IA Explicable (XAI) y Ciencia de Datos

Jose Dario Menendez Acosta

Facultad de Ingeniería Eléctrica y Electrónica, Universidad Nacional de Ingeniería

CBS06 M Inteligencia Artificial II

Ing. Yury Oscar Tello

23 de Setiembre de 2025

Tabla de contenido

Resumen

Este documento explora la convergencia transformadora entre la ciberseguridad y la ciencia de datos, impulsada por la creciente complejidad de las amenazas digitales. Se detallan los cuatro tipos de análisis de datos —descriptivo, diagnóstico, predictivo y prescriptivo— que fundamentan las estrategias de ciberseguridad modernas. Se aborda el desafío de los modelos de "caja negra" y se establece la necesidad crítica de la Inteligencia Artificial Explicable (XAI) para generar confianza, mejorar la toma de decisiones y cumplir con los requisitos normativos.

Se propone un marco de cinco capas para el modelado explicable y se discuten las aplicaciones prácticas del descubrimiento de conocimiento. Adicionalmente, se examina el rol del Blockchain para asegurar la integridad de los datos de entrenamiento y se presenta un marco para evaluar la complejidad y robustez de las explicaciones de XAI. Finalmente, se analizan las vulnerabilidades de seguridad inherentes a los propios métodos de XAI, concluyendo que el futuro de la ciberseguridad depende de una IA que no solo sea potente, sino también transparente, robusta y segura.

La Convergencia de la Ciberseguridad y la Ciencia de Datos: Hacia un Modelado Explicable con IA

En el contexto actual de la computación, la tecnología y las operaciones de ciberseguridad están en constante evolución, y la ciencia de datos se ha convertido en el motor principal de este cambio. La construcción de modelos basados en datos que extraen patrones de incidentes de ciberseguridad es clave para automatizar y gestionar de manera inteligente un sistema de seguridad. La ciberseguridad juega un rol cada vez más importante en una era donde los paisajes digitales crecen continuamente y la tecnología permea casi todos los aspectos de nuestras vidas. Formalmente, la ciberseguridad se define como un conjunto de tecnologías y procesos diseñados para proteger computadoras, redes, programas y datos de ataques, daños o accesos no autorizados (Sarker, 2024).

La creciente interconexión de sistemas y el uso de tecnologías basadas en datos han hecho que tanto organizaciones como individuos sean más vulnerables a las ciberamenazas. A medida que la complejidad de la ciberseguridad aumenta, los métodos de defensa tradicionales a menudo resultan insuficientes para proteger a las organizaciones contra adversarios dinámicos y adaptables. La necesidad de enfoques innovadores para salvaguardar los activos digitales es impulsada por la conectividad creciente y las amenazas sofisticadas (Sarker, 2024).

Este documento explora la ciencia de datos en ciberseguridad, una disciplina de vanguardia que combina el análisis de datos, el aprendizaje automático y la experiencia en el dominio para mejorar las medidas de ciberseguridad. Se profundiza en la intrincada intersección de ambas áreas, con un enfoque en el análisis avanzado, la extracción de conocimiento y la creación de modelos explicables. El objetivo es demostrar cómo esta convergencia está remodelando los paradigmas de detección de amenazas, respuesta a incidentes y mitigación general de riesgos. El descubrimiento de conocimiento y reglas

juega un papel fundamental en la aplicación de análisis avanzados para desarrollar modelos resilientes (Sarker, 2024).

La adopción de la ciencia de datos en ciberseguridad es cada vez más importante a medida que las organizaciones se esfuerzan por adelantarse a los ciberadversarios. Mediante el uso de conocimientos basados en datos, se puede fortalecer la ciber-resiliencia, fomentar una mejor comprensión de las amenazas emergentes y habilitar estrategias de defensa proactivas. Un aspecto crucial es la necesidad de transparencia y modelado explicable, que permite a los profesionales y partes interesadas comprender las decisiones tomadas por los sistemas de análisis avanzados (Sarker, 2024).

