que la organización pueda explicar y comprender la naturaleza de las soluciones proporcionadas por la IA.
- **Seguridad:** Como clientes habrá muchas medidas que solo el proveedor podrá llevar a cabo, por ello es crucial que disponga de medidas sólidas en infraestructuras y sistemas.
- **Cumplimiento normativo:** Deberá proporcionar mecanismos que nos permitan verificar el cumplimiento de las regulaciones y normativas aplicables, especialmente en lo que respecta a la protección de datos y la privacidad.
- **Acuerdos de nivel de servicio (SLA):** Donde se establezca claramente la relación contractual y queden recogidos los posibles cambios en las políticas de uso, en el soporte o en una posible terminación repentina del soporte, ya que todo ello podría afectar a la operativa de la organización.
- **Evaluación previa a la contratación:** Se debería realizar una prueba que permita comprobar la calidad de la tecnología, la capacidad para dar respuesta a las necesidades requeridas y poder valorar el desempeño de la solución en la organización.

3.1.1. ANÁLISIS DE RIESGOS Y PRINCIPALES MEDIDAS A CONSIDERAR

Se plantea el análisis desde dos vertientes distintas:

Riesgos asociados a la tecnología, donde todos, o la mayoría, quedarán del lado del proveedor:

Riesgo	Medidas mitigantes
Mala calidad de los contenidos o data sets, sobre todo en aquellos casos en los que el proveedor o bien no es los suficientemente transparente o no se responsabiliza sobre el control, la revisión, o la idoneidad de los contenidos.	Seleccionar un proveedor y tecnología en base a su: ▪ Reputación. ▪ Experiencia y conocimiento técnico (debería disponer de un equipo sólido de científicos e ingenieros). ▪ Transparencia de los algoritmos. ▪ Compromiso y adhesión a códigos éticos o certificaciones en IA.
Medidas de seguridad inapropiadas, que puede desembocar en un ciberataque con el resultado de destrucción compromiso o indisponibilidad de los sistemas informáticos.	▪ Requerir certificaciones de seguridad contrastadas (ENS, ISO27001, ...). ▪ Establecer acuerdos contractuales (Acuerdos de nivel de servicio o SLA) que aborden: seguridad, privacidad, transparencia, responsabilidad, tiempos de respuesta, actualizaciones, calidad de los data sets, continuidad del servicio... ▪ Evaluar los riesgos del proveedor, a través de sus análisis de riesgos, en su: sistema, tecnología o cadena de suministro (proveedores o subcontratas). ▪ Requerir informes de seguridad periódicos.

Riesgos asociados al uso, siendo estos los que quedan en el lado de la organización que lo integra:

Riesgo	Medidas mitigantes
Fuga de información confidencial, el uso de estos sistemas conlleva que los usuarios pueden verse tentados a compartir, con el sistema IA, información sensible de la organización, o datos de carácter personal, con la finalidad de resolver un problema. Es un punto crítico ya que incluso esta fuga, podría tener consecuencias penales.	▪ Establecer políticas de uso de la IA. ▪ Aplicar cláusulas de confidencialidad y deber de secreto para trabajadores y colaboradores. ▪ Implantar planes de sensibilización y concienciación sobre el uso tecnologías IA. ▪ Concienciar y formar, de forma periódica, en tecnologías IA, seguridad de la información y protección de datos.
Brechas de privacidad, mediante la revelación de información personal a la IA sin la autorización de los titulares de los datos	▪ Aplicar cláusulas de confidencialidad y deber de secreto para trabajadores y colaboradores de todo tipo. ▪ Concienciar y formar, de forma periódica, en tecnologías IA, seguridad de la información y protección de datos.
Brechas de seguridad, por la descarga de aplicaciones, herramientas o códigos basados en IA y de origen incierto.	▪ Establecer de políticas de uso de la IA. ▪ Concienciar y formar, de forma periódica, en tecnologías IA, seguridad de la información y protección de datos

Derechos de autor, muchas IA no garantizan los derechos de propiedad intelectual lo cual pueden provocar que se utilicen contenidos, generados por la IA, que puedan tener la consideración de plagio, vulnerando derechos de terceros.	▪ Incorporar políticas de uso ético de las IA. ▪ Concienciar y formar, de forma periódica, en tecnologías IA, seguridad de la información y protección de datos.
Protección del usuario, este riesgo está alineado con los requisitos del futuro Reglamento de IA de la UE, en concreto se deberá informar al consumidor de que está interactuando con un sistema de IA. Además de evitar prácticas que puedan suponer una discriminación del consumidor.	▪ Incorporar políticas de uso ético de las IA. ▪ Informar a los consumidores de la tecnología IA utilizada. ▪ Auditar los procesos, especialmente los de atención al cliente.
Principios éticos, hay que tener en cuenta que actualmente las IA no tienen la capacidad de diferenciar entre lo correcto y lo incorrecto, lo bueno y lo malo, o lo verdadero o lo falso. Esto puede provocar que la información procesada se ve afectada por sesgos o por formad de discriminación.	▪ Incorporar políticas de uso ético de las IA. ▪ Supervisar los resultados proporcionados por la IA.
Falta de control, al ser una solución proporcionada por un tercero, se puede tener una falta de control sobre el desarrollo y mantenimiento de la tecnología. Esto puede dificultar la identificación y corrección de vulnerabilidades o la implantación de medidas de seguridad específicas.	▪ Establecer acuerdos contractuales (Acuerdos de nivel de servicio o SLA) que aborden: seguridad, privacidad, transparencia, responsabilidad, tiempos de respuesta, actualizaciones, calidad de los data sets, continuidad del servicio... ▪ Supervisar los resultados proporcionados por la IA.
Dependencia tecnológica, si la organización se vuelve altamente dependiente de una solución IA de un tercero, la interrupción del servicio o un problema con el proveedor podría tener consecuencias negativas en la operativas de la organización	▪ Diversificar o disponer de otros proveedores, con características similares. ▪ Establecer planes de contingencias o de continuidad de negocio.
Transparencia, si la toma de decisiones o la generación de resultados no es fácilmente comprensible se pueden plantear problemas de credibilidad, de confianza o de rendición de cuentas a los stakeholders.	▪ Incorporar políticas de uso ético de las IA
Suplantación de identidad y accesos no autorizados, provocados por una inadecuada gestión de permisos o roles de usuarios pudiendo exponer la información de la organización a terceros no autorizado.	▪ Establecer políticas de usuarios que incluyan: contraseñas robustas, sistemas de MFA, revisión de roles, logs de accesos o control de accesos y sesiones, entre otras medidas. ▪ Concienciar y formar, de forma periódica, en tecnologías IA, seguridad de la información y protección de datos.
Interceptación de información, acceso a la información gestionada con la IA.	▪ Cifrar la información en tránsito (TSL, VPN, ...). ▪ Cifrar los datos, o contenedores de datos, resultado del uso de la IA. ▪ Realizar copias de seguridad de los datos.

3.2. PARTICIPACIÓN EN PROYECTOS DE IA CON APRENDIZAJE

Este es el enfoque en el que la organización decide desarrollar modelos propios de IA.

Las soluciones basadas en Inteligencia Artificial y aprendizaje automático engloban un conjunto de metodologías y soluciones capaces de dotar a sistemas automáticos de capacidades inspiradas en la inteligencia humana: análisis de información, toma de decisiones o generación de contenidos.

En el contexto de la inteligencia artificial, los algoritmos son descripciones de una serie de pasos que permiten definir los parámetros del modelo que va a realizar las tareas que podrán solucionar un problema. El modelo es el producto final que puede realizar las tareas definidas.

Tradicionalmente, los distintos campos de la inteligencia artificial engloban las técnicas de aprendizaje automático, que en sus implementaciones con redes neuronales se denomina aprendizaje profundo.

Los modelos de aprendizaje automático que puede desarrollar la organización se dividen en las siguientes categorías según su entrenamiento:

- **Aprendizaje supervisado:** Donde partimos de conjuntos de datos de entrenamiento y conjuntos de pruebas con sus respuestas informadas.
- **Aprendizaje no supervisado:** Donde partimos de conjuntos de datos no informados.
- **Aprendizaje semi-supervisado:** Aquellos casos en los que se usan conjuntos de entrenamiento parcialmente informados.
- **Aprendizaje por refuerzo:** Algoritmos que basan su aprendizaje a partir de la información del entorno en el que se ejecutan.

El procesamiento de lenguaje natural incluye modelos que permiten la comprensión del lenguaje humano por parte de sistemas automáticos. Estos modelos se entrenan de forma supervisada para realizar tareas específicas, por ejemplo, clasificaciones sobre el contenido de textos y traducciones automáticas.

3.2.1. ANÁLISIS DE RIESGOS Y PRINCIPALES MEDIDAS A CONSIDERAR

3.2.1.1. IDENTIFICACIÓN Y PREPARACIÓN DE LOS DATOS.

Las fuentes de datos utilizadas en proyectos de Inteligencia Artificial (IA) pueden variar ampliamente dependiendo del proyecto que queramos abordar y del tipo de aplicación de IA que tengamos previsto utilizar.

IDENTIFICACIÓN DE LOS DATOS

Es importante entender la información que será procesada, realizando una identificación de los datos:

Tipos de datos:

» **Estructurados:** Organizados en filas y columnas y que fundamentalmente son datos de texto como son bases de datos, hojas de cálculo, etc.

» **No estructurados:** Son datos que no tienen un formato definido y no se pueden almacenar fácilmente en una tabla. Podemos encontrar ejemplos como:

- Texto – Correos electrónicos, comentarios...
- Audio – Grabaciones fundamentales para los modelos basados en LNP...
- Imagen – Muy útiles para el entrenamiento de coches autónomos, por ejemplo...
- Video – permite el entrenamiento de modelos que tenga como finalidad la identificación facial...

Fuente origen de los datos:

» **Fuentes Internas:** Muchas organizaciones tienen bases de datos internas que contienen datos relevantes para sus operaciones comerciales. Estas bases de datos pueden incluir registros de clientes, registros de transacciones, etc., que son muy útiles para el entrenamiento de modelos.

» **Fuentes Externas:** Existen numerosas fuentes de datos públicas disponibles en línea que pueden ser utilizadas en proyectos de IA como pueden ser datos de investigación académica, datos gubernamentales, datos de encuestas.

MEDIDAS DE SEGURIDAD EN LA GESTIÓN DE LOS DATOS

Controles normativos para el aseguramiento de la Calidad de los datos:

- Procedimiento para gestionar y garantizar una adecuada gobernanza de los datos (garantizar la exactitud, integridad, fiabilidad, veracidad, actualización y adecuación del conjunto de datos).
- Mecanismos de supervisión de los procesos de recopilación, tratamiento, conservación y utilización de los datos.
- Análisis y cuantificación de la muestra utilizada para el entrenamiento del modelo y verificación de la adecuación del tamaño de la muestra.
- Análisis de la muestra utilizada para el entrenamiento del modelo y verificación de la representatividad del conjunto final de datos con relación a la población del contexto.
- Verificación de que la distribución de las variables es adecuada.
- Procedimientos para analizar, medir y detectar posibles desequilibrios en la cantidad de datos.
- Análisis de compensación, estableciendo la relación entre la cantidad y tipología de datos a ser recogidos/descartados y aquellos necesarios para garantizar la efectividad y eficiencia del componente.
- Análisis del tamaño de la muestra para la conservación de datos con propósito de auditoría**.**

Uso proporcional de información y determinación del origen de las fuentes de datos

- Identificación del contexto de origen y las fuentes de datos utilizados para el entrenamiento y validación del modelo.
- Documentar y justificar el proceso de elección de las fuentes de datos utilizadas para el entrenamiento.
- Establecer base legitimadora para el uso de datos personales para cada una de las etapas del ciclo de vida.
- Justificar la recogida y el empleo de datos personales, que no son necesarios en la etapa de entrenamiento, que permite realizar una comprobación del comportamiento del modelo en las etapas posteriores de verificación y validación del componente.
- En el caso de usar datos personales sensibles, se debe evaluar la necesidad de su uso y que existe una circunstancia que justifica levantar la prohibición general de su tratamiento.

Controles en la Determinación de los Destinatarios de los datos

» Obligaciones de información a los interesados con relación al tratamiento de datos ya sean datos obtenidos de los interesados directamente u obtenidos de terceros

» En el caso de transferir datos a terceros países es necesario valorar si estos países se encuentran dentro del EEE (Espacio Económico Europeo) o fuera de él. En el caso de transferencias internacionales será necesario valorar la existencia o ausencia de una decisión de adecuación de la Comisión.

» Los destinatarios de los datos deberán aparecer identificados en la actividad o actividades del Registro de Actividades de Tratamiento en las que se inscribe el uso del componente IA.

Controles sobre la limitación de la conservación de datos

» Definir una política de conservación de la muestra de datos de entrenamiento, así como una política de conservación en el registro de actividades y aplicar estrategias de privacidad (minimización, ocultación, separación o abstracción de datos) para su explotación.

» Fijar procedimientos para verificar la aplicación de criterios, medidas y plazos de conservación.

» Procedimiento de revisión del análisis de la necesidad y proporcionalidad de la conservación de los datos.

Controles sobre las Categorías de los Interesados

» Identificar las categorías de los interesados y las consecuencias a corto y largo plazo que la implementación del componente IA puede suponer.

