Universidad Tecnológica Nacional
AJEA – Actas de Jornadas y Eventos Académicos de UTN
Trabajo Completo Publicado en AJEA – Full Text Published in AJEA
https://rtyc.utn.edu.ar/index.php/ajea/article/view/1629
Libro de actas de resúmenes DOI: https://doi.org/10.33414/ajea.1640.2024
Colección JIT en AJEA https://rtyc.utn.edu.ar/index.php/ajea/JITs

Jóvenes Investigadores Tecnológicos
Facultad Regional Rafaela 26 y 27 de octubre de 2023

Área: Medio ambiente, contingencias y desarrollo sustentable / Categoría: Estudiante de grado / Regional: Facultad Regional Rafaela

Detección de intrusiones basados en firmas con Minería de Datos y Machine Learning.

Signature-based Intrusion Detection with Data Mining and Machine Learning.

Presentación: 17/10/2023

Guillermo Dolan
Universidad Tecnológica Nacional Facultad Regional Rosario Argentina
guillermopatdolan@gmail.com
Bautista Guerra
Universidad Tecnológica Nacional Facultad Regional Rosario Argentina
bautistaguerra.it@gmail.com
Ornella Colazo
Universidad Tecnológica Nacional Facultad Regional Rosario Argentina
ornecolazo@gmail.com
Maximiliano Mansilla
Universidad Tecnológica Nacional Facultad Regional Rosario Argentina
mmansilla02@outlook.com
Lucía Morena Fabbri
Universidad Tecnológica Nacional Facultad Regional Rosario Argentina
fabbriluciam@gmail.com

Resumen

Una de las problemáticas en torno a la seguridad de la información a las que estamos expuestos el día de hoy son las amenazas cibernéticas. Cada sitio web en el que navegamos o aplicación que se conecta a internet que utilizamos está expuesta a riesgos que afectan a los datos sensibles que son almacenados o manipulados.

Las empresas que brindan soluciones de software deben de resguardar la seguridad que ofrecen sus soluciones y de respetar las Leyes que las obligan a cuidar el tratamiento de los datos sensibles.

Siendo muchas las amenazas cibernéticas, nos encontramos ante la necesidad de evaluar todo el ciclo en que se materializan los ataques informáticos. En particular, el presente trabajo se enfoca en realizar un análisis de una posible manera de aportar seguridad a los sistemas en cuanto a las primeras fases de un ataque informático: Investigación y recopilación de información y obtención de acceso.

Palabras clave: Minería de Datos - Machine Learning - Sistemas de Detección de Intrusiones - Amenazas informáticas

Abstract

AJEA – Actas de Jornadas y Eventos Académicos de UTN
Libro de actas de resúmenes DOI: https://doi.org/10.33414/ajea.1640.2024
Texto Completo Publicado en AJEA – Full Text Published in AJEA
https://rtyc.utn.edu.ar/index.php/ajea/article/view/1629

One of the issues surrounding information security that we face today is cyber threats. Every website we navigate or internet-connected application we use is exposed to risks that affect sensitive data that is stored or manipulated.

Companies providing software solutions are challenged to ensure the security of their offerings and comply with laws that require them to protect the handling of sensitive data.

Compliance with data security or confidential information requirements will not only allow compliance with the mentioned laws but also instill confidence in the customers for whom a specific solution is developed or provided, or for the specific audience it targets.

With many cyber threats, we need to assess the entire cycle in which cyber-attacks materialize. In particular, this work focuses on analyzing a possible way to contribute to the security of systems regarding the early stages of a cyber-attack: Research and information gathering and gaining access.

Specifically, the implementation of Signature-Based Intrusion Detection Systems will be analyzed, using Data Mining and Machine Learning.

Keyword: Data Mining – Machine Learning – Intrusion Detection System – Cyber Threat

Introducción

En actualidad se habla de amenazas informáticas como una fuente o causa potencial de eventos o incidentes no deseados que pueden resultar en daño a los insumos informáticos de la organización y ulteriormente a ella misma (Zambrano, S. M. Q., & Valencia, D. G. M. 2017). Pero en realidad no se tiene en cuenta la importancia que se le debería dar a dichas amenazas, dado que la protección contra amenazas informáticas muchas veces es dejada de lado por cuestiones de tiempo y costos.