Tipos de Análisis en la Ciencia de Datos para la Ciberseguridad

En los procesos del mundo real, es común y fundamental plantearse preguntas clave para comprender y resolver científicamente un problema particular, tales como: "¿Qué pasó en el pasado?", "¿Por qué pasó?", "¿Qué pasará en el futuro?" y "¿Qué acción se debe tomar?". Basado en estas preguntas, se pueden distinguir cuatro tipos de análisis (Sarker, 2024).

Análisis Descriptivo

En el modelado de ciberseguridad, el análisis descriptivo es crucial para comprender y caracterizar el panorama de seguridad digital actual de una organización. Este tipo de análisis implica la interpretación de datos históricos de ciberseguridad, como registros de incidentes, informes de inteligencia de amenazas y patrones de actividad de la red. Analizar incidentes históricos, identificar vectores de ataque comunes y reconocer patrones en actividades maliciosas es fundamental para comprender la naturaleza de las amenazas.

Mediante técnicas de visualización y resúmenes estadísticos, este enfoque simplifica la presentación de tendencias, permitiendo a las organizaciones fortalecer sus posturas de seguridad de manera proactiva (Sarker, 2024).

Análisis Diagnóstico

El análisis diagnóstico en ciberseguridad se utiliza para profundizar en los patrones y anomalías identificados, con el fin de determinar las causas raíz de los incidentes de seguridad. Esta fase incluye la investigación detallada de las amenazas detectadas, el análisis forense y el estudio de los vectores de ataque. El análisis diagnóstico permite la identificación de Indicadores de Compromiso (IOCs) a partir de registros del sistema y tráfico de red, así como el desarrollo de estrategias de remediación específicas, brindando a las organizaciones la oportunidad de abordar problemas subyacentes y mitigar riesgos futuros (Sarker, 2024).

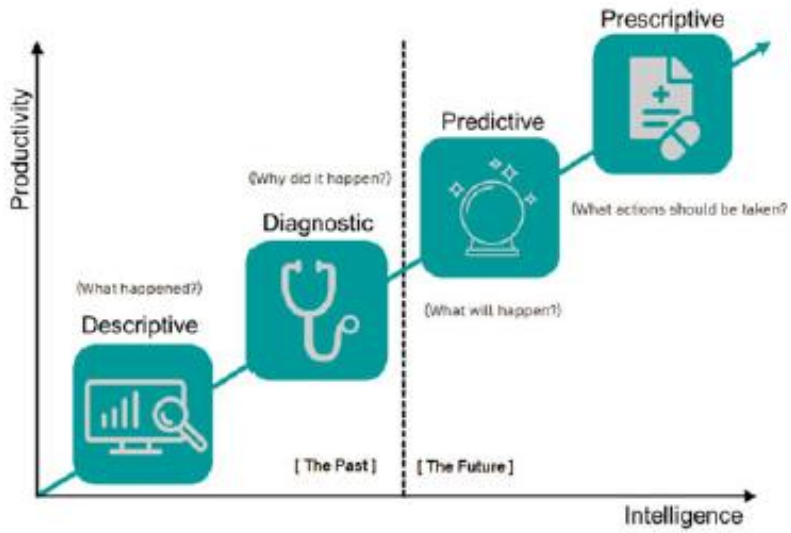
Análisis Predictivo

El análisis predictivo se utiliza para anticipar ciberamenazas y brechas de seguridad utilizando datos históricos y algoritmos avanzados. Este tipo de análisis emplea patrones, tendencias y anomalías de incidentes pasados para predecir posibles ataques o vulnerabilidades futuras. Se utilizan modelos de aprendizaje automático, como algoritmos de detección de anomalías, para identificar amenazas emergentes que podrían no ser evidentes a través de medidas de seguridad tradicionales. Este enfoque prospectivo permite a las organizaciones implementar medidas preventivas de manera proactiva y asignar recursos de manera efectiva (Sarker, 2024).