» Definir los procedimientos necesarios para analizar el contexto social en el que se enmarca el uso del componente y recabar información a través de personas, grupos u organizaciones afectadas

3.2.1.2. DATOS DE ENTRENAMIENTO Y DATOS DE PRUEBA

Para realizar una evaluación correcta del modelo de inteligencia artificial utilizada debemos tener en cuenta varios aspectos a la hora de dividir los datos de entrenamiento y test:

- **Tamaño de la muestra:** Debemos tener suficientes datos en ambos conjuntos. Si el conjunto de entrenamiento es demasiado pequeño, el modelo puede tener dificultades para aprender patrones y generalizar adecuadamente. Por otro lado, si el conjunto de prueba es demasiado pequeño, la evaluación del rendimiento puede no ser representativa de cómo el modelo se comportará en datos nuevos.
- **Representatividad de los datos:** Ambos conjuntos de datos deben ser representativos de los datos reales a los que se enfrentará el modelo una vez que esté en producción. Deben tener la suficiente variabilidad y diversidad que podemos encontrarnos en la situación real.
- **Aleatoriedad:** La división de los datos para ambos conjuntos debe ser totalmente aleatoria ya que garantiza que los conjuntos sean independientes evitando sesgos no deseados. Utilizar formulas aleatorias ayuda a generalizar los resultados dando mayor robustez ante la entrada de datos nuevos.
- **Proporcionalidad.** La división entre ambos conjuntos de datos dependerá de la complejidad del proyecto y del volumen de datos disponibles. La proporción más común utilizada sería 70% para los datos de entrenamiento y un 30% para el test.

  Los datos que utilicemos para entrenar y evaluar el modelo deben representar el estado actual de los datos que se encontrarán en producción. Si tenemos datos que pueden cambiar en el tiempo, se hace necesario actualizarlos periódicamente en los conjuntos de entrenamiento y prueba para asegurarse de que el modelo se ajuste a los datos más recientes.

MEDIDAS DE SEGURIDAD – GESTIÓN DE LOS DATOS

Análisis y minimización de datos

» Identificar y documentar los criterios de depuración previa de los datos originales a lo largo de las diferentes iteraciones en el proceso de entrenamiento del componente.

» Argumentar y documentar las técnicas y buenas prácticas de limpieza de datos.

» Documentar la estructura y propiedades del conjunto de datos tratados, en número de sujetos y extensión de datos utilizados.

» Categorización previa de los datos utilizados, organizándolos en datos no personales y personales.

Distinción de procesos de entrenamiento y validación

» Implementar estrategias de segregación y disociación sobre la información adicional que no es necesaria para el entrenamiento pero que será necesaria en los procesos de verificación y validación del comportamient

» Preprocesar y depurar previamente de los datos utilizados para el entrenamiento y validación del componente IA, detectando las posibles anomalías.

Seudonimización/diccionario de datos

» Determinar las variables relevantes para el modelo, identificando las variables asociadas a categorías especiales de datos y las variables proxy.

» Determinar y aplicar los criterios de minimización de los datos aplicables a las diferentes etapas del componente IA usando estrategias de ocultación, separación, abstracción, anonimización y seudonimización de los datos.

» Asociar un diccionario de datos a las bases de datos que permita su análisis y comprensión.

» Análisis del grado de anonimización de los datos y del posible riesgo de reidentificación.

3.2.1.3. SELECCIÓN DEL MODELO DE IA

Una vez que tenemos los objetivos que tenemos que conseguir e identificados los datos que vamos a utilizar para el modelo que vamos a desarrollar llega el momento elegir el modelo de Inteligencia Artificial que vamos a utilizar. ¿Qué debemos tener en cuenta a la hora de elegir el modelo que vamos a utilizar?

- Elección de la medida de éxito del modelo. El indicador que determinará que vamos por buen camino debe ser observable y debe estar alineado con los objetivos de negocio que queremos conseguir. La elección de esta medida dependerá de la naturaleza del problema que vamos a abordar, los objetivos, el contexto en el que nos encontramos y el impacto comercial de las decisiones basadas en el modelo

- Selección de algoritmos y técnicas que se ajusten al problema al que nos estamos enfrentando. Podemos encontrar diferentes técnicas como, por ejemplo:
    * Redes Neuronales.
    * Árboles de decisión.
    * SVM (Support Vector Machines).
    * Métodos basados en reglas.

A la hora de elegir el que vamos a utilizar hay que tener en cuenta las ventajas, desventajas del algoritmo o técnica elegida, así como los distintos casos de uso que se adapten mejor a nuestras necesidades.

En cuanto a la disponibilidad de bibliotecas y marcos de trabajo para el desarrollo del modelo, se deben valorar los que se encuentran disponibles en ese momento que mejor se ajusten al modelo que queremos desarrollar. Estas bibliotecas disponen de una amplia gama herramientas y utilidades para que se desarrolle y mantenga correctamente el modelo de IA. A modo de ejemplo, podemos encontrar bibliotecas de código abierto tales como: TensorFlow, PyTorch, scikit-learn , Keras y Café.

RIESGO NORMATIVO - IDENTIFICACIÓN Y TRANSPARENCIA DEL COMPONENTE

A continuación, se enumeran recomendaciones para mitigar el riesgo normativo derivado de una incorrecta identificación y transparencia del componente de IA:

Inventario de componentes IA

» El componente IA está identificado en la documentación con un nombre o código, la identificación de la versión y la fecha de creación.

» Se dispondrá de una firma digital de todo el conjunto que garantice su integridad.

» Histórico de versiones de la evolución del componente IA utilizado, incluyendo los parámetros usados en el entrenamiento del componente.

Identificación de responsabilidades

» Registrar los datos identificativos y de contacto de la/s persona/s o institución/instituciones responsables de las etapas del ciclo de vida del componente IA bajo auditoría y/o, corresponsables, representantes del responsable y de los encargados.

» Especificar en el contrato asociado a las etapas de tratamiento bajo auditoría del reparto de responsabilidades desde el punto de vista de protección de datos personales.

» Inscripción en el Registro de Actividades de Tratamiento de los responsables respectivos, y/o los encargados, del tratamiento de datos personales bajo auditoría.

» Determinar si existe un delegado de Protección de Datos, y en caso afirmativo, obtener la identificación de este y comunicación ante la Autoridad de Control.

Transparencia

» Documentar el origen de los datos y el mecanismo para informar.

» Identificar y justificar las características de los datos usados para entrenar al componente IA.

» Elección del modelo más adecuado teniendo en cuenta criterios de eficiencia, calidad y precisión del componente IA usando criterios de simplicidad e inteligibilidad.

» Poner a disposición de las partes interesadas la información sobre los metadatos del componente IA, su lógica y las consecuencias que pueden derivarse de su empleo.

» Se dispone de la documentación suficiente para comprender la lógica del componente IA utilizado y realizar la trazabilidad de su comportamiento respecto a cada conjunto de datos de entrada.

» Prever mecanismos para minimizar los perjuicios ante un comportamiento erróneo del componente IA.

RIESGO NORMATIVO - PROPÓSITO DEL COMPONENTE IA

Identificar los fines y usos

» Documentar el objetivo que se persigue con el uso del componente IA, tanto en términos cuantitativos como cualitativos, con una descripción clara de lo que se pretende conseguir mediante su empleo en el marco del tratamiento.

» Identificar la relación entre el objetivo que se persigue con el uso del componente de IA y las condiciones que garantizan la licitud de dicho tratamiento.

» Identificar las dinámicas, actividades y/o procesos en el marco de la organización en la que se integra la etapa del ciclo de vida del componente IA bajo auditoría, delimitando, en la medida de lo posible, el contexto de su uso.

» Categorizar los usuarios potenciales del componente IA.

» Describir otros posibles usos y usuarios secundarios junto con la base de legitimación que justifica su utilización.

Identificación de contexto

» Documentar los contextos jurídico, social, económico, organizacional, técnico, científico, o de cualquier otra clase que esté relacionado con la inclusión del componente IA

» Definir la estructura organizacional y/o contractual entre las partes y así el reparto de tareas y responsabilidades.

» Describir los factores condicionantes de la efectividad del componente.

» Definir los requisitos de los operadores humanos que tendrán misión de supervisar e interpretar la operación del componente IA.

» Documentar la interacción del componente IA con otros componentes, sistemas o aplicaciones, propias o de terceros y el reparto de responsabilidades de mantenimiento, actualización y minimización de los problemas de privacidad del sistema.

» Definir los baremos o umbrales para interpretar y utilizar los resultados ofrecidos por el componente IA utilizado.

» Identificar aquellos contextos para los que no está recomendado incluir el componente IA en un tratamiento al no poder cumplir con su objeto o propiedades, o un nivel de fiabilidad y/o exactitud inadecuada respecto a la relevancia que podría tener el tratamiento en el interesado.

Análisis de proporcionalidad y necesidad

» Evaluar el empleo del componente IA en el marco del tratamiento frente a otras posibles opciones con relación a los derechos y libertades de los interesados.

» Si se trata de un nuevo desarrollo, realizar un análisis comparado de la eficacia y la adecuación de los resultados del componente IA frente a otros componentes más probados.

» Justificar las motivaciones y argumentos que conducen a abordar este problema a través del empleo de un componente IA.

» Si se trata de un problema conocido, documentar y justificar los motivos que han conducido a un cambio en el esquema de funcionamiento anterior, describiendo, en su caso, los nuevos objetivos que se persiguen mediante el empleo del componente IA.

» Análisis y gestión del riesgo para los derechos y libertades de los interesados que introduce en el tratamiento el procesamiento de los datos mediante el componente IA.

3.2.1.4. ENTRENAMIENTO DEL MODELO

Una vez que elegido el modelo de Inteligencia Artificial a utilizar se inicia su entrenamiento. En esta fase se usan los datos de entrenamiento para ajustar los parámetros del modelo para que el modelo aprenda a realizar predicciones o a tomar decisiones.

Se detallan los pasos que componen un entrenamiento de un modelo:

a. **Preparación de los datos de entrenamiento de forma adecuada:** Deberemos limpiar y normalizar los datos, codificar las variables que necesitemos, elegir las características más relevantes y etiquetarlas.

b. **Definir la arquitectura y diseño del modelo:** En este punto, es necesario incluir el número y la configuración de las capas, la elección de funciones de activación, la selección de optimizadores y la definición de hiperparámetros. Como ejemplos de hiperparámetros podemos encontrar.

- Tasa de Aprendizaje – Cuanto de lento o rápido aprende el modelo.
- Tamaño del lote – Número de ejemplos de entrenamiento que se utilizan.
- Funciones de activación - Son funciones matemáticas que se aplican a la salida de cada neurona en una red neuronal.

c. **Inicializar los parámetros del modelo:** Deberemos asignar los pesos y sesgos iniciales de forma aleatoria o utilizando valores preentrenados que hayamos utilizado en modelos previos.

d. **Alimentar los datos de entrenamiento al modelo:** El modelo realiza cálculos basados en los datos de entrada y genera predicciones o resultados. Estos resultados se comparan con las etiquetas o resultados conocidos para medir la diferencia entre las predicciones del modelo y los valores reales.

e. **Cálculo de función de pérdida:** Esta función nos permite ver lo bien que se están haciendo las predicciones comparadas con los valores reales. En general, las funciones de pérdida pueden clasificarse como:

- Error lineal o local - Se calcula obteniendo la diferencia entre el valor real que hay que predecir y el valor predicho.
- Error medio cuadrático (MSE) o error global - Esta función nos permitirá conocer de manera global el porcentaje de error cometido

f. **Optimizar los parámetros del modelo:** El objetivo del entrenamiento es ajustar los parámetros del modelo para minimizar la función de pérdida. Utiliza algoritmos de optimización, como el descenso del gradiente, permite que se puedan ajustar los pesos y sesgos del modelo y mejorar gradualmente su rendimiento.

g. **Iterar el proceso de entrenamiento:** Repite los pasos anteriores para varias iteraciones ya que de esta forma el modelo aprende a partir de los datos y mejora su rendimiento. Esto implica alimentar el conjunto de entrenamiento al modelo repetidamente y ajustar los parámetros en función de las retroalimentaciones proporcionadas por la función de pérdida.

h. **Validación cruzada y ajuste de hiperparámetros:** En la actualidad podemos dos técnicas principales validación cruzada:

- **Técnica del Train-Test Split:** Esta técnica descompone de manera aleatoria una serie de datos y se divide en dos partes. La primera, que es entre el 70% y 80% de los datos de la serie, sirve para el entrenamiento del modelo de aprendizaje automático y la segunda, entre el 20% y 30% de los datos restantes, te permite probarla para la validación.
- **Método K-Folds:** Esta técnica tiene un parámetro único llamado "K", que hace referencia al número de grupos en el que se dividirá la muestra. El valor K no debe ser ni demasiado bajo ni demasiado alto, un valor más alto lleva a un modelo con menos sesgo, pero una varianza demasiado amplia puede llevar a un ajuste excesivo. Sin embargo, un valor más bajo es prácticamente lo mismo que utilizar el método Train-Test Split.

El entrenamiento se considera completo cuando se alcanza un criterio de éxito preestablecido.

RIESGO NORMATIVO – FUNDAMENTOS DEL COMPONENTE I

Identificar la política de Desarrollo del componente

» Establecimiento de las políticas de desarrollo de productos y sistemas.

» Diseño de un proceso de revisión y un control de versiones de las políticas.

Adecuación de los modelos teóricos base

- Estudio y análisis sobre el marco teórico y experiencias previas similares sobre las que se fundamenta el desarrollo del componente IA.
- Están determinadas, argumentadas y documentadas las ideas de base e hipótesis que se toman en consideración para la creación y desarrollo del modelo.
- Procedimiento de revisión crítica y contrastada de los razonamientos derivados de la aceptación de hipótesis importantes para el desarrollo del componente IA.
- Análisis cuidadoso de cara a establecer presunciones adecuadas sobre las posibles variables proxy que intervienen en el componente IA.