Las empresas que brindan soluciones de Software en Argentina deben ser conscientes de cuan importante es la seguridad que deben de implementar a los productos o servicios que brindan, para cumplir con toda la normativa legal que se les exige, como por ejemplo la Regulación General de Protección de Datos Europea (The Council of the European Union, 2016) para aquellos software que desarrollan para el exterior, y la Ley 25.326 Ley de Protección de Datos Personales (Infoleg, s.f.) para el caso de software que sea desarrollado para Argentina, entre otras, como así también brindar confianza a los clientes para quienes desarrollan o brindan soluciones.

Como una manera de aportar a la seguridad de las soluciones que desarrollan estas empresas de Software, nos enfocaremos a hablar de los Sistemas de Detección de Intrusiones (Intrusion Detection System (IDS) en Inglés), que son sistemas de supervisión que detectan actividades sospechosas y generan alertas al detectarlas (Check Point, s.f.). Tomará relevancia para nuestro trabajo los Sistemas de Detección de Intrusiones Basados en Firmas (Signature-Based Detection).

En la actualidad técnicas avanzadas como el Machine Learning (Aprendizaje Automático), ciencia de desarrollo de algoritmos y modelos estadísticos que utilizan los sistemas de computación con el fin de llevar a cabo tareas sin instrucciones explícitas (Amazon Web Services, s.f.), y el Data Mining (Minería de Datos) proceso técnico, automático o semiautomático, que analiza grandes cantidades de información dispersa para darle sentido y convertirla en conocimiento (Iberdrola, s.f.), están siendo

AJEA – Actas de Jornadas y Eventos Académicos de UTN
Libro de actas de resúmenes DOI: https://doi.org/10.33414/ajea.1640.2024
Texto Completo Publicado en AJEA – Full Text Published in AJEA
https://rtyc.utn.edu.ar/index.php/ajea/article/view/1629

aplicadas en varios casos de uso. A continuación se citan algunos ejemplos: Detección de Spam, Antivirus, Pronósticos del clima, Vehículos autónomos y robots, análisis de imágenes HD (Iberdrola. s.f.), Comercio, Medicina, Astronomía, Geología (ESIC, s.f.), por mencionar solo algunos. Pero además tanto el Machine Learning como el Data Mining, están siendo aplicados también en ámbitos de seguridad informática, para combatir las amenazas que se encuentren en la Internet.

Por ello y ya mencionado anteriormente, nos enfocaremos en los IDS basados en firmas, para ver una manera de cómo haciendo uso del Data Mining y el Machine Learning, se obtiene un enfoque para la detección de amenazas informáticas.

El resto del trabajo está organizado de la siguiente manera: en la sección 2 se mencionan las técnicas de Minería de Datos, en la sección 3 se menciona a las categorías de Machine Learning, en la sección 4 se explica el proceso de detección de amenazas con Data Mining y Machine Learning, en la sección 5 explicamos las ventajas y desventajas de la detección de intrusiones basadas en firmas, contrastándola con la detección basada en anomalías, en la sección 6 hacemos agradecimientos a quienes colaboraron a nuestro trabajo, y en la sección 7 dejamos una breve conclusión del trabajo realizado.

Técnicas de Minería de Datos

Actualmente en la Minería de Datos existen tres tipos de técnicas, las cuales se ajustan a las necesidades de la tarea que se requiera realizar: técnicas descriptivas, predictivas, prescriptivas (Amazon Web Services, s.f.).

Técnicas descriptivas:

En cuanto a las técnicas descriptivas tenemos dos: Minería de reglas de asociación, Agrupación en Clusters. Con respecto a la primera, la podemos definir como el proceso de encontrar relaciones entre dos conjuntos de datos diferentes y aparentemente no relacionados (Amazon Web Services. s.f.). La segunda consiste en agrupar varios puntos de datos en función de sus similitudes y el resultado de la minería de datos es un conjunto de clústeres en el que cada colección es distinta de otros grupos, pero los objetos de cada clúster son similares de alguna manera (Amazon Web Services. s.f.).