Análisis Prescriptivo

El análisis prescriptivo en ciberseguridad se utiliza para proporcionar recomendaciones y estrategias procesables con el fin de optimizar las defensas de seguridad. El objetivo es sugerir el curso de acción más efectivo para mitigar y responder a posibles ciberamenazas, basándose en los conocimientos derivados de los análisis diagnóstico y predictivo. El análisis prescriptivo guía a los equipos de ciberseguridad sobre cómo implementar controles de seguridad específicos, ajustar políticas o adoptar nuevas tecnologías para abordar vulnerabilidades de forma proactiva (Sarker, 2024).

En resumen, los análisis descriptivo y diagnóstico examinan el pasado para aclarar qué sucedió y por qué, mientras que los análisis predictivo y prescriptivo utilizan datos históricos para predecir qué sucederá en el futuro y qué pasos se deben tomar para influir en esos efectos (Sarker, 2024).



Diferencia de los tipos de análisis

El Desafío de la "Caja Negra" y la Necesidad de la IA Explicable (XAI)

A pesar del alto rendimiento predictivo de muchos modelos de Inteligencia Artificial (IA), especialmente los de aprendizaje profundo (*deep learning*), a menudo son considerados "cajas negras" (*black boxes*) por los expertos en ciberseguridad. Esto se debe

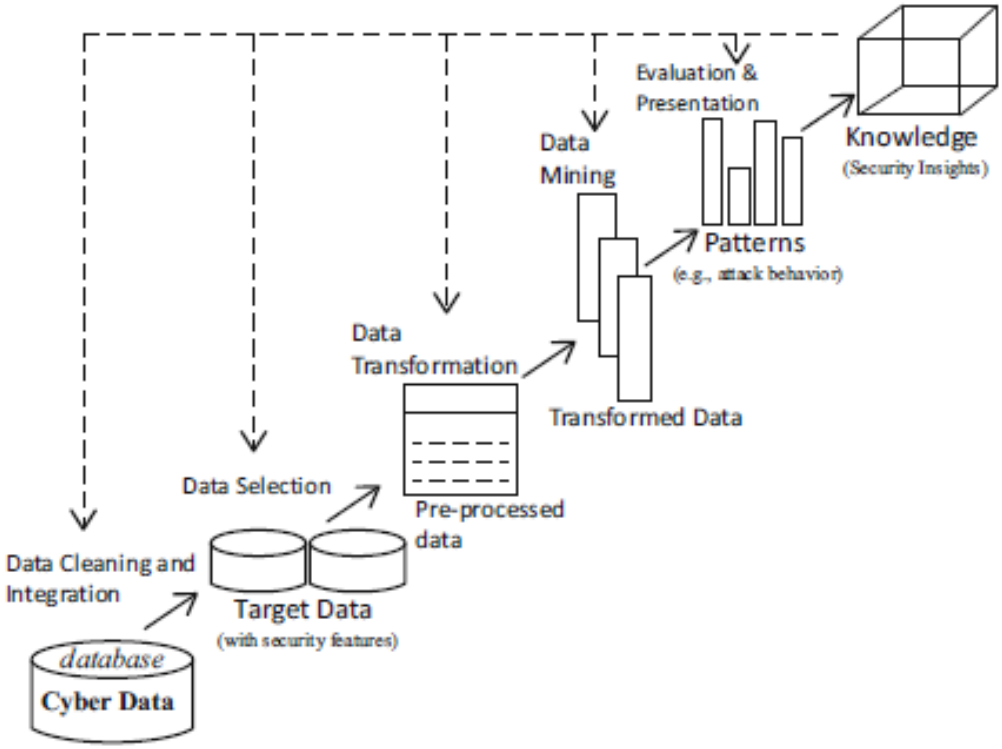
a que su funcionamiento interno es tan complejo que resulta difícil o imposible para los humanos comprender el razonamiento detrás de sus decisiones, lo que los hace incapaces de justificar sus resultados (Charmet et al., 2022).