Adecuación del marco metodológico

- Documentar el marco metodológico de definición del modelo y creación del componente IA.
- Determinar, en función del análisis del problema a resolver y de manera justificada, el modelo de desarrollo a utilizar.
- Seleccionar y definir las métricas con respecto de las cuales medir el comportamiento del componente IA.
- Tener un procedimiento de registro y seguimiento de las desviaciones en el comportamiento del componente IA respecto de las métricas definidas que permite realizar una monitorización.

Identificación de la arquitectura básica del componente

- Fase de Análisis: Incluir, como parte del catálogo de requisitos, aquellos específicos para garantizar la privacidad y la protección de los datos personales.
- Fase de Programación: Seguir y documentar los principios, códigos y buenas prácticas de codificación utilizados para garantizar que el código sea legible, seguro, fácil de mantener y robusto.
- Identificar la arquitectura básica del componente IA, incluyendo información sobre la técnica de aprendizaje automático utilizada, el tipo o tipos de algoritmos probados.
- Tener un procedimiento sistemático de documentación de la implementación del componente que garantiza el registro y posterior adquisición de toda la información necesaria para identificarlo.

Si no se tiene posibilidad de acceder al código del componente IA, evaluar aplicar un proceso de ingeniería inversa u otro método alternativo como el uso de pruebas de conocimiento cero.

3.2.1.5. EVALUACIÓN DEL MODELO

La evaluación de un modelo entrenado es un paso crítico para determinar su rendimiento y su capacidad para realizar predicciones precisas con datos que no se han utilizado previamente.

Para medir el éxito de un modelo entrenado hay deben elegirse las métricas más adecuadas para el modelo que se ha estado utilizando

Como ejemplos podemos encontrar:

- Matriz de Confusión: Es una representación matricial de los resultados de las predicciones de cualquier prueba binaria que se utiliza a menudo para describir el rendimiento del modelo de clasificación sobre un conjunto de datos de prueba cuyos valores reales se conocen.
- Exactitud (Accuracy): Mide la proporción de predicciones correctas realizadas por el modelo sobre el total de predicciones.
- Precisión (Precision): Calcula la proporción de verdaderos positivos sobre la suma de verdaderos positivos y falsos positivos. Es útil cuando se busca minimizar los falsos positivos
- Sensibilidad o tasa de verdaderos positivos (Recall o True Positive Rate): Representa la proporción de verdaderos positivos sobre la suma de verdaderos positivos y falsos negativos. Es útil cuando se busca minimizar los falsos negativos.
- Especificidad (Specificity): Calcula la proporción de verdaderos negativos sobre la suma de verdaderos negativos y falsos positivos. Es útil cuando se busca minimizar los falsos positivos.
- Valor F1 (F1-score): Es una medida de la precisión y la sensibilidad del modelo y se calcula como la media armónica de ambos valores.
- Pérdida (Loss): Es una medida de qué tan lejos están las predicciones del modelo de los valores reales. Se busca minimizar esta métrica.
- Curva de características operativas del receptor (ROC): Al trazar la tasa positiva verdadera (sensibilidad) frente a la tasa de falsos positivos (1 - especificidad), obtenemos la curva de Característica Operativa del Receptor (ROC). Esta curva nos permite visualizar el equilibrio entre la tasa de verdaderos positivos y la tasa falsos positivos
- Pérdida Logarítmica: Mide el desempeño de un modelo de clasificación en el que la entrada de la predicción es un valor de probabilidad entre 0 y 1. La pérdida logarítmica aumenta a medida que la probabilidad predicha se aleja de la etiqueta real. El objetivo de cualquier modelo de aprendizaje automático es minimizar este

valor. Por lo tanto, una pérdida logarítmica menor es mejor, con un modelo perfecto teniendo una pérdida logarítmica de 0.

» Índice Jaccard o coeficiente de similitud Jaccard: Es una estadística utilizada para comprender las similitudes entre los conjuntos de muestras. La medición enfatiza la similitud entre conjuntos de muestras finitas y se define formalmente como el tamaño de la intersección dividido por el tamaño de la unión de los dos conjuntos etiquetados.

» Gráfico de Kolmogorov Smirnov: Es una medida del grado de separación entre las distribuciones positivas y negativas.

» Gráfico de ganancia y elevación: Es una medida de la eficacia de un modelo de clasificación calculado como la relación entre los resultados obtenidos con y sin el modelo. Los gráficos de ganancia y elevación son ayudas visuales para evaluar el rendimiento de los modelos de clasificación. Sin embargo, en contraste con la matriz de confusión que evalúa los modelos en toda la población, el gráfico de ganancia o elevación evalúa el rendimiento del modelo en una porción de la población. Cuanto mayor sea la elevación (es decir, cuanto más lejos esté de la línea de base), mejor será el modelo.

» Coeficiente de Gini: Es una métrica popular para los valores de clase desequilibrados. El coeficiente oscila entre 0 y 1, donde 0 representa la igualdad perfecta y 1 la desigualdad perfecta. Aquí, si el valor de un índice es mayor, entonces los datos estarán más dispersos.

Comprender lo bien que un modelo de aprendizaje automático va a funcionar con datos no vistos es el propósito final de trabajar con estas métricas de evaluación. Métricas como la exactitud, la precisión, la exhaustividad son buenas formas de evaluar los modelos de clasificación para conjuntos de datos equilibrados, pero si los datos están desequilibrados y hay una disparidad de clases, entonces otros métodos como el ROC/AUC o el coeficiente de Gini funciona mejor en la evaluación del rendimiento del modelo.

CONTROLES DE VALIDACIÓN DEL ALGORITMO

Verificación y validación

» Documentar el proceso de verificación y validación, las técnicas empleadas, el conjunto de pruebas y comprobaciones realizadas, los resultados obtenidos y las acciones propuestas.

» Establecer una guía o estándar para realizar un procedimiento sistemático de verificación y validación del componente IA.

» Implementar mecanismos de control y supervisión necesarios para garantizar el cumplimiento de los objetivos con eficacia.

» Definir las métricas y criterios respecto a los cuales se realizará las comprobaciones en el proceso de verificación y validación.

» Definir la estrategia para evaluar la corrección del componente IA.

» Asegurar la cualificación del personal involucrado en las tareas de verificación y validación.

Verificación y Validación del Componente IA, debe tomarse en cuenta, como mínimo:

» Revisión e inspección para detectar y corregir, de manera temprana, defectos en los requisitos, defectos de diseño, especificaciones incorrectas o desviaciones en el desarrollo respecto de los criterios aplicables

» Incluir en el plan las pruebas del tipo:

- Caja blanca a nivel del diseño de la red o del componente IA.
- Caja blanca a nivel de implementación y código.
- Caja negra necesarias para comprobar que la funcionalidad del componente IA garantizada, su comportamiento es el esperado y que la integridad de la información utilizada se mantiene.

» Incluir las pruebas necesarias para testear la seguridad, con relación a la protección de los derechos y libertades si el componente IA se implementa en sistemas robóticos, industria 4.0, o de internet de las cosas.

» Incluir la comprobación de valores límite y casos de prueba extremos que pueden llevar al componente a funcionar de una manera no esperada.

» Incluir en la validación un proceso de depuración documentado para corregir los errores, carencias o inconsistencias detectadas.

Rendimiento

» Establecimiento de métricas o conjunto de métricas agregadas para determinar su precisión, exactitud, sensibilidad u otro parámetro de rendimiento relativo a la aplicación del principio de exactitud de los datos.

» Análisis e interpretación de los valores de las tasas de falsos positivos y falsos negativos que arroja el componente IA para determinar la precisión, la especificidad y la sensibilidad del comportamiento del componente.

» Evaluar el nivel y definición de los parámetros de rendimiento que se requieren para el componente de IA

- Comparar los valores de rendimiento entre distintas opciones de componentes IA en el marco de un proceso de elección del componente más adecuado para un tratamiento.
- Definir y determinar las variables de salida prestando especial atención a aquellas que constituyen categorías especiales de datos.
- Adoptar las medidas necesarias para garantizar que los datos utilizados son exhaustivos y están actualizados.
- Determinar los parámetros y sus valores de corte para que el modelo tome en cuenta determinadas variables de cara a obtener resultados que sean significativos.
- Procedimientos para detectar si la respuesta del componente IA a los datos de entrada es errónea o supera un umbral de error determinado.
- Realizar el reajuste de la dimensionalidad del modelo para que exista un equilibrio entre la complejidad y la capacidad de generalización.

Coherencia

- Procedimiento para verificar y actuar si se producen variaciones significativas en los resultados obtenidos respecto de las salidas esperadas.
- Establecer un umbral de cara a determinar cuándo un resultado obtenido difiere del esperado ante datos de entrada idénticos o similares (variaciones significativas).
- Determinar si el componente IA se comporta de manera distinta frente a individuos que se diferencian entre sí en características asociadas a categorías especiales de datos o en los valores que toman las variables proxy.
- Analizar los efectos en los resultados de salida del componente IA ante variaciones de las variables con baja prevalencia en el conjunto de datos de entrenamiento.
- Adoptar medidas para garantizar la independencia del componente.
- Verificar que no existe correlación entre los resultados y las variables adicionales asociadas a sujetos que no forman parte de las variables de proceso y que pudieran determinar la existencia de sesgos.

Estabilidad y robustez

» Identificar los factores dentro del contexto de funcionamiento del componente, cuya variación puede afectar a las propiedades del componente IA.

» Evaluar el comportamiento del componente IA ante casos de uso o entornos imprevistos.

» Estimar los tiempos en los que es necesario una reevaluación, reajuste o reinicio del componente para ajustarlo a desviaciones en los datos de entrada o cambios en los criterios de toma de decisiones.

» Documentar si el componente IA se ha construido siguiendo un enfoque estático o bien un enfoque dinámico o de aprendizaje continuo. Si es continuo, se deberá evaluar el grado de adaptabilidad a nuevos datos o tipos de datos de entrada y definirse los procedimientos y mecanismos de supervisión para verificar que las conclusiones extraídas siguen siendo válidas.

Trazabilidad

» Generar un sistema de control de versiones de todos los elementos del componente IA: conjuntos de datos utilizados, del código del componente, librerías empleadas y de cualquier otro elemento asociado al componente.

» Generar un procedimiento formal sujeto a una reevaluación del riesgo en función de aquellos cambios que puedan producirse en la implementación del componente IA a lo largo de su ciclo de vida.

» Establecer mecanismos de monitorización y supervisión del componente IA, tales como ficheros de logs y registros de resultados.

» Registrar aquellas incidencias y comportamientos anómalos previos detectados y corregidos.

» Mecanismos de monitorización disponibles para operadores humanos para su seguimiento y verificación.

» Fijar un procedimiento para asegurar la intervención humana en la toma de decisiones, tanto de oficio, ante resultados discrepantes en relación con el comportamiento esperado,

» Adoptar mecanismos para que los resultados y las decisiones tomadas puedan llegar a depender, de manera exclusiva, de la responsabilidad de seres humanos.

» Registrar los eventos que tengan valor para la ciberseguridad.

Seguridad

- » Análisis de los riesgos para los derechos y libertades de las personas a la luz del cual se puedan determinar los requisitos de seguridad y privacidad del componente IA utilizado en el marco del tratamiento.
- » Definir en el origen los requisitos relacionados con la protección de los datos y la seguridad
- » Seguir los estándares y buenas prácticas disponibles para el desarrollo y la configuración segura del componente.
- » Implementación de las medidas necesarias para garantizar la protección de los datos tratados, en particular, las orientadas a garantizar la confidencialidad, y la integridad para proteger la implementación del componente de manipulaciones accidentales o intencionadas.
- » Implementar medidas para garantizar la resiliencia del componente y su capacidad para resistir ataques.
- » Procedimentar la monitorización del funcionamiento del componente y detectar, de manera temprana, posibles fugas de datos, accesos no autorizados u otros tipos de brechas de seguridad.
- » Usuarios, operadores y administradores tendrán la información necesaria y conocen sus deberes y responsabilidades en materia de seguridad.
- » Registrar entre los eventos de seguridad las actuaciones sobre la configuración, la enseñanza, el uso y cualquier otra información relevante o que pueda llegar a serla, para la ciberseguridad.

3.2.1.6. AJUSTE Y MONITORIZACIÓN DEL MODELO

Para realizar un ajuste óptimo en un modelo de Inteligencia Artificial podemos recurrir a las siguientes acciones:

- » Usar más información para el entrenamiento: El uso de un gran conjunto de datos 'nuevos' generalmente ayuda al modelo a elegir la señal de manera eficiente, sin embargo, esta técnica puede no funcionar todas las veces. Si agregamos muchos datos ruidosos y los datos relevantes son escasos, incluso tener una gran cantidad de datos totales no ayudará al modelo a predecir con precisión los valores.

» Técnica de validación cruzada: La validación cruzada es un estándar de oro aplicado para estimar la precisión del modelo en datos no vistos. Si tienes los datos, usar un conjunto de datos de validación también es una práctica excelente. Una forma estándar de encontrar un error de predicción fuera de muestra es usar una validación cruzada de 5 veces.

» Detección temprana: Cuando entrena iterativamente un modelo hasta cierto número de iteraciones, el rendimiento del modelo mejora. Después el modelo se sobrecarga y tendrá un rendimiento bajo en los conjuntos de datos de prueba. Por lo tanto, debes detener las iteraciones de entrenamiento del modelo antes de que exista un ajuste excesivo en el modelo.