Técnicas predictivas:

En cuanto a las técnicas predictivas vamos a hacer mención de tres: la primera es la Clasificación, técnica compleja de minería de datos que utiliza ejemplos para entrenar un modelo de Machine Learning, para clasificar los datos en distintas categorías. Utiliza métodos estadísticos, como árboles de decisión (Amazon Web Services. s.f.). En segundo lugar tenemos la regresión que sirve para ubicar relaciones y calcular probabilidades con base en datos. Esto significa que puede utilizarse para predecir valores numéricos (HubSpot. s.f.). Y por último mencionamos a la detección de anomalías que se encarga de detectar valores atípicos a través del rastreo o clasificación de datos. En algunos casos, los algoritmos pueden detectar irregularidades y predecir su resultado o las consecuencias, gracias al aprendizaje obtenido de otros casos similares (HubSpot. s.f.).

Técnicas prescriptivas:

En cuanto a las técnicas prescriptivas, se las llama así dado que establecen reglas o comandos dependiendo de los resultados del análisis de la información. Aquí poseemos dos tipos de técnicas descriptivas: Automatización, Optimización. Dentro del primer tipo tenemos a los árboles de decisión, que son modelos predictivos y de aprendizaje automático que generan

AJEA – Actas de Jornadas y Eventos Académicos de UTN
Libro de actas de resúmenes DOI: https://doi.org/10.33414/ajea.1640.2024
Texto Completo Publicado en AJEA – Full Text Published in AJEA
https://rtyc.utn.edu.ar/index.php/ajea/article/view/1629

respuestas a ciertos problemas, cediendo responsabilidades a las tecnologías (HubSpot. s.f.). En cuanto a las técnicas de Optimización, las mismas generan simulaciones para la toma de decisiones frente al resultado de una analítica de los datos, por lo tanto, obtienen una mejor respuesta basada en casos anteriores (HubSpot. s.f.).

Categorías de Machine Learning

Dentro del Machine Learning tenemos diferentes categorías de aprendizaje: supervisado, no supervisado, y de refuerzo. A continuación hacemos una breve descripción de las mismas:

Aprendizaje supervisado:

El aprendizaje supervisado comienza con un conjunto establecido de datos y una cierta comprensión de cómo se clasifican estos datos. El aprendizaje supervisado tiene la intención de encontrar patrones en datos que se pueden aplicar a un proceso de análisis. Estos datos tienen características etiquetadas que definen el significado de los datos (IBM. s.f.). Algunos algoritmos de este tipo de aprendizaje son: Decision Table, Naive Bayes (Raona. s.f.), Random Forest (Espinosa-Zúñiga, J. J. 2020).

Aprendizaje no supervisado:

El aprendizaje no supervisado se utiliza cuando el problema requiere una cantidad masiva de datos sin etiquetar, analizándolos sin intervención humana, mediante un proceso iterativo (IBM. s.f.).

Aprendizaje de refuerzo:

El aprendizaje de refuerzo es un modelo de aprendizaje conductual. El algoritmo recibe retroalimentación del análisis de datos, conduciendo el usuario hacia el mejor resultado. El aprendizaje de refuerzo difiere de otros tipos de aprendizaje supervisado, porque el sistema no está entrenado con el conjunto de datos de ejemplo. Más bien, el sistema aprende a través de la prueba y el error (IBM. s.f.).

Detectando amenazas con Data Mining y Machine Learning

Como ya fue mencionado al inicio de este trabajo, nos enfocamos en los IDS basados en firmas. Aquí es donde entran en juego tanto las técnicas de Data Mining como de Machine Learning.

Los Sistemas de Detección de Intrusiones Basados en Firmas poseen la característica que cuando se produce un ciberataque, se procede al tratamiento del mismo, y se obtiene una firma. Una firma es un patrón o cadena que corresponde a un ataque o amenaza conocidos (Liao et al. 2013), con lo que las técnicas de IDS basadas en firmas funcionaran cuando exista una firma de intrusión en la base de datos del IDS (Rene, C. I., & Abdullah, J. 2017).

Por ello, los Sistemas de Detección de Intrusiones Basados en Firmas son efectivos para detectar tipos conocidos de ataques sin generar un número abrumador de falsas alarmas, en contraste con los Sistemas de Detección de Intrusiones Basados en Anomalías, que generan un gran número de falsos positivos (Rene, C. I., & Abdullah, J. 2017).