Esta falta de transparencia es un problema crítico en la ciberseguridad, donde la confianza y la capacidad de verificar las decisiones son primordiales. La Inteligencia Artificial Explicable (XAI) ha surgido como un campo de investigación para abordar este desafío, con el objetivo de hacer que los modelos de IA sean interpretables. La explicabilidad es fundamental, ya que permite a los profesionales comprender las decisiones, a los usuarios finales confiar en ellas y a los desarrolladores mejorar los algoritmos identificando sus fortalezas y debilidades (Charmet et al., 2022).

Además, el cumplimiento de regulaciones legales, como la Ley de IA de la Comisión Europea, exige que los modelos de aprendizaje automático, especialmente en campos de alto riesgo, sean explicables y robustos (Calzarossa et al., 2025).

El Proceso de Modelado en la Ciencia de Datos

El modelado en ciencia de datos consiste en crear representaciones matemáticas o modelos basados en datos para hacer predicciones, clasificar información o descubrir patrones y conocimientos (Sarker, 2024).



Procedimiento general del proceso de descubrimiento de conocimiento

Comprensión del Problema de Negocio

Este es el paso fundamental del modelado. Implica colaborar con las partes interesadas para obtener una comprensión integral de los objetivos, desafíos e indicadores clave de rendimiento de la organización, sentando las bases para seleccionar técnicas de modelado apropiadas y adaptar el análisis para ofrecer conocimientos procesables (Sarker, 2024).

Comprensión de los Datos

La ciencia de datos es impulsada por la disponibilidad de datos, por lo que es esencial comprenderlos antes de construir un modelo. Los conjuntos de datos del mundo real a

menudo tienen ruido, valores faltantes o problemas de consistencia que deben manejarse eficazmente para obtener conocimientos valiosos (Sarker, 2024).

Preprocesamiento y Exploración de Datos

El análisis exploratorio de datos es un método para resumir conjuntos de datos con métodos visuales para identificar sus características clave y descubrir tendencias iniciales. Es importante limpiar y transformar los datos brutos antes del procesamiento para garantizar su calidad, lo que implica reformatear la información, corregir datos y fusionar conjuntos de datos (Sarker, 2024).

Modelado y Evaluación con Aprendizaje Automático

Este paso representa el núcleo del proceso analítico. Después de explorar y preprocesar los datos, implica seleccionar un algoritmo de aprendizaje automático apropiado. El modelo se entrena en un conjunto de datos específico y su efectividad se prueba rigurosamente utilizando métricas como exactitud, precisión, recall o F1-score (Sarker, 2024).

Producto de Datos y Automatización

Un resultado de la ciencia de datos es típicamente un producto de datos, como una predicción, un servicio o una herramienta que procesa datos y genera resultados. En este paso, un modelo se traduce en aplicaciones prácticas y se integra en sistemas automatizados para garantizar el análisis continuo, la adaptación y la entrega oportuna de resultados (Sarker, 2024).

Un Marco de 5 Capas para el Modelado Explicable en Ciberseguridad

Motivado por los procesos de descubrimiento de conocimiento y modelado de ciencia de datos, se propone una arquitectura de cinco capas para un modelado de ciberseguridad explicable y basado en reglas (Sarker, 2024).

Capa 1: Módulo de Recolección de Datos

Este módulo obtiene, agrega y refina diversos conjuntos de datos de ciberseguridad de fuentes como registros de red, eventos del sistema y fuentes de inteligencia de amenazas, priorizando las medidas de privacidad y seguridad para garantizar su idoneidad para la minería de reglas (Sarker, 2024).

Capa 2: Módulo de Preparación y Aumentación de Datos

Esta fase une la brecha entre los datos brutos y las reglas perspicaces, mejorando la calidad y cantidad del conjunto de datos mediante técnicas avanzadas de preprocesamiento y estrategias de aumentación de datos para diversificar artificialmente el conjunto de datos (Sarker, 2024).

Capa 3: Módulo de Minería de Reglas

Este módulo extrae reglas de seguridad procesables a partir de los conjuntos de datos preprocesados mediante algoritmos avanzados como la minería de reglas de asociación o la inducción de árboles de decisión. Las reglas generadas proporcionan transparencia para la toma de decisiones, mejorando la interpretabilidad (Sarker, 2024).