» Regularización – La regularización se hace para simplificar el modelo de IA y consiste en muchos métodos. La técnica de regularización utilizada depende del tipo de modelo, por ejemplo, si el modelo es un árbol de decisión, la regularización podría ser podar el árbol; si el modelo es una regresión puede agregar una penalización a la función de costo para la regularización.

Además del ajuste, es necesario la evaluación periódica de la validez del modelo, así como el mantenimiento de los procesos de control establecidos.

MEDIDAS DE SEGURIDAD PARA LA PROTECCIÓN DEL ENTORNO

Desarrollo de técnicas de pseudoanonimización y protección

» Colaborar con las áreas técnicas para que sea posible la incorporación de más información al modelo, manteniendo garantías suficientes de protección de la información sensible.

» Colaborar en el desarrollo de infraestructuras de tratamiento de datos seguros que permitirán la realización de validaciones cruzadas sin sobrecostes.

4 LA IA COMO HERRAMIENTA ADVERSA

4.1. PRINCIPALES USOS MALICIOSOS DE LA IA

En este apartado se identifican tipos de ataques en los que los ciberdelincuentes utilizan herramientas de IA, detallando el uso que realizan de la misma para así poder diseñar contramedidas en nuestros procesos de protección y detección.

4.1.1. INGENIERÍA SOCIAL

TIPOS HABITUALES DE ATAQUE:

- **Phishing:** El envío masivo de e-mails simulando ser un negocio u organización legítima y solicitando que se abra un archivo infectado o que se visite una página web.
- **Smishing:** El phishing se realiza con mensaje de texto.
- **Vishing:** El fraude se produce a través de una llamada telefónica.

LIMITACIÓN ACTUAL DE LAS TÉCNICAS:

- Algunos ataques de phishing actuales son reconocibles por parte de los usuarios porque no utilizan un lenguaje "normal" o incluso pueden contener faltas de ortografía. El uso de estas herramientas va a permitir que cualquier ciberdelincuente desde cualquier parte del mundo, pueda preparar mensajes mucho más creíbles, incluso sin conocer el idioma de la víctima.

VALOR PROPORCIONADO POR IA:

» La capacidad de la IA para analizar y procesar grandes volúmenes de datos podrá utilizarse para mejorar los ataques de ingeniería social. Al analizar perfiles de redes sociales y otras fuentes de información pública, un atacante podría obtener información personal detallada sobre un individuo objetivo y utilizarla para diseñar un ataque más personalizado y mejor dirigido, aumentando las posibilidades de éxito.

» Podrán recopilar más información del entorno del objetivo, y customizar el mensaje haciendo que los ataques de phishing sean más creíbles y que aumente el número de personas que pican.

» Las herramientas basadas en Inteligencia Artificial generativa van a suponer un avance para los equipos de ciberdelincuentes que realizan estos ataques de ingeniería social ya que tienen una gran capacidad para generar información en un lenguaje que se adapta muy bien al entorno que se le pida.

4.1.2. GENERACIÓN DE RANSOMWARE ASISTIDO POR LA IA

LIMITACIÓN ACTUAL DE LAS TÉCNICAS:

- Dificultad y necesidad de personal especializado para el desarrollo de técnicas de evasión.

VALOR PROPORCIONADO POR IA:

» La IA puede ser utilizada tanto para desarrollar programas de software legítimos como para desarrollar malware de distintos tipos lo cual implica una nueva y peligrosa generación de ciberamenazas.

» Los ciberdelincuentes pueden entrenar a su propia IA para evitar técnicas de detección de comportamientos maliciosos en el código generado, lanzando ataques y probando su detección para mejorarlos de forma continua y hacerlos cada vez más eficaces e indetectables.

» A través de IA, mediante por ejemplo un chatbot como ChatGPT, un ciberdelincuente puede crear capacidades básicas de búsqueda y cifrado de archivos en base al algoritmo de cifrado escogido con el objetivo de utilizarlo para desarrollar ransomware contra un objetivo.

PRINCIPALES TÉCNICAS QUE APROVECHAN LAS MEJORAS DEL IA:

- La generación de malware polimórfico puede llegar a ser indetectable ya que se producen mutaciones de código dinámicamente por lo que es mucho más difícil de identificar con herramientas actuales.
- Este malware se caracteriza porque no muestre un comportamiento malicioso mientras esté almacenado en el disco y que no contenga una lógica sospechosa mientras esté almacenado en memoria. Esta característica lo hace altamente indetectable para las soluciones de seguridad que usan detección basada en firmas pudiendo pasar desapercibido.

4.1.3. DENEGACIÓN DE SERVICIO (DDOS) IMPULSADOS POR IA

LIMITACIÓN ACTUAL DE LAS TÉCNICAS:

- Dificultad y necesidad de personal especializado para el desarrollo de técnicas de ataque.

VALOR PROPORCIONADO POR IA:

» Como parte de los ataques, se suelen utilizar servidores Command and Control que controlan una serie de bots comprometidos (una Botnet, es una red de equipos infectados con software malicioso controlados remotamente y utilizados para el envío de tráfico como parte del ataque).

» Utilizando sistemas de machine learning, las máquinas atacantes pueden aprender buscando y probando los mejores tipos de ataque, cambiando los vectores de ataque aleatoriamente para mejorarlos y hacerlos eficaces. Así, la máquina atacante es capaz de identificar patrones para crear un algoritmo con los datos que debe tratar y luego hacer predicciones.

PRINCIPALES TÉCNICAS QUE APROVECHAN LAS MEJORAS DEL IA:

- Evolución de técnicas de ataque avanzado y mayor facilidad en la coordinación de bots de manera global por parte de los atacantes.

4.1.4. ATAQUES DE FUERZA BRUTA

LIMITACIÓN ACTUAL DE LAS TÉCNICAS:

- Tiempo necesario para la comprobación y riesgo de identificación al utilizar patrones de intento conocidos/fácilmente detectables.

VALOR PROPORCIONADO POR IA:

» En el ataque por fuerza bruta basado en inteligencia artificial un atacante utiliza técnicas de aprendizaje automático para automatizar el proceso de adivinar contraseñas/claves. Para ello recopilará y analizará grandes cantidades de datos que se utilizarán para entrenar al algoritmo de aprendizaje automático y luego este algoritmo generará las contraseñas/claves con más probabilidad de que sean las correctas.

» El uso de inteligencia artificial aumenta la eficacia de este ataque frente al uso de técnicas convencionales porque es más rápido y tiene mayor precisión, ya que el método consiste en generar contraseñas/claves con más probabilidad de que sean correctas. Llegar a adivinar credenciales/claves puede llevar mucho tiempo en algunos casos, por lo que el uso de inteligencia artificial para agilizar esta tarea es una ventaja para el atacante.

» Frente a estas técnicas clásicas, la IA utiliza otros métodos: las contraseñas generadas, como se ha dicho, parten de recopilar información de la cual aprende el algoritmo para predecir qué contraseñas son más probables de ser utilizadas y probar estas contraseñas primero. De esta forma se maximizan las probabilidades de acierto y se tardaría menos tiempo en que se materializase el acceso.

PRINCIPALES TÉCNICAS QUE APROVECHAN LAS MEJORAS DEL IA (TODAS LAS DE FUERZA BRUTA):

- Ataque simple de fuerza bruta: prueba de contraseñas comunes o haciendo uso de información accesible online (como redes sociales).
- Pulverización de contraseñas donde se prueba una contraseña en muchas cuentas diferentes y así puede que las políticas de bloqueo no adviertan el ataque.
- Ataque de diccionario en la cual para un usuario seleccionado se eligen contraseñas desde un diccionario

(ampliando con caracteres especiales) o los ataques de fuerza bruta inversa, donde desde una contraseña existente se prueba a averiguar el usuario.

* Nota: en este contexto entendemos como ataque de fuerza bruta el intento de averiguar credenciales (usuario/contraseña) aplicando el método de prueba y error. También se aplica al caso de claves de cifrado en criptografía, donde el ataque por fuerza bruta consiste en ir probando todas las posibles combinaciones

4.1.5. DENEGACIÓN DE SERVICIO (DDOS) IMPULSADOS POR IA

Por ataques asistidos podemos entender p.ej. la creación de scripts enfocados a mercados de la Dark Web o el desarrollo de Amenazas persistentes avanzadas (APT).

LIMITACIÓN ACTUAL DE LAS TÉCNICAS:

- Dificultad y necesidad de personal especializado para la identificación de información relevante sobre la víctima.

VALOR PROPORCIONADO POR IA:

» Las amenazas avanzadas persistentes (APT por sus siglas en inglés) hacen referencia a todos aquellos ataques dirigidos contra un objetivo determinado, y que se lanzan de manera continuado en el tiempo y meticulosa. El que este tipo de ataques complejos se pueda dilatar en el tiempo implica que los atacantes avancen lentamente en las fases del ataque, no alcanzando los umbrales que podrían hacer saltar una alerta por parte de la víctima. Adicionalmente, el que se desarrollen de una forma tan minuciosa, implica que el atacante recopila y estudia en detalle la información del objetivo, obteniendo una visión completa de los sistemas y procedimientos de la víctima.

» En este tipo de ataques la Inteligencia artificial desarrolla un papel clave tanto en el análisis de la información de la víctima, como en la generación de artefactos que faciliten el ataque en sí.

PRINCIPALES TÉCNICAS QUE APROVECHAN LAS MEJORAS DEL IA:

- Mejora de las capacidades de procesamiento de información y agilidad en todas las técnicas utilizadas.

4.1.6. EXPLOTACIÓN MASIVA DE VULNERABILIDADES

LIMITACIÓN ACTUAL DE LAS TÉCNICAS:

- Dificultad y necesidad de personal especializado.

VALOR PROPORCIONADO POR IA:

» La Inteligencia Artificial facilita enormemente la generación de código de forma rápida y efectiva, ahorrando horas de trabajo de los programadores. Estas facilidades pueden ser aprovechadas por los cibercriminales para lanzar o explotar campañas dirigidas contra un objetivo concreto de manera rápida, por ejemplo, tras el descubrimiento de una nueva vulnerabilidad sobre la cual no existe aún contramedida (denominada comúnmente como Zero-day). Este tipo de programas informáticos maliciosos (Malware) pueden ser luego compartidos en la Dark/Deep web donde los cibercriminales comparten libremente (o venden) información específica de los objetivos que han estado siguiendo.

PRINCIPALES TÉCNICAS QUE APROVECHAN LAS MEJORAS DEL IA (TODAS LAS DE FUERZA BRUTA):

- Mejora de las capacidades de procesamiento de información y agilidad en todas las técnicas utilizadas.

4.1.7. ATAQUES DE RECONOCIMIENTO FACIAL Y BIOMETRÍA

Si bien este tipo de ataque encaja en el capítulo de "Ataque de Ingeniería Social", debido a su alto impacto por la irrupción de los sistemas de IA generativa, se trata de forma específica.

VALOR PROPORCIONADO POR IA:

» Los ataques realizados mediante inteligencia artificial sobre sistemas de reconocimiento facial y biometría aprovechan las capacidades de la IA para engañar o manipular estos sistemas. Estos ataques se basan en la generación de imágenes, voz o vídeos falsos que engañan a los algoritmos de reconocimiento facial y biometría, haciendo que identifiquen erróneamente a personas o permitan el acceso no autorizado a sistemas protegidos.

» La técnica conocida como "deepfake" es una de las formas más destacadas de ataque utilizando IA en el ámbito de la manipulación de imágenes y vídeos. Los deepfakes consisten en la creación de contenido multimedia sintético utilizando algoritmos de aprendizaje profundo. Estos algoritmos analizan y aprenden los patrones y características de un conjunto de datos, y luego pueden generar imágenes o vídeos realistas que parecen auténticos, pero son completamente falsos.

» Los deepfakes se han utilizado para manipular imágenes y vídeos en los que se superponen los rostros de personas reales en situaciones falsas. Esto ha planteado preocupaciones significativas en cuanto a la seguridad y la confianza en los sistemas de reconocimiento facial y biometría. Los atacantes pueden utilizar deepfakes para suplantar identidades, realizar fraudes o incluso difamar a personas inocentes al hacer que aparezcan en situaciones comprometedoras.

» Muchos de estos ataques tienen como objetivo la suplantación de identidades, no solo los ataques de deepfake, los ataques mediante la manipulación de características de imágenes o voz tienen como objetivo final engañar a sistemas biométricos y suplantar identidades concretas.

4.1.8. RECONOCIMIENTO Y ESCANEO AUTOMATIZADOS

LIMITACIÓN ACTUAL DE LAS TÉCNICAS:

- Automatización y optimización de técnicas de reconocimiento y escaneo automático.

VALOR PROPORCIONADO POR IA:

» Las labores de reconocimiento y escaneo basados en Inteligencia artificial permiten a los atacantes automatizar las labores de búsqueda y beneficiarse de la potencia que ofrece la inteligencia artificial de cara a encontrar de una manera más fácil y rápida los activos y las vulnerabilidades asociados a los mismos, permitiendo automatizar la correlación de las vulnerabilidades.

» Adicionalmente, los atacantes pueden utilizar la Inteligencia Artificial como capa intermedia durante el escaneo, facilitando su anonimización y dificultando que sean descubiertos por la víctima.

PRINCIPALES TÉCNICAS QUE APROVECHAN LAS MEJORAS DEL IA (TODAS LAS DE FUERZA BRUTA):

- **Escaneo activo:** Utilización de herramientas automáticas para el escaneo de la infraestructura de la víctima, de cara a enumerar y encontrar vulnerabilidades en la misma. La inteligencia artificial facilita estos escaneos de manera que sean más efectivos.