Metodología

AJEA – Actas de Jornadas y Eventos Académicos de UTN
Libro de actas de resúmenes DOI: https://doi.org/10.33414/ajea.1640.2024
Texto Completo Publicado en AJEA – Full Text Published in AJEA
https://rtyc.utn.edu.ar/index.php/ajea/article/view/1629

Para la detección de intrusiones basadas en firmas es necesario realizar un completo análisis de tráfico de la red para que se logre la máxima efectividad posible. Cuando se lleva a cabo el análisis mencionado, se obtiene información de múltiples fuentes, como son registros del sistema, de aplicaciones, mensajes de alarma, entre otros (Patel, J., & Panchal, K. 2015). Por ello, con toda la información obtenida, poseemos diferentes fuentes de datos, formatos, y mucha redundancia, y necesitamos descubrir propiedades, patrones, relaciones en los datos que hasta el momento eran desconocidas, como así también extraer características, con lo que entonces entra en juego la Minería de Datos, puesto que la misma tiene una gran ventaja en extracción de datos en grandes volúmenes, como este es el caso (Patel, J., & Panchal, K. 2015). Aquí se aplican, por un lado la técnica descriptiva de Agrupación en Clusters, y por el otro, técnicas predictivas como la clasificación y regresión (Patel, J., & Panchal, K. 2015).

Realizado todo el proceso de análisis y extracción de características, el siguiente paso consiste en la clasificación de si el contenido que se estuvo analizando es benigno o malicioso.

Por ello es que se procede a usar Machine Learning, aplicando diversos algoritmos como lo son: Decision Table, Naive Bayes, Random Forest (Almseidin, M., Alzubi, M., Kovacs, S., & Alkasassbeh, M. 2017).

En la figura 1 podemos observar un resumen de todo el proceso descrito anteriormente.

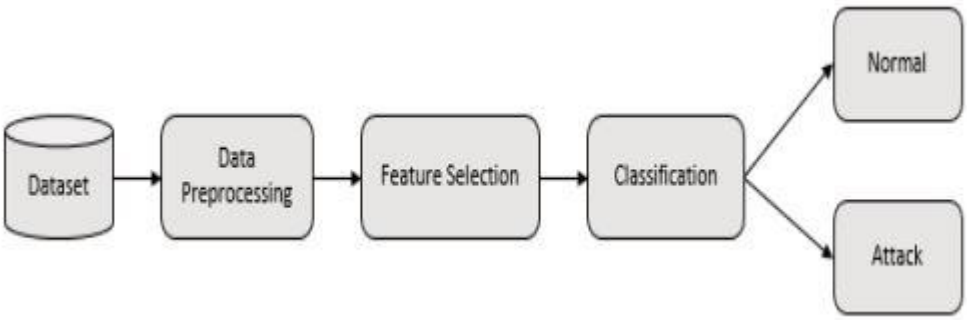

Figura 1. Resumen del proceso de aplicación de Minería de Datos y Machine Learning (Patel, J., & Panchal, K. 2015)

Ventajas y Desventajas de los Sistemas de Detección de Intrusiones Basada en Firmas

Desde hace un tiempo que la Internet está repleta de amenazas cibernéticas creadas por ciberdelincuentes que buscar obtener información confidencial, robo de dinero, espiar a la competencia en el caso de una organización que busca espiar a otra, entre otras cosas. Pero los años fueron pasando y las amenazas cibernéticas fueron evolucionando, con lo que hoy se debe de estar alerta por nuevos posibles ataques. Por ello haremos mención de algunas ventajas y desventajas del Sistema de Detección de Intrusiones Basados en Firmas.

Una de las desventajas que posee es que no puede detectar ataques novedosos o de 'día cero' (Buczak, A. L., & Guven, E. 2015) que son los más presentes al día de hoy, con lo que pone en riesgo a toda una organización. Además, cumple una función reactiva y no proactiva puesto que no puede prevenir ataques (Protecciondatos-lopd.com. s.f.), con lo que éstas son algunas de las causas por las cuales hoy en día se apuesta por un Sistema de Detección de Intrusiones Basado en Anomalías. Sin embargo, y ya dicho en la sección 3, el Sistema de Detección de Intrusiones Basado en Firmas tiene una tasa muy baja de generar falsos positivos, con lo que lo consideramos una ventaja, y esto evita que se disparen alarmas al azar, o en muy grandes cantidades.