Capa 4: Módulo de Gestión de Reglas

Este módulo es responsable de organizar, optimizar y ajustar las reglas derivadas, priorizándolas según su relevancia e impacto. Monitorea continuamente el panorama de

ciberseguridad para admitir actualizaciones y ajustes dinámicos de las reglas (Sarker, 2024).

Capa 5: Módulo de Resultados Explicables

Este componente clave traduce las reglas derivadas a explicaciones legibles por humanos, proporcionando una justificación transparente de por qué se tomaron ciertas decisiones. Esto asegura que los resultados del modelo sean accesibles para las partes interesadas humanas (Sarker, 2024).

Tecnologías Complementarias: Blockchain para la Integridad de los Datos

La eficacia de los modelos de IA depende en gran medida de la calidad y la integridad de los datos de entrenamiento. Un desafío crítico es que un actor malicioso, como un proveedor de servicios en la nube, podría proporcionar información falsa (un ataque interno) para degradar la capacidad de detección de amenazas del modelo. Para abordar este problema, se propone la integración de Blockchain, que permite el intercambio inmutable de datos entre múltiples proveedores en la nube mediante un consenso como el Clique Proof-of-Authority (C-PoA). Este mecanismo garantiza que los datos utilizados para entrenar los modelos de IA sean auténticos y no hayan sido manipulados, proporcionando una base sólida y confiable para la detección de amenazas (Kumar et al., 2024).

Aplicaciones de Ciberseguridad Basadas en el Descubrimiento de Conocimiento

El descubrimiento de conocimiento y la minería de reglas proporcionan soluciones efectivas de ciberseguridad y, a su vez, sientan las bases para los análisis de explicabilidad,

ayudando a los profesionales a interpretar y mejorar la toma de decisiones automatizada (Sarker, 2024).

Detección de Anomalías o Intrusiones

El descubrimiento de conocimiento es crucial para detectar anomalías e intrusiones. Mediante el análisis de datos históricos, estas metodologías definen las operaciones normales del sistema y de la red. Cualquier desviación de estas reglas establecidas puede ser identificada como una anomalía, lo que permite detectar ataques como Denegación de Servicio Distribuido (DDoS), fingerprinting, amenazas internas o suplantación de identidad (spoofing) (Sarker, 2024). Este enfoque proactivo permite a las organizaciones identificar rápidamente irregularidades y accesos no autorizados, ayudando a prevenir brechas de seguridad antes de que ocurran (Sarker, 2024).

Categorización o Clasificación de Ataques

Estas técnicas son valiosas para clasificar ataques al analizar datos históricos para revelar patrones que indican metodologías de ataque distintas. Con base en el conocimiento histórico y reglas extraídas, los sistemas pueden clasificar los ataques en categorías específicas, como los basados en malware, denegación de servicio o phishing (Sarker, 2024). Un ejemplo destacado es la detección de phishing, donde los modelos de aprendizaje automático emplean características extraídas de las URL y del código fuente HTML de las páginas para diferenciar sitios web maliciosos de los legítimos (Calzarossa et al., 2025).

Predicción de Amenazas y Estrategias de Mitigación

El análisis de conjuntos de datos vastos y diversos revela patrones ocultos que pueden predecir amenazas y vulnerabilidades emergentes. Esta capacidad predictiva es vital para

adelantarse a los adversarios, permitiendo a las organizaciones fortalecer sus defensas y asignar recursos estratégicamente. Con base en este conocimiento, los equipos de ciberseguridad pueden desarrollar estrategias de mitigación efectivas, como aplicar parches o fortalecer protocolos, para neutralizar vulnerabilidades antes de que sean explotadas, cultivando así una postura de seguridad proactiva (Sarker, 2024).