* **Análisis de tráfico:** Los atacantes analizan el tráfico generado por la víctima en busca de información. Como ejemplo, un simple mail enviado al atacante ofrece determinada información sobre Infraestructura, direcciones IP, taxonomía de las cuentas de correo de los empleados, la herramienta de filtrado de correo que se utiliza… La inteligencia Artificial facilita la extracción de este tipo de información.

* **Acceso a la infraestructura:** Los atacantes podrían intentar conectarse a la Infraestructura de la víctima de cara a enumerar los activos que la misma, y localizar las vulnerabilidades desde la red interna, obteniendo así una mayor visibilidad de los activos.

4.2. RECOMENDACIONES PARA PREVENIR EL USO MALICIOSO DE LA IA

4.2.1. FORMACIÓN Y CONCIENCIACIÓN

La formación y concienciación es el primer paso en cualquier programa de mejora de ciberseguridad, en la IA también.

Es importante que se actualicen los programas de formación y concienciación y que en ellos se incorporen píldoras de formación en las que se establezcan ejemplos reales y concretos en los que el colaborador pueda ver cómo los atacantes han podido aprovechar las nuevas técnicas.

Se deben actualizar los materiales de formación y concienciación para hacer que los usuarios distingan si un mensaje recibido es legítimo o no, y que, sean conscientes de que, en caso de duda, es necesario que confirmen por un canal alternativo la autenticidad del mensaje.

Dentro de esta formación es importante incorporar "tips" concretos sobre el uso de técnicas de ingeniería social para engañar a víctima y ganarse su confianza:

- Suplantan el logo e imagen de la empresa a la que tratan de suplantar.
- Suplantan el teléfono de la entidad a la que suplantan al hacer Smishing o Vishing.
- Se tratan de ganar la confianza de la víctima dando información que tienen disponible para hacer el mensaje más creíble: por ejemplo, cuando saben el nombre del usuario empiezan la comunicación con dicho nombre.
- Utilizan las redes sociales (especialmente LinkedIn) con el fin de solicitar información sobre la organización.
- Incluso si no saben información como si la víctima es cliente o no de un banco, asumen que lo es y se dirigen a ella afirmándolo. De esta manera, el que no es cliente de la entidad, no hará caso del email, pero el que de la casualidad y lo sea, tendrá más posibilidades de picar.
- Utilizan cualquier dato de actualidad para hacer que el mensaje sea más creíble. Por ejemplo, en época de la declaración de la renta, envían emails suplantando a la agencia tributaria para simular que la víctima tiene derecho a un cobro y facilite sus datos bancarios.
- En ataques más dirigidos los atacantes llevan a cabo una investigación y recolección de datos del objetivo incluyendo redes sociales personales y laborales.

4.2.2. REFORZAR EL PROCESO DE IDENTIFICACIÓN Y AUTORIZACIÓN

Teniendo en cuenta que los métodos de adivinación son más eficientes cuando se usa la IA, la forma de protegerse basada en robustez de contraseñas debe mejorar a la convencional. Además, la actividad de los usuarios (uso de contraseñas similares en distintas cuentas) puede facilitar el conocimiento de esta.

Por lo tanto, es necesario rediseñar los procesos de identificación y autorización para que se basen en mecanismos de autenticación reforzada. Esto es especialmente importante en los aplicativos de negocio/transaccionales (donde además debe de ser relativamente sencillo utilizar nuevas técnicas de inteligencia artificial para definir patrones de uso) de modo que se puedan además limitar en lo posible los privilegios de los usuarios.

De manera general, se debe de considerar una obligación no negociable el uso de mecanismos de segundo factor de autenticación en toda conexión proveniente de redes no confiables.

En el caso de que sea necesario el uso de contraseñas "tradicionales", es importante acompañar el paso de cambio de contraseña de recomendaciones claras para la definición de contraseñas seguras (p.e. uso de primeras letras de una frase) y usar claves más largas, por ejemplo, de 15 caracteres (y por supuesto manteniendo las buenas prácticas de usar caracteres alfanuméricos, mayúsculas y minúsculas, no usar palabra de diccionario ni palabras personales (nombres de mascotas, años de nacimiento, etc.).

4.2.3. MEDIDAS TÉCNICAS DE PROTECCIÓN

Según hemos visto, es importante estar familiarizado con frameworks generales de protección (como NIST) y desarrollar un framework de cumplimiento de estos que incorporen una estrategia de defensa en profundidad. A continuación, se enumeran algunas de las principales medidas a tener en cuenta utilidad a la hora de mitigar los nuevos riesgos que surgen a partir de la IA.

- Mantener actualizados los equipos y programas es también clave a la hora de protegerse de cualquier vulnerabilidad que pueda ser explotada por parte de los atacantes. Para esto, es fundamental tener un plan de parcheado de equipos que pruebe y vaya instalando los parques de seguridad del fabricante -tanto de hardware como de software- en todos los dispositivos corporativos.
- Protección perimetral adecuada al tipo de servicios expuestos (AntiDDoS, WAF, CDN).
- La protección del correo electrónico para detectar mensajes de spear phishing que contienen archivos maliciosos o enlaces a sitios maliciosos.
- Asegurar un control adecuado de la navegación (tanto de usuarios como de servidores).
- Utilizar medidas de protección antimalware así como guías de bastionado seguro que incluyan la limitación de permisos de los usuarios, limitación de uso de powershell, etc.
- Limitar acceso a correo electrónico particular desde entorno corporativo (o protegerlo a través de soluciones de sandboxing adecuadas).

4.2.4. SERVICIOS DE INTELIGENCIA

En función de las capacidades de cada empresa (se debe de evaluar la contratación de proveedores especializados que nos permitan aprovechar sinergias con otros clientes) es de interés disponer de servicios de:

» **Descubrimiento y reconocimiento de activos:** Estos servicios nos permiten tener controlada la "foto" que publicamos, así como detectar vulnerabilidades expuestas en el perímetro.

» **Fuga de credenciales:** Detectar la presencia de credenciales/información técnica asociada a nuestro entorno en foros de inteligencia.

4.2.5. MEDIDAS ESPECÍFICAS PARA EL USO DE BIOMETRÍA

En respuesta a estos ataques, se están desarrollando técnicas de detección de deepfakes y se están implementando mejoras en los algoritmos de reconocimiento facial y biometría para hacerlos más resistentes a los ataques de inteligencia artificial. Estas son algunas de las técnicas desarrolladas:

» **Detección de anomalías:** Se pueden implementar algoritmos de detección de anomalías para identificar patrones inusuales en los datos de reconocimiento facial y biometría. Estos algoritmos pueden detectar ataques como deepfakes o manipulaciones de características y generar alertas cuando se detecte actividad sospechosa.

» **Verificación multimodal:** La verificación multimodal combina múltiples modalidades biométricas, como reconocimiento facial, de voz o de huellas dactilares, para aumentar la seguridad. Al requerir la verificación de múltiples modalidades, se dificulta la falsificación y se mejora la precisión del sistema.

» **Autenticación en tiempo real:** Se pueden implementar técnicas de autenticación en tiempo real que verifiquen la autenticidad de la persona en el momento del acceso. Esto puede incluir el uso de desafíos aleatorios, como solicitar a la persona que realice acciones específicas o responda a preguntas de seguridad en tiempo real.

» A través de la IA se puede además **recopilar información** sobre el contexto del ataque para así proporcionar medidas mitigantes más eficientes de una forma más rápida.

5 PRINCIPALES USOS DE LA IA EN LA PROTECCIÓN FRENTE ATAQUES

Como hemos visto en el capítulo anterior los ciberataques han aumentado su eficacia y complejidad gracias, en parte, a que los ciberdelincuentes ya utilizan técnicas de inteligencia artificial. Esto hace que los incidentes se vuelven cada día más complejos, difíciles de detectar y en muchos casos superen las capacidades de las soluciones tradicionales de seguridad. En este entorno se hace imprescindible utilizar la inteligencia Artificial para jugar con las mismas armas en la prevención, detección y respuesta a las diferentes amenazas.

Los principales usos de la IA se realizan en:

PREVENCIÓN	**A través del análisis de patrones de comportamiento, aprendiendo la línea base de comportamiento normal y detectando anomalías y desviaciones sospechosas.**
DETECCIÓN	**Analizando grandes volúmenes de datos en tiempo real e identificando tanto los ataques conocidos como los desconocidos con mayor rapidez**
RESPUESTA	**Realizando análisis entre diferentes fuentes y amenazas de manera rápida y eficiente que automaticen la respuesta y/o aporten datos relevantes a los analistas para eliminar la amenaza.**

Además, las Inteligencia artificial puede ayudar a correlacionar grandes cantidades de datos para identificar las fuentes de los ataques, los vectores utilizados y los datos comprometidos que son imprescindibles para realizar las actividades de análisis forense. Del análisis de los datos también se pueden extraer nuevos algoritmos y modelos para mantener los sistemas de defensa actualizados frente a las últimas tendencias en un ciclo de mejora continua.

Sin embargo, es importante destacar que la Inteligencia Artificial no es autosuficiente. Requiere del análisis de expertos en ciberseguridad y del uso de otras tecnologías de seguridad para una protección integral.

Se enumeran a continuación los principales beneficios que aporta la IA en la protección frente amenazas:

MAYOR PRECISIÓN Y VELOCIDAD EN LA DETECCIÓN DE AMENAZAS

» La inteligencia artificial (IA) ha demostrado ser una herramienta invaluable en la ciberseguridad al proporcionar una detección más precisa y rápida de amenazas. A través de algoritmos avanzados y técnicas de aprendizaje automático, la IA puede analizar grandes volúmenes de datos en tiempo real.

» Esto permite identificar patrones y comportamientos anómalos en la red y los sistemas que podrían ser indicativos de una actividad maliciosa. Además, la IA puede aprender de patrones previos y amenazas conocidas para mejorar aún más su capacidad de detección.

REDUCCIÓN DE FALSOS POSITIVOS Y NEGATIVOS EN LOS SISTEMAS DE SEGURIDAD

» Uno de los desafíos clave en los sistemas de seguridad es la generación de falsos positivos (alertas erróneas) y falsos negativos (amenazas no detectadas). La inteligencia artificial aborda este desafío al entrenar algoritmos con conjuntos de datos exhaustivos y diversos que representan una amplia gama de amenazas y escenarios.

» Al aprovechar técnicas de aprendizaje automático, la IA puede identificar patrones sutiles y correlaciones en los datos de seguridad que pueden indicar una amenaza real. Esto permite una mayor precisión en la detección y reduce los falsos positivos, minimizando así el tiempo y los recursos desperdiciados en la investigación de alertas incorrectas. Además, la IA puede ajustar y mejorar sus algoritmos a medida que se recopilan más datos y se obtienen resultados reales, lo que conduce a una reducción adicional de los falsos negativos.

RESPUESTA Y MITIGACIÓN MÁS RÁPIDAS ANTE INCIDENTES

» La integración de la IA en la ciberseguridad brinda una respuesta y mitigación más rápidas ante los incidentes. La IA puede monitorizar de forma continua la actividad de la red y los sistemas, analizando grandes volúmenes de datos en tiempo real para identificar posibles amenazas.

» En caso de detectar un incidente, la IA puede generar alertas automáticas y proporcionar información detallada sobre la naturaleza del ataque, lo que permite una respuesta inmediata. Además, la IA puede colaborar con los equipos de seguridad al proporcionar recomendaciones específicas basadas en su análisis de datos históricos y conocimientos previos. Esto agiliza la toma de decisiones y la implementación de medidas de mitigación adecuadas, lo que resulta en una reducción significativa del tiempo necesario para contener y mitigar los impactos de un incidente de seguridad

MEJORA DE LA CAPACIDAD DE DEFENSA PROACTIVA Y ADAPTATIVA

» La inteligencia artificial mejora la capacidad de defensa proactiva y adaptativa en la ciberseguridad al permitir una detección temprana de amenazas y una adaptación continua a las tácticas de los ciberdelincuentes. Mediante el análisis de datos históricos y el uso de algoritmos de aprendizaje automático, la IA puede identificar patrones y comportamientos que indican la presencia de una amenaza potencial. Esto permite a los sistemas de seguridad anticiparse a los ataques y tomar medidas preventivas antes de que se produzcan. Además, la IA puede adaptarse y aprender de forma continua, actualizando sus modelos de detección y adaptándose a las nuevas técnicas y tácticas utilizadas por los ciberdelincuentes. Esto garantiza que los sistemas de seguridad se mantengan actualizados y sean capaces de enfrentar las amenazas emergentes de manera efectiva.

OPTIMIZACIÓN DE RECURSOS Y REDUCCIÓN DE COSTOS EN LA SEGURIDAD CIBERNÉTICA

» La integración de la IA en la ciberseguridad ofrece beneficios significativos en términos de optimización de recursos y reducción de costes. Al automatizar tareas rutinarias, como la clasificación de eventos de seguridad, el análisis de registros y la generación de informes, la IA libera a los profesionales de seguridad para que se centren en actividades más estratégicas y de alto valor. Esto aumenta la eficiencia operativa y permite un mejor uso de los recursos humanos disponibles. Además, al reducir los falsos positivos y negativos, la IA evita que los recursos se desperdicien en la investigación de amenazas que resultan ser falsas o en la falta de detección de amenazas reales. Esto permite una asignación más precisa de los recursos disponibles y ayuda a minimizar los impactos financieros de los ataques cibernéticos. La IA también puede ayudar a identificar y priorizar las áreas de mayor riesgo, lo que facilita una asignación eficiente de los recursos de seguridad y contribuye a una reducción de costos a largo plazo.