AJEA – Actas de Jornadas y Eventos Académicos de UTN
Libro de actas de resúmenes DOI: https://doi.org/10.33414/ajea.1640.2024
Texto Completo Publicado en AJEA – Full Text Published in AJEA
https://rtyc.utn.edu.ar/index.php/ajea/article/view/1629

Conclusión

Visto y considerando que los Sistemas de Detección de Intrusiones Basados en Firmas poseen varias desventajas, creemos que a pesar de eso, podrían ser una manera más de aportar a la seguridad que poseen las soluciones que brindan las empresas desarrolladoras de Software. Hoy la manera de brindar seguridad está muy ligada y depende de cuan peligrosas sean las amenazas cibernéticas que sean creadas por ciberdelincuentes.

Es tarea de las empresas que brindan soluciones de software de llevar a cabo tareas de Investigación, no solo para estar al día con las amenazas que se encuentran en la Internet, sino además de encontrar manera de escoger la mejor solución en cuanto a seguridad, y así implementarlas a sus software.

Por otro lado en revisión de la bibliografía consultada acerca de los Sistemas de Intrusiones Basados en Firmas surgen una serie de interrogantes: ¿Cuánto tiempo más se seguirán usando estos IDS?, ¿Desaparecerán en un lapso de tiempo no muy lejano?, ¿Se implementarán otros algoritmos de Machine Learning para la detección de intrusiones que mejorarán el modo de detección de estos IDS?

La primera pregunta surge dado que estamos en una constante y veloz evolución de amenazas cibernéticas, y estos IDS no son lo más adecuado si se quiere estar al día para detectar nuevos ataques o ataques de día cero (zero-day). La segunda pregunta surge dado que creemos que habrá más tendencia de uso hacia los IDS basados en anomalías ya que se necesita y se necesitará un enfoque de detección de tiempo real, y estar preparados para los nuevos ataques que puedan llegar a ocasionarse. La tercera pregunta nos surge dado que hoy solo se utilizan algunos algoritmos de Machine Learning, pero podría darse el caso de que el día de mañana quizá se puedan mejorar estos Sistemas de Intrusiones Basados en Firmas con la ayuda de otras técnicas de esta ciencia, haciendo que no se dejen en desuso y se sigan implementando. Ahora solo resta esperar cómo van a evolucionar los Sistemas de Detección de Intrusiones, y si se alinearán de la manera más efectiva posible para combatir las amenazas en la Internet.

Agradecimientos

El presente trabajo fue realizado en el contexto del proyecto: TOECRO0008583 - ”Modelización de un Sistema de Diagnóstico de Riesgos de Seguridad de la Información para su Integración a Sistemas de Gestión de Calidad” de la Universidad Tecnológica Nacional radicado en la Facultad Regional Rosario, presentado en el CoNaIISI 2022 realizado en la ciudad de Concepción del Uruguay, los días 3 y 4 de Noviembre.

Agradecemos además a nuestro equipo de Investigación conformado por docentes y alumnos que siempre está para las dudas que nos van surgiendo, y para asesorarnos con la experiencia que ellos poseen.

Referencias bibliografías

Zambrano, S. M. Q., & Valencia, D. G. M. (2017). Seguridad en informática: consideraciones. Dominio de las Ciencias, 3(3), 676–688.

The Council of the European Union. (2016). The general data protection regulation. [En línea] Disponible en: https://www.consilium.europa.eu/en/policies/data-protection/data-protection-

AJEA – Actas de Jornadas y Eventos Académicos de UTN
Libro de actas de resúmenes DOI: https://doi.org/10.33414/ajea.1640.2024
Texto Completo Publicado en AJEA – Full Text Published in AJEA
https://rtyc.utn.edu.ar/index.php/ajea/article/view/1629

regulation/#:~:text=The%20GDPR%20establishes%20the%20general,data%20processing%20operations%20they%20perform.

Infoleg. (s.f.). Protección de los datos personales. [En línea] Disponible en: https://servicios.infoleg.gob.ar/infolegInternet/anexos/60000-64999/64790/texact.htm.

Check Point. (s.f.). ¿Qué es un sistema de detección de intrusos (IDS)?. [En línea] Disponible en: https://www.checkpoint.com/es/cyber-hub/what-is-an-intrusion-detection-system-ids/

Amazon Web Services. (s.f.). ¿Qué es el machine learning?. [En línea] Disponible en: https://aws.amazon.com/es/what-is/machine-learning/.