Análisis Diagnóstico y Respuesta a Incidentes

Estas técnicas son cruciales para el análisis diagnóstico, ya que permiten la investigación de la causa raíz y la caracterización de los incidentes de seguridad mediante el análisis de datos. El conocimiento adquirido se utiliza para desarrollar reglas robustas que optimizan las estrategias de respuesta a incidentes, facilitando la identificación y clasificación rápida de las amenazas. La minería de reglas agiliza la respuesta al automatizar acciones basadas en condiciones predefinidas, lo que mejora la eficiencia en la contención y mitigación de incidentes y fortalece la resiliencia cibernética de la organización (Sarker, 2024).

La Seguridad de la Propia XAI

Si bien la XAI es una solución al problema de la caja negra, también introduce sus propios desafíos de seguridad, ya que las explicaciones pueden ser un objetivo para los atacantes. Se han identificado varias propiedades de seguridad de la XAI que pueden ser comprometidas (Charmet et al., 2022).

Equidad (Fairness)

Un modelo es justo si su resultado es irrelevante para las características sensibles de un individuo. Un atacante, generalmente el propietario del modelo, puede manipular las

explicaciones para ocultar sesgos discriminatorios, un ataque conocido como *fairwashing*, para engañar a un auditor haciéndole creer que el modelo es justo (Charmet et al., 2022).

Integridad

Las explicaciones pueden ser alteradas sutilmente para engañar a un experto humano sin que este se dé cuenta de la manipulación. Por ejemplo, se pueden generar explicaciones arbitrarias o no informativas para una predicción sin cambiar la predicción en sí (Charmet et al., 2022).

Privacidad

Los métodos de XAI pueden filtrar información sensible sobre los datos de entrenamiento a través de ataques de inversión de modelo (reconstruir datos privados a partir de las explicaciones) o ataques de inferencia de membresía (determinar si un individuo específico formó parte del conjunto de entrenamiento) (Charmet et al., 2022).

Robustez

Las explicaciones pueden ser frágiles, lo que significa que una pequeña perturbación en la entrada (como en un ejemplo adversario) puede cambiar drásticamente la explicación sin alterar la predicción del modelo, lo que socava la confianza en su fiabilidad (Charmet et al., 2022).

Conclusión

La ciberseguridad moderna es impulsada por la ciencia de datos, aprovechando análisis avanzados (descriptivo, diagnóstico, predictivo y prescriptivo) para crear modelos basados en datos que gestionen y mitiguen las ciberamenazas en constante evolución. No obstante, el alto rendimiento de los modelos de IA a menudo viene con el costo de ser "cajas negras", lo que hace que la Inteligencia Artificial Explicable (XAI) sea crucial para garantizar la transparencia, la confianza del usuario y el cumplimiento normativo, como lo exige la Ley de IA de la Comisión Europea. La XAI se implementa a través de marcos estructurados, como la arquitectura de cinco capas para el modelado explicable basado en reglas, pero su propia naturaleza introduce vulnerabilidades significativas que comprometen la privacidad, la integridad y la robustez de las explicaciones.

Referencias

Calzarossa, M. C., Giudici, P., & Zieni, R. (2025). An assessment framework for explainable AI with applications to cybersecurity. *Artificial Intelligence Review, 58*, 150. https://doi.org/10.1007/s10462-025-11141-w

Charmet, F., Tanuwidjaja, H. C., Ayoubi, S., Gimenez, P. F., Han, Y., Jmila, H., Blanc, G., Takahashi, T., & Zhang, Z. (2022). Explainable artificial intelligence for cybersecurity: A literature survey. *Annals of Telecommunications, 77*, 789–812. https://doi.org/10.1007/s12243-022-00926-7

Kumar, P., Javeed, D., Kumar, R., & Islam, A. K. M. N. (2024). Blockchain and explainable AI for enhanced decision making in cyber threat detection. *Software: Practice and Experience, 54*(8), 1337-1360. https://doi.org/10.1002/spe.3319

Sarker, I. H. (2024). Cybersecurity Data Science: Toward Advanced Analytics, Knowledge, and Rule Discovery for Explainable AI Modeling. En *AI-Driven Cybersecurity and Threat Intelligence* (pp. 101-118). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-031-54497-2_6