A continuación, se detalla de manera más técnica los que se consideran son los usos más avanzados de IA en los procesos de seguridad y donde como profesionales, consideramos se debe de plantear la inversión en proyectos que se apoyen en tecnologías de IA.

5.1. LOG PARSING O NAME ENTITY RECOGNITION

La principal fuente de información de ciberseguridad son los eventos y logs emitidos por los dispositivos y aplicaciones. Estos logs se usan para detectar amenazas de ciberseguridad como fuente de datos en técnicas avanzadas de inteligencia artificial o en técnicas tradicionales aplicando reglas de correlación asociadas a casos de uso.

El primer paso en la detección o análisis de incidentes de sería recoger y estructurar los logs.

Los logs se escriben de forma secuencial en ficheros y, aunque cada vez más sistemas los generan con un formato estructurado (JSON, XML, CEF, etc…), otras aplicaciones escriben los registros de una manera no estructurada y difícil de ingerir como en el ejemplo anterior.

El método tradicional de estructurar los logs es mediante el uso de expresiones regulares, llamadas regexp que, siendo una solución efectiva y potente, tiene grandes inconvenientes. Las expresiones regulares son difíciles de definir y mantener, son sensibles a pequeñas modificaciones y computacionalmente es costoso. El uso de modelos de inteligencia artificial resuelve estos problemas.

Los sistemas de inteligencia artificial son modelos estadísticos capaces de detectar las entidades mediante representaciones de los datos observados en función de sus características. Los sistemas de reconocimiento de entidades entrarían dentro del campo de procesamiento del lenguaje natural que mediante entrenamiento pueden ser capaces de estructurar los logs.

En 2018 la universidad de Edith Cowan publicó un paper con el análisis de un modelo basado en BLSTM entrenado para realizar el log parsing con unas métricas excelentes: 99.98% precisión, 99.94% recall, 99.96% F1 score, and 99.98% accuracy.

El uso de modelos de NLP para realizar tareas de logs parsing no ha quedado relegado sólo a un paper. En 2021 Splunk, uno de los lideres en SIEM según Gartner publicó el análisis de tres tipos de modelos NER: Conditional Random Fields (CRF), una shallow Network (LSTM) y dos deep Network pre-entrenadas basadas en transformers (BERT y miniBERT).

Este análisis realizado en un entorno más realista obtiene unos resultados prometedores. Se puede observar cómo los modelos basados en redes generativas, los antecesores de ChatGPT-4 (BERT y MiniBERT) tiene unas métricas inferiores a otros modelos, sin embargo, los avances en redes generativas como ChatGPT-4 indica que este análisis tendría un resultado mejor tanto en métrica como en coste..

5.2. DETECCIÓN MEDIANTE MODELOS DE IA

Los sistemas tradicionales de detección basados en SIEM, introducidos en el año 2005, permiten agrupar los logs y los eventos bajo un mismo sistema. Sobre estos sistemas se definen casos de uso que, en los SIEM más básicos, no son más que reglas deterministas que, mediante la correlación de los eventos detectan ataques o incidentes de seguridad. Sin embargo, la correlación no implica causalidad por lo que se generan alertas que después de un análisis por parte de los analistas son falsos positivos o positivos contenidos, además de tener un sobreajuste muy grande.

Los sistemas basados en IA son expertos en identificar anomalías y detectar amenazas que pasan desapercibidas para las medidas de seguridad convencionales. Los algoritmos de aprendizaje automático pueden analizar grandes cantidades de datos, incluido el tráfico de la red, el comportamiento del usuario y los registros del sistema, para establecer patrones normales e identificar desviaciones que puedan indicar un ataque cibernético. AI aprender continuamente de nuevos datos, los sistemas de IA pueden adaptarse y evolucionar para detectar amenazas emergentes de manera efectiva.

Los modelos generativos actuales como ChatGPT han abierto un abanico de posibilidades y ya se están reentrenando con información específica de ciberseguridad como por ejemplo la información de MITRE ATT&CK para mejorar su respuesta ante este tipo de problemas.

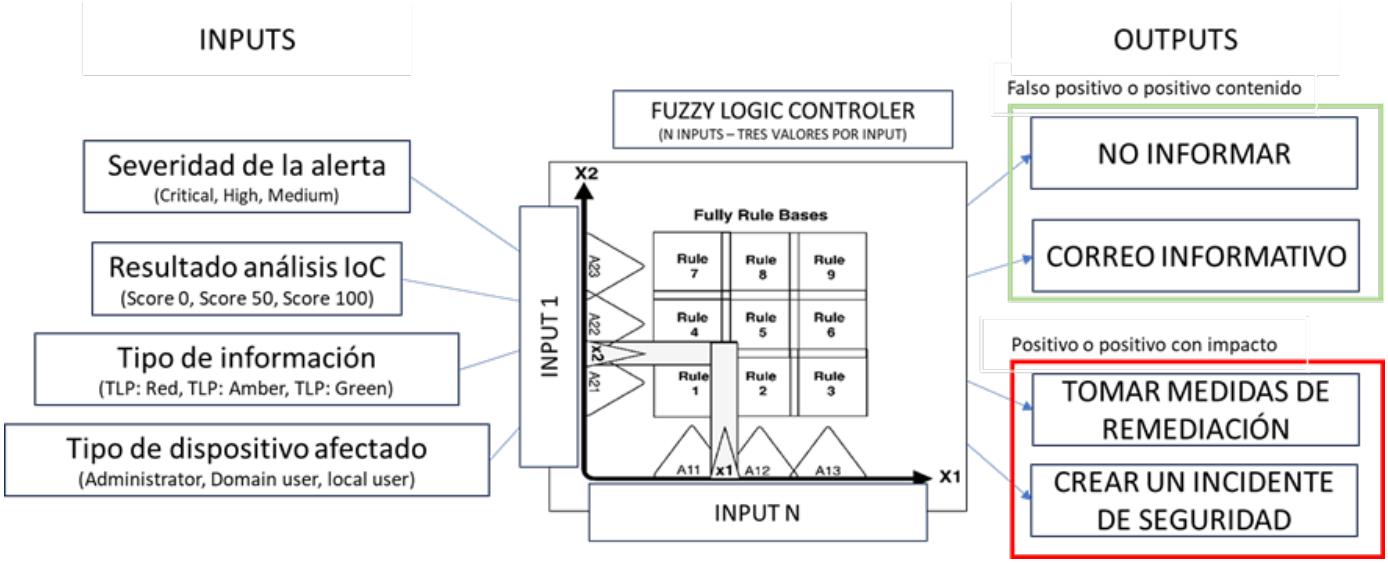
5.3. PLAYBOOKS: TOMA DE DECISIONES Y RESPUESTA A INCIDENTES

La detección de amenazas es sólo una parte de la seguridad. También se necesita un análisis inteligente y eficaz ante posibles incidentes. Debido al volumen cada vez mayor de alertas es inviable tratarlas de forma manual por esa razón los sistemas de orquestación y automatización son necesarios.

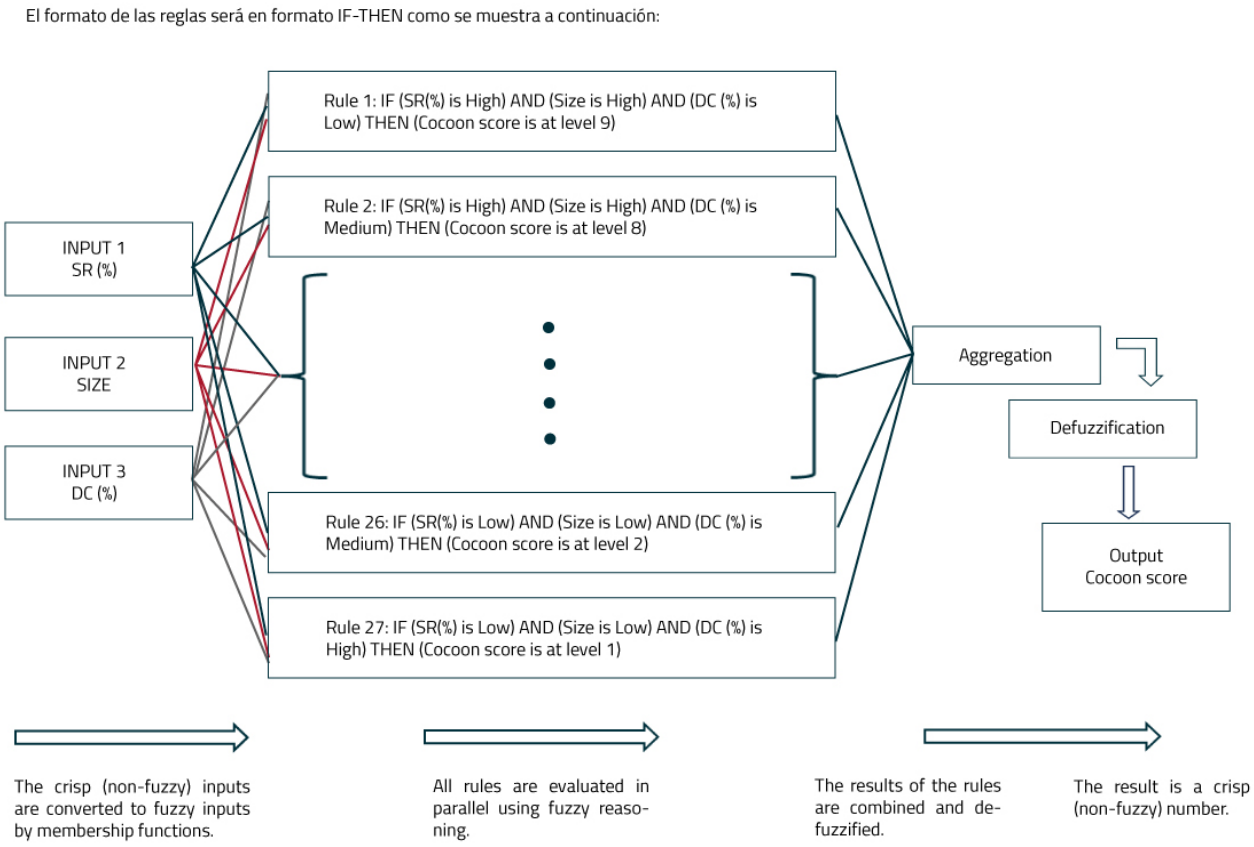
Implementar playbooks eficaces es la base del éxito ya que permiten automatizan gran parte del análisis, por ejemplo, el enriquecimiento mediante consultas a herramientas de inteligencia de amenazas. Sin embargo, existen acciones que necesita la intervención de un analista ya que la información disponible es incompleta, ambigua o incierta.

La lógica difusa es una rama de la inteligencia artificial que propone un enfoque alternativo a la lógica clásica. En la lógica clásica se utiliza valores booleanos (verdadero/falos), los sistemas de detección ya han decidido que la amenaza es real y ha generado una alerta. En el campo de la ciberseguridad, la lógica difusa va a ayudar a modelar el comportamiento del usuario y de las amenazas categorizando la información en un rango determinado entre dos polos opuestos. Permite manejar la falta de precisión de ciertos términos, como "poco" o "mucho", "Low" o "Medium" y contemplar puntos intermedios. La lógica difusa imita la manera de tomar decisiones de los seres humanos lo que permite una mayor flexibilidad en la respuesta a situaciones imprevistas o inciertas.

La siguiente figura representa los posibles inputs al modelo. El controlador de lógica difusa analizará las reglas y se tomará una decisión.



El formato de las reglas será en formato IF-THEN como se muestra a continuación:



En la ilustración anterior se ha representado la decisión del modelo en dos niveles: falso positivo/positivo contenido o positivo/positivo con impacto. El segundo nivel que sería otro modelo de lógica difusa que se implementaría en un paso posterior en el playbooks en el que los inputs condicionarían la remediación que se debe aplicar.

EJEMPLO PLAYBOOK BASADO EN IA PARA DDOS

- INPUTS:
  - » Volumen de tráfico generado.
  - » Número de IPs origen.
  - » Número de IPs destino.
  - » Tipo de servicio que recibe el ataque.
  - » Etc.

- OUTPUT, LA REMEDIACIÓN QUE APLICAR:
  - » Mitigar todo el tráfico.
  - » Aplicar traffic shaping.
  - » No mitigar
  - » Etc.