Iberdrola. (s.f.). ¿Qué es el data mining?. [En línea] Disponible en: https://www.iberdrola.com/innovacion/data-mining-definicion-ejemplos-y-aplicaciones#:~:text=%C2%BFQU%C3%89%20ES%20EL%20'DATA%20MINING,sentido%20y%20convertirla%20en%20conocimiento.

Bedu. (s.f.). ¿Qué es machine learning y en dónde se aplica?. [En línea] Disponible en: https://bedu.org/blog/tecnologia/que-es-machine-learning-y-en-donde-se-aplica.

ESIC. (s.f.). Minería de datos: qué es, como es el proceso y a qué áreas se puede aplicar. [En línea] Disponible en: https://www.esic.edu/rethink/tecnologia/mineria-datos-proceso-areas-se-puede-aplica.

Amazon Web Services. (s.f.). ¿Qué técnicas de mineria de datos existen?. [En línea] Disponible en: https://aws.amazon.com/es/what-is/data-mining/#:~:text=La%20miner%C3%ADa%20de%20datos%20es,relaciones%20ocultas%20en%20sus%20datos.

Amazon Web Services. (s.f.). Minería de reglas de asociación. [En línea] Disponible en: https://aws.amazon.com/es/what-is/data-mining/#:~:text=La%20miner%C3%ADa%20de%20datos%20es,relaciones%20ocultas%20en%20sus%20datos.

Amazon Web Services. (s.f.). Agrupación en clústeres. [En línea] Disponible en: https://aws.amazon.com/es/what-is/data-mining/#:~:text=La%20miner%C3%ADa%20de%20datos%20es,relaciones%20ocultas%20en%20sus%20datos.

Amazon Web Services. (s.f.). Clasificación. [En línea] Disponible en: https://aws.amazon.com/es/what-is/data-mining/#:text=La%20miner%C3%ADa%20de%20datos%20es,relaciones%20ocultas%20en%20sus%20datos.

AJEA – Actas de Jornadas y Eventos Académicos de UTN
Libro de actas de resúmenes DOI: https://doi.org/10.33414/ajea.1640.2024
Texto Completo Publicado en AJEA – Full Text Published in AJEA
https://rtyc.utn.edu.ar/index.php/ajea/article/view/1629

HubSpot. (s.f.). Técnicas de regresión. [En línea] Disponible en: https://blog.hubspot.es/marketing/mineria-datos.

HubSpot. (s.f.). Técnicas de detección de anomalías. [En línea] Disponible en: https://blog.hubspot.es/marketing/mineria-datos.

IBM. (s.f.). Aprendizaje supervisado. [En línea] Disponible en: https://www.ibm.com/mx-es/analytics/machine-learning.

Raona. (s.f.). Los 10 algoritmos esenciales en machine learning. [En línea] Disponible en: https://www.raona.com/los-10-algoritmos-esenciales-machine-learning/.

Espinosa-Zúñiga, J. J. (2020). Aplicación de algoritmos random forest y xgboost en una base de solicitudes de tarjetas de crédito. Ingeniería, investigación y tecnología, 21(3).

IBM. (s.f.). Aprendizaje no supervisado. [En línea] Disponible en: https://www.ibm.com/mx-es/analytics/machine-learning.

IBM. (s.f.). Aprendizaje de refuerzo. [En línea] Disponible en: https://www.ibm.com/mx-es/analytics/machine-learning.

Liao, H.-J., Lin, C.-H. R., Lin, Y.-C., & Tung, K.-Y. (2013). Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications, 36(1), 16–24.

Rene, C. I., & Abdullah, J. (2017). Malicious code intrusion detection using machine learning and indicators of compromise. International Journal of Computer Science and Information Security (IJCSIS), 15(9).

Patel, J., & Panchal, K. (2015). Effective intrusion detection system using data mining technique. Journal of Emerging Technologies and Innovative Research, 2(6), 1869–1878.

Almseidin, M., Alzubi, M., Kovacs, S., & Alkasassbeh, M. (2017). Evaluation of machine learning algorithms for intrusion detection system. En 2017 IEEE 15th international

Buczak, A. L., & Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153–1176.

Protecciondatos-lopd.com. (s.f.). Ventajas y desventajas de un sistema de detección de intrusos. [En línea] Disponible en: https://protecciondatos-lopd.com/empresas/sistema-deteccion-intrusiones-ids/#Ventajas_y_desventajas_de_un_sistema_de_deteccion_de_intrusos.