5.4. DETECCIÓN DE PHISHING

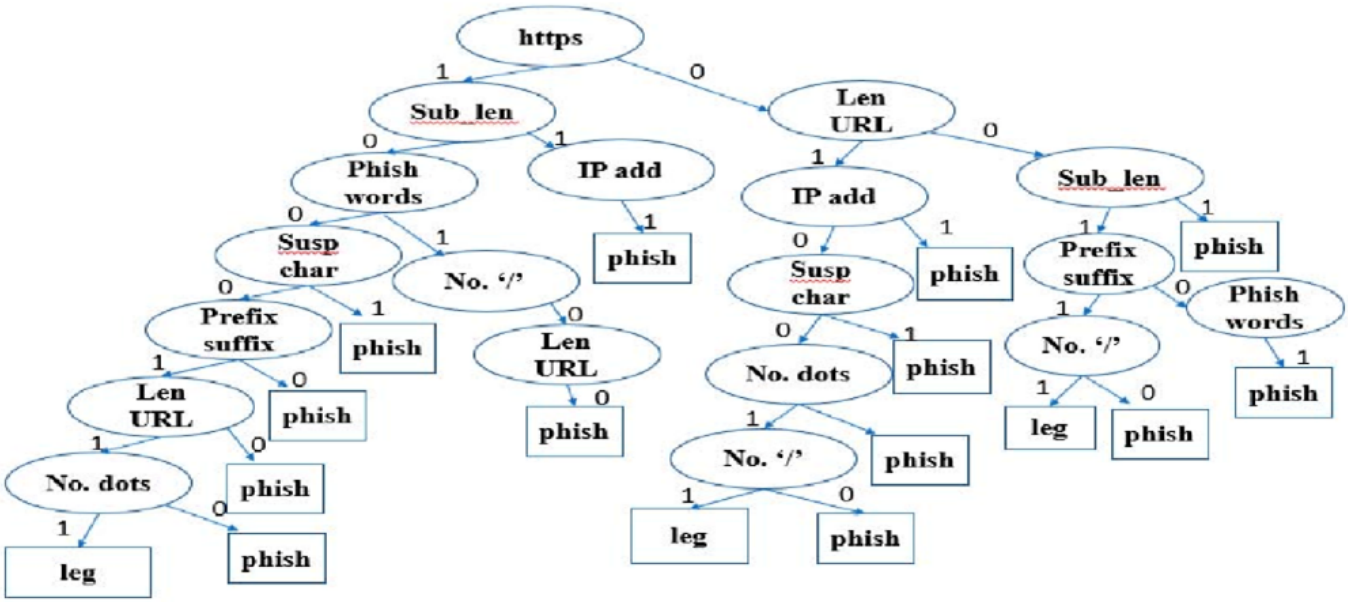
La detección de phishing se puede realizar encadenando técnicas de inteligencia artificial. Los primeros modelos permitirán extraer características del email que se emplearán en el modelo final que es el encargado de clasificará si el email es phishing o no.

El cuerpo del mensaje es una variable importante que analizar. Por norma habitual las campañas de phishing masivo generan los correos de manera automática y empleando palabras clave, por ejemplo, relacionadas con la urgencia en realizar una acción. En una primera iteración es posible usar una red generativa, tipo ChatGPT, para que analice es cuerpo del mensaje e indique si ha sido generado automáticamente y si detecta palabras clave en email de phishing.

Los enlaces que aparecen en los correos redirigen a destinos fraudulentos. Estos dominios se han podido generar automáticamente mediante una técnica llamada DGA, otro paso previo que se puede hacer es utilizar modelos de Deep Learning basados en NLP que cuantifican la similitud semántica entre dominios y que nos indica la probabilidad de que sea o no un dominio DGA.

Estas características junto con otras relevantes que se incluyen en el correo (dirección IP, longitud de la url, etc.) se utilizan en el entrenamiento de un modelo supervisado de clasificación del tipo árbol de decisión. Durante el entrenamiento se generará un modelo que se usará para realizar la predicción sobre los nuevos correos.

El modelo se puede representar tal y como se muestra en la siguiente imagen:



Este tipo de modelos tiene una ventaja respecto a otros modelos que podrán ser óptimos. Como se puede ver en la imagen anterior, el árbol se puede pintar y la predicción siempre estará asociada a los valores de las características extraídas del correo para su análisis. Esto permite que el modelo sea explicable y que el analista sepa en cada momento porque se ha clasificado el email como phishing o no. Esta característica es importante ya que a la hora de tomar una decisión en algunos casos se exige que se explique la razón por la que se ha clasificado de esa manera.

.

5.5. IDENTIFICACIÓN DE MALWARE Y RANSOMWARE

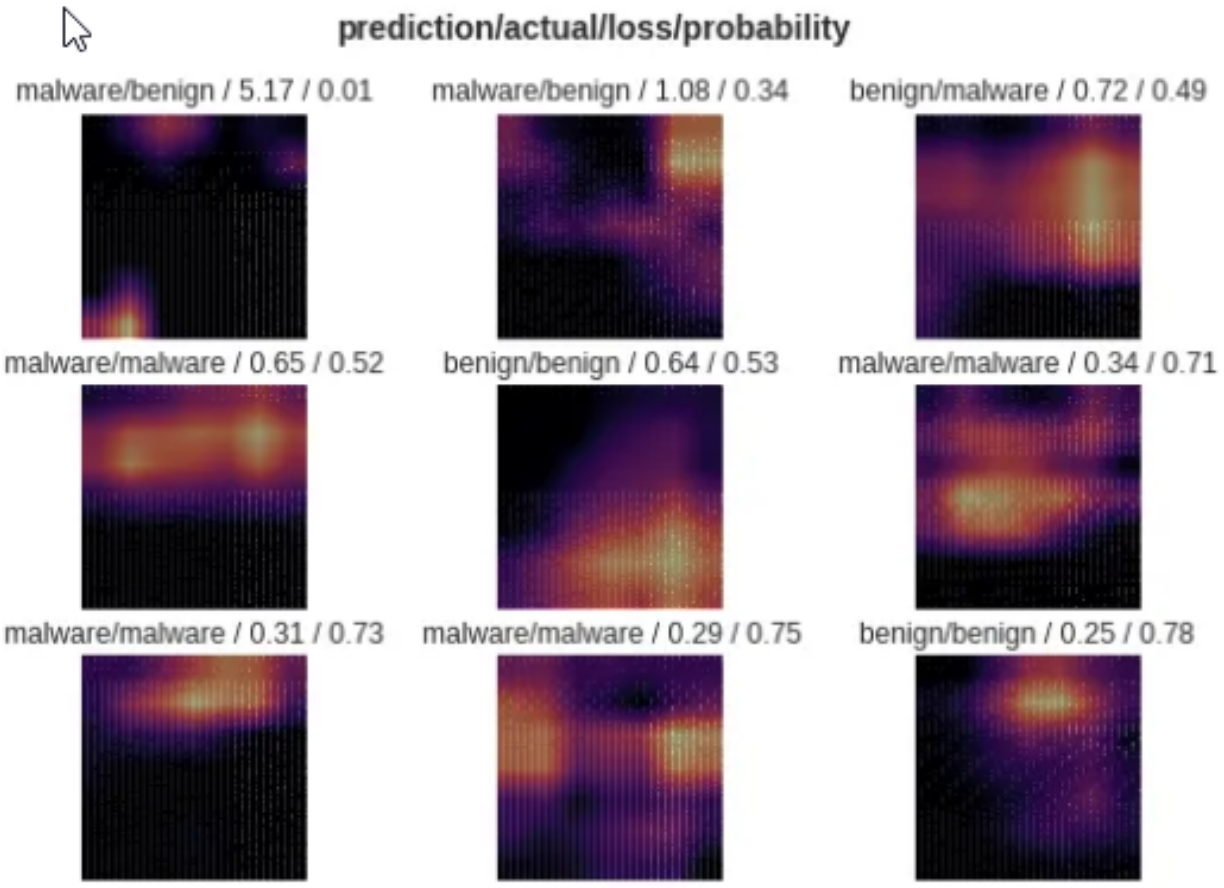
La protección de los sistemas ante ataques del malware/ransomware es uno de los mayores retos de seguridad en la actualidad. El análisis es manual de forma estática, extrayendo información útil del código o de forma dinámica, ejecutando la muestra en una sandbox.

Con modelos de inteligencia artificial se puede hacer un análisis de malware/ransomware de una manera más optima e incluso que nos permita detectar nuevos actores o variaciones de los que existen.

El análisis se basa en la potencia de las redes neuronales convolucionales y su capacidad de extraer características inapreciables para el ser humano. El primer paso es la preparación de los datos ya que los ficheros a analizar se deben transformar en imágenes para que puedan ser explotados por la CNN (Red Neuronal Convolucional).

Una vez definida la CNN a utilizar, en este caso la conocida como "attention-VGG16 network", se comienza a entrenar el modelo con el conjunto de training que contiene muestras de ficheros de malware/ransomware y ficheros lícitos. El análisis de las métricas se realizará con el conjunto de test y se reentrenará el modelo adaptando los parámetros del modelo hasta obtener unas métricas de precisión, recall y accuracy adecuados.

Los resultados de los test realizados se pueden ver en la siguiente figura. Las imágenes que aparecen no corresponden con la transformación del fichero en imagen. Se ha aplicado una técnica que ilumina con colores cálidos las características de la imagen en la que se ha basado el modelo para realizar la predicción. Con una muestra más alta y una clasificación basada en el tipo de malware/ransomware se podría sacar una conclusión sobre las características, la zona, sobre la que se ha fijado el modelo para tomar la decisión.



5.6. PREDICCIÓN Y ANTICIPACIÓN DE AMENAZAS EMERGENTES

Los modelos de aprendizaje máquina utilizados para la detección de amenazas se dividen en dos; supervisado o no supervisados. En ambos casos una vez entrenados los modelos el sistema permanece estático hasta que se vuelva a reentrenar. Sin embargo, en ciberseguridad para las amenazas emergentes o ataques de Zero-day los datos de entrenamiento no están disponibles antes de que se produzca el ataque.

Zero-Shot learning es un tipo de aprendizaje automático en el que se crea un modelo de aprendizaje profundo preentrenado para generalizar las categorías iniciales. Zero-shot trata de imitar a los humanos y encontrar similitudes de forma natural entre las clases reentrenando el modelo para identificarlos.

El objetivo principal del aprendizaje Zero-shot es obtener la capacidad de predecir los resultados sin muestras de entrenamiento; la máquina debe ser capaz de reconocer los objetos de las clases que no se entrenaron. El aprendizaje Zero-Shot se basa en la transferencia de conocimiento que está contenido en las instancias alimentadas durante el entrenamiento.

Se propone el aprendizaje Zero-Shot para aprender capas y propiedades semánticas intermedias, y luego aplicarlo para predecir una nueva clase de datos invisibles. Por lo tanto, aprovechamos las capacidades de Zero-shot Learning (ZSL) que pueden manejar de manera efectiva las clases invisibles en comparación con las técnicas tradicionales de aprendizaje automático.

5.7. AUTENTICACIÓN Y CONTROL DE ACCESO BASADO EN IA

En 2021, los incidentes de robo de identidad representaron más de 6,1 billion dollars en pérdidas financieras y el número de quejas de consumidores por robo de identidad aumentó un 3,3%, hasta algo más de 1,43 millones.

Los ciberdelincuentes cada vez son más sofisticados, los métodos de autenticación tradicionales, como las contraseñas y los PIN, ya no son suficientes para proteger nuestros datos sensibles. La necesidad de métodos de autenticación más seguros, fiables y cómodos ha llevado la utilización de la IA para autenticación.

El reconocimiento facial es un método de autenticación biométrica que verifica la identidad de una persona midiendo y reconociendo patrones faciales a partir de una imagen. La tecnología de reconocimiento facial utiliza IA y machine Learning para comparar estos puntos de datos con una base de datos o la foto de un documento de identidad.

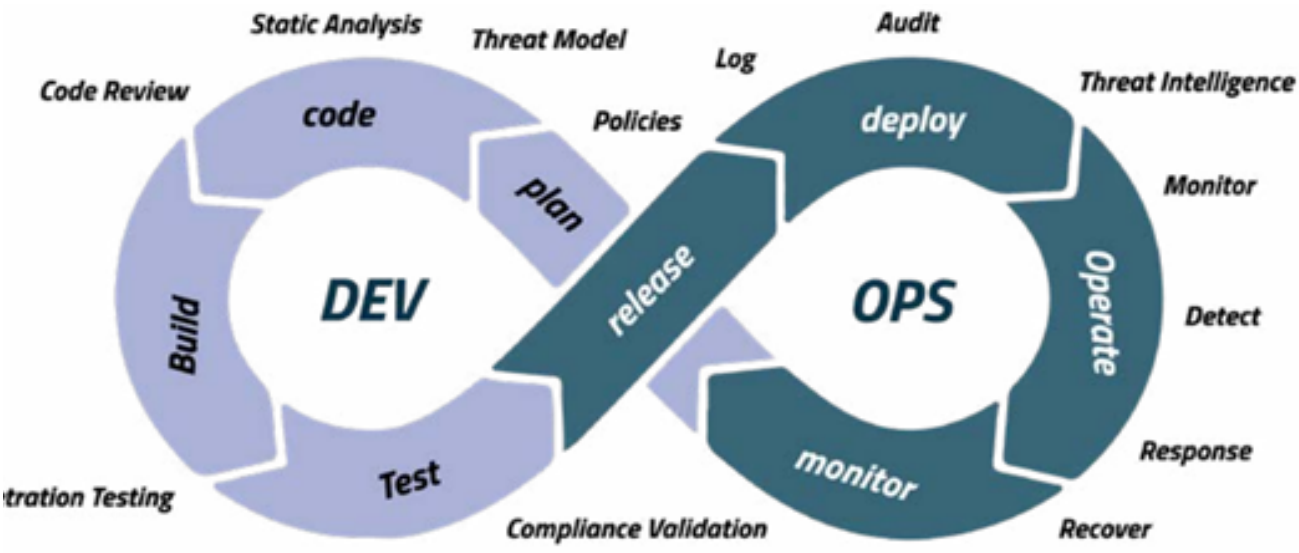
Los datos biométricos de un rostro, como el tamaño de los ojos, el espaciado, la forma de los pómulos, la longitud de la nariz, la anchura de la nariz y el tamaño de las pupilas, se transcodifican en formato digital para compararlos con la foto de un documento de identidad del usuario.

Estos métodos son más seguros. Las contraseñas pueden olvidarse o ser robadas, mientras que tus datos biométricos son únicos y, mientras los deepfakes lo permitan, son más difíciles de replicar. Esto significa que es mucho más difícil para alguien obtener acceso no autorizado a tus cuentas o información. Y además los datos biométricos son únicos por lo que es más difícil que alguien acceda sin autorización a tus sistemas.

5.8. AUTOMATIZACIÓN DE LA SEGURIDAD DEVSECOPS

El ciclo de vida de desarrollo integrado con la seguridad requiere de la automatización de ciertos procesos en el despliegue de software. Para ello mediante la filosofía DevSecOps, podemos integrar una serie de herramientas que validen que cada cambio que sube a producción cumpla con el estándar de ciberseguridad y evite un riesgo a los clientes y a la empresa. Dentro de este tipo de pruebas automáticos encontraremos diferentes soluciones que permiten securizar nuestra aplicación desde un modo básico a un modo experto y como la IA puede aportar mejora en el proceso.

Puede encontrar información adicional sobre DevSecOps en la "Guía de iniciación en la Seguridad aplicada a DevOps" publicada por ISMS Forum.



Fuente: Guía de iniciación en la seguridad aplicada a DevOps (ISMS Forum)

5.8.1. AUTOMATIZACIÓN ESTÁTICA O DINÁMICA

A continuación, se definen los conceptos de SAST, SCA y DAST relativos a los análisis de proyectos

SAST O STATIC ANALYSIS SECURITY TESTING

Abarca el análisis de un proyecto de forma estática evaluando las posibles vulnerabilidades de seguridad más comunes como las explicadas en el TOP 10 OWASP. Este enfoque deriva en una fase posterior de análisis manual por el equipo de desarrollo y por un equipo de seguridad para certificar que las vulnerabilidades localizadas no sean falsos positivos.

Abarca el análisis de un proyecto de forma dinámica, es decir, se trata de una prueba ofensiva sobre una aplicación en ejecución. Este enfoque permite simular un ataque cercano a uno real ya que simula como un atacante lanzaría una prueba contra nuestros servicios con la diferencia de automatizar el proceso y permitir bloquear despliegues con vulnerabilidades a producción.

DAST O DYNAMIC ANALYSIS SECURITY TESTING

SCA OSA OPEN SOURCE) ANALYSIS

Abarca el análisis de las dependencias o librerías de código abiertoconsumidas por nuestros proyectos y como resultado informa de las librerías con vulnerabilidades abiertas. Tipos de riesgos que pueda levantar este proceso: exposición de datos o sistemas internos, exposición a una demanda judicial por uso indebido de una librería, falta de control sobre nuestra aplicación debido a una librería sin mantenimiento, etc.

5.8.2. PROBLEMAS O GAPS DE LOS ANALIZADORES AUTOMÁTICOS

Los tres métodos automatizados anteriores traen muchos beneficios, pero tienen una serie de problemas o gaps:

» **Falsos positivos:** Los analizadores de código estático evalúan de forma aislada una aplicación sin conocer el contexto que rodea. Ello lleva a presuponer que una vulnerabilidad es explotable cuando en determinados casos una herramienta superior protege de ese tipo de ataque. Este tipo de situaciones generan mucho ruido en equipos de desarrollo que invierten tiempo en solventar vulnerabilidades que quizá no sean explotables.

» **Triajes con un alto coste económico y temporal:** Una vez actúan dichas herramientas generan resultados que son necesarios de evaluar por un equipo experto y que dependiendo del tamaño de los repositorios puede tener un alto coste para una compañía para validar que los resultados sean explotables o falsos positivos.

» **Alejados de la realidad de un atacante:** Al evaluar pequeñas piezas por separado, estas herramientas se alejan de la realidad de un atacante enfrentándose a una web o servidor. Ciertamente previenen, pero no aseguran el Zero risk.

» **Costes altos en repositorios de código grandes:** El código legacy es un mal de muchas grandes compañías y que conllevan un alto coste en herramientas y servidores para analizar el código contenido.

» **Full scan vs Incremental Scan problemática:** Muchas herramientas de análisis automático tienen un gran dilema y es cuando realizar un scan completo a una aplicación y cuando un scan parcial a solo el cambio modificado por un desarrollador. Esta decisión tiene consecuencias tanto económicas como de riesgo.

5.8.3. ¿CÓMO PODRÍA LA IA MEJORAR ESTOS PROCESOS?

SECURIZANDO LAS DEPENDENCIAS

Las vulnerabilidades provenientes de librerías "open source" tienden a ser un gran problema en riesgos y en consumo de tiempo para el desarrollador de revisar y actualizar. Para ello mediante la IA podemos automatizar dicho proceso y permitir la actualización de librerías de un modo automatizado y seguro. Dicha IA elaboraría un cambio del código de forma automática cada vez que detectara una librería con una vulnerabilidad y la aparición de un "fix".

SECURIZANDO LOS SECRETOS

Mediante la implantación de una IA que estudie y entienda el contexto de las aplicaciones de una compañía es viable deducir cuando una password es realmente productiva o meramente un código de pruebas o malentendido léxico.

AUTOMATIZACIÓN DE LOS TRIAJES

Antes comentábamos que las compañías necesitan gastar millones en equipos especialistas en seguridad para evaluar los resultados de las herramientas SAST y DAST. Siendo un equipo cross a una organización resulta complicado que tengan el contexto completo y puedan evaluar siempre correctamente, por ello una IA entrenada podría suplir dicho rol dentro de las compañías y reducir entorno a un 90% los falsos positivos. IBM trabajó hace ya tiempo dicho campo para conseguir una reducción significativa de falsos positivos.

APRENDIZAJE MEDIANTE ATAQUES

Mediante la IA podemos establecer un modelo que aprenda de ataques directos contra nuestras aplicaciones. De dichos ataques la IA aprendería patrones y establecería una capa de testing extra más cercana a la realidad permitiendo reducir el número de incidentes en producción.

PEER PROGRAMMING JUNTO A UNA IA

Actualmente existen herramientas como Copilot que autogeneran código junto al desarrollador, dicha autogeneración se basa en el aprendizaje interno de la empresa de miles de repositorios públicos. Esto permite recomendar buenas prácticas de programación, de performance y de seguridad. De este modo aplicaríamos un 4 eyes durante el desarrollo de software mejorando la calidad del desarrollo.

5.9. PRUEBAS SOBRE APLICACIONES

Las pruebas de regresión que se deben realizar al desplegar un nuevo código de una aplicación o simplemente buscar errores en una aplicación es un trabajo arduo y costoso. En este aspecto la inteligencia artificial nos puede facilitar la tarea. Se pueden desplegar modelos que analicen los ficheros de logs para detectar posibles fallos o se pueden utilizar las nuevas redes generativas como Copilot que tiene la funcionalidad de "Explain codec" o utilizar su función de chat para que te proponga mejoras en el código.

6 MOTIVOS POR LOS QUE PUEDE FALLAR UNA IMPLANTACIÓN Y USO DE UNA IA.

Aun teniendo claros los objetivos, y los resultados a obtener, habiendo realizado la elección de un proveedor o una tecnología adecuada todos los proyectos de IA están expuestos a una serie de obstáculos que pueden dificultar o incluso hacer fracasar la implantación y uso de una IA, esto son, entre otros:

- **Falta de integración con los sistemas existentes**: La ausencia de compatibilidad, o la dificultad para integrar la solución de IA con los sistemas y procesos de la organización, puede limitar su efectividad.
- **Falta de soporte y mantenimiento adecuados**: La carencia de soporte adecuado, por parte del proveedor de IA, la deficiencia en las actualizaciones o un mantenimiento irregular pueden desencadenar en problemas de estabilidad y rendimiento de la solución.
- **Falta de capacidad interna para comprender y utilizar la IA**: No contar con personal capacitado o formado, que comprenda y utilice eficazmente la solución de IA, podría producir efectos indeseados en su empleo, rendimientos poco adecuados o incapacidad para alcanzar las metas fijadas.
- **Falta de escalabilidad**: Si la solución elegida no es escalable o no puede adaptarse a las nuevas necesidades y requisitos de la organización esto puede limitar su utilidad y funcionalidad a largo plazo.
- **Falta de planificación y gestión adecuada**: Como en cualquier proyecto, una dirección ineficiente dificultará el cumplimiento de los plazos, desviará los presupuestos o no permitirá alcanzar los resultados esperados.
- **Falta de colaboración entre equipos:** El uso o despliegue de una IA implica la colaboración de variados equipos y perfiles (científicos de datos, expertos, asesores legales, …) esto puede provocar una falta de comunicación o colaboración entre ellos y desencadenar en un conjunto de malentendidos o desencuentros que provoquen retrasos, desvió de los objetivos marcados o incapacidad de poner en marcha el sistema.
- **Falta de confianza de los clientes:** El uso de IA siempre plantea desafíos éticos y morales que, sin una adecuada gestión, pueden dañar la confianza que los usuarios depositan en la organización, pudiendo provocar efectos negativos innecesarios.
- **Incompatibilidad con los requisitos y objetivos de la organización:** Si la solución de IA no cumple con los requisitos y objetivos de la organización, su implantación puede resultar inadecuada y no proporcionar los resultados esperados.

- **Rendimientos sesgados**: Cuando los conjuntos de datos o algoritmos pueden perpetuar sesgos o discriminación se presentan dilemas éticos, consecuencias legales, o reputacionales de la organización, provocadas por una codificación incorrecta y no supervisada o por datos erróneos o espurios.

- **Rendimientos deficientes o inexactos:** Cuando no se alcanzan unos resultados precisos, esperados o confiables, se puede generar una falta de confianza en la propia tecnología, lo que dificulta su adopción y uso dentro de la organización, además de exponer la reputación de la misma.

- **Resistencia organizativa o cultural:** La gestión del cambio es compleja en cualquier organización y máxime cuando hablamos de una tecnología tan disruptiva como es la IA, la implantación de soluciones IA provoca cambios en los roles y en los procesos. Por un lado, el desconocimiento de la tecnología puede provocar una falta de confianza en ella, y por otro lado se pueden generar temores en el personal que puede ver peligrar sus puestos de trabajo. En ambos casos, sin una comunicación y formación adecuada se puede provocar una oposición a la implantación de la solución, lo cual dificultaría su adopción en la organización.

- **Adopción por moda:** La IA se ha convertido en una moda, muchas organizaciones sienten el impulso de adoptar IA por puro marketing sin saber bien qué están implantando o sin darse cuenta de que su necesidad se podría resolver con herramientas más simples que no tengan las letras "IA".

7 REFERENCIAS

Este documento blanco ha sido realizado por un grupo de profesionales que han utilizado las siguientes fuentes de referencia (además de las fuentes específicas referidas en los diferentes puntos).

Estudios de mercado:

- McKinsey, The state of AI: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review
- Informe de Panorama de Amenazas de ciberseguridad 2022. Enisa https://administracionelectronica.gob.es/pae_Home/pae_Actualidad/pae_Noticias/Anio2022/Noviembre/Noticia-2022-11-04-Informe-Panorama-Amenazas-ciberseguridad-2022.html

Otras referencias de interés

- Programa Nacional de Algoritmos Verdes. que impulsa una IA verde por diseño (Green by Design): https://portal.mineco.gob.es/RecursosNoticia/mineco/prensa/noticias/2022/20221213_plan_algoritmos_verdes.pdf]
- Machine learning for human learners: opportunities, issues, tensions and threats - Mary E. Webb, Andrew Fluck, Johannes Magenheim, Joyce Malyn-Smith, Juliet Waters, Michelle Deschênes, Jason Zagami: https://link.springer.com/article/10.1007/s11423-020-09858-2
- WormGPT: https://slashnext.com/blog/wormgpt-the-generative-ai-tool-cybercriminals-are-using-to-launch-business-email-compromise-attacks/
- Jailbreak Chat: https://www.jailbreakchat.com/
- MITRE ATLAS: https://atlas.mitre.org/
- The Prompt Report: https://www.thepromptreport.com/
- BugBounty OpenAI: https://bugcrowd.com/openai

Identificación y preparación de datos:

- Recopilación de datos para aprendizaje automático e inteligencia artificial: una guía completa - Shaip
- Tipos de fuentes de datos (questionpro.com)
- Requisitos para Auditorías de Tratamientos que incluyan IA (aepd.es)

Selección de modelo de inteligencia artificial

- Machine Learning: Cómo Desarrollar un Modelo desde Cero | by Victor Roman | Ciencia y Datos | Medium
- ¿Cuáles son los 5 mejores marcos de IA de código abierto? (containerize.com)
- Requisitos para Auditorías de Tratamientos que incluyan IA (aepd.es)

Entrenamiento del modelo

- Inteligencia artificial fácil - Machine Learning y Deep Learning prácticos - Funciones de pérdida (Loss function) | Editiones ENI (ediciones-eni.com)
- Cross validation: qué es y su relación con machine learning (tec.mx)
- Requisitos para Auditorías de Tratamientos que incluyan IA (aepd.es)

Evaluación del modelo

- Métricas De Evaluación De Modelos En El Aprendizaje Automático (datasource.ai)
- Sobreajuste y Subajuste en Machine Learning - Aprende IA
- https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf
- https://github.com/jiep/offensive-ai-compilation/blob/main/README.md#-image-
- https://www.mobbeel.com/blog/deepfakes-garantizar-autenticidad/

ENERO 2024

INTELIGENCIA ARTIFICIAL Y CIBERSEGURIDAD

www.ismsforum.es
info@ismsforum.es
(+34) 915 63 50 62

@ISMSForum
ISMS Forum