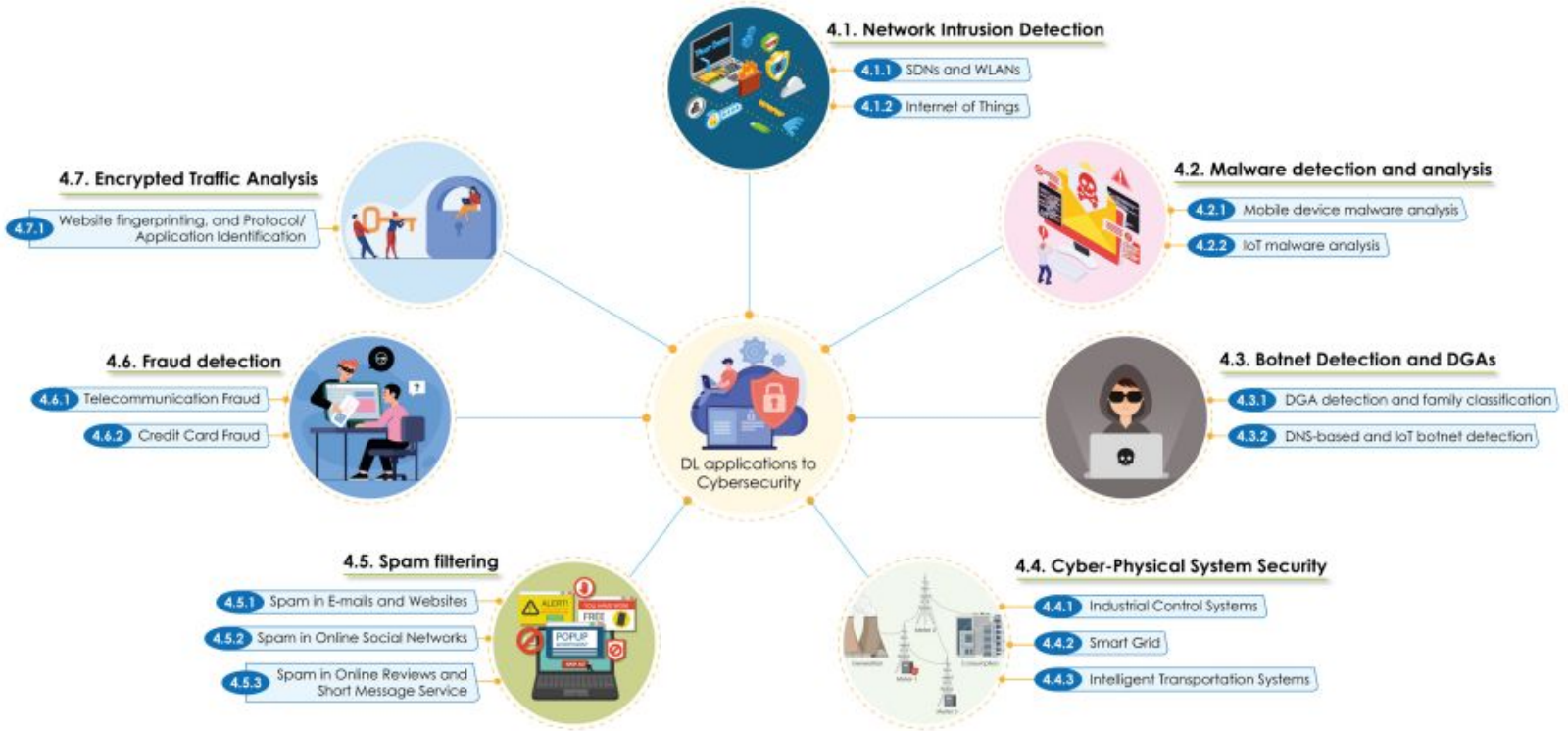
Machine Learning y Seguridad: Detección de Intrusiones en Redes

Juan José Acevedo Serna
Carlos Andrés Rodríguez Torres
Óliver Eduardo Chan Dorado

Contenido

Introducción y Estado del Arte

Introducción a la Ciberseguridad y Técnicas Tradicionales



Detección de Intrusiones en la Red (IDS)

- **Monitoreo y análisis** de eventos en un sistema o red para **detectar incidentes de seguridad**.
- **Aprendizaje automático** donde se modelan el comportamiento normal de usuarios, aplicaciones y tráfico para **detectar desviaciones significativas**.

Detalles Técnicos

Recolección de Datos

Dataset del tráfico en la red

- KDD CUP '99
- NSL-KDD
- DARPA
- CICIDS-2017

Extracción de características

- Duración de conexión
- Bytes y paquetes enviados/recibidos
- Protocolo (TCP, UDP, ICMP)
- Comportamientos estadísticos

Etiquetado

- BENIGNO
- MALIGNO
  - DoS Hulk
  - DDoS
  - PortScan
  - Brute Force
  - Web Attack
  - Infiltration
  - Bot
  - Heartbleed

Preprocesamiento

Limpieza de Datos	Características Categóricas	Normalización	División de Datos
Eliminar entradas incompletas, incorrectas o innecesarias. **Eliminar**: valores nulos, duplicados y datos inválidos. Asegura la calidad y fiabilidad del conjunto de datos.	Convertir categorías a valores **numéricos** para que puedan ser procesados por el modelo. Por ejemplo, convertir protocolos (TCP, UDP, ICMP) en números (One-Hot Encoding o Label Encoding).	**Normalización/Estandarización**: usar MinMaxScaler, StandardScaler para escalar las características numéricas.	Dividir el dataset en 3 subconjuntos: ● Entrenamiento ● Validación ● Test

Entrenamiento del Modelo

Random Forest

Utiliza **múltiples árboles de decisión** para generar predicciones. La decisión final se toma con base en la **mayoría** de las predicciones de los árboles.

Support Vector Machine

Utiliza un **hiperplano** en un espacio de alta dimensión para clasificar muestras. La clasificación se realiza encontrando el margen más grande entre las muestras y el hiperplano

Gaussian Naive Bayes

Es un **modelo probabilístico**, asume **independencia** entre atributos dentro de cada clase. Es **rápido** y **efectivo** para clasificar datos en **tiempo real**.

Entrenamiento del Modelo

Regresión Logística

Este modelo se utiliza para **predecir la probabilidad** de que un evento ocurra. Es simple y efectivo para problemas de **clasificación binaria** en ciberseguridad.

Deep Neural Networks

Las DNN son capaces de **aprender representaciones complejas** de datos. Su uso en la detección de intrusiones permite identificar **patrones** sutiles en **grandes conjuntos** de datos.

Pruebas y Evaluación del Modelo

Luego de entrenar el modelo se **evalúan** con un **conjunto de prueba** no visto previamente, simulando escenarios reales. Las métricas clave son el **recall**, para asegurar la detección efectiva de ataques (**baja tasa de falsos negativos**), y la precisión, para evitar falsos positivos en el tráfico benigno.

Benign → 0,

DoS → 1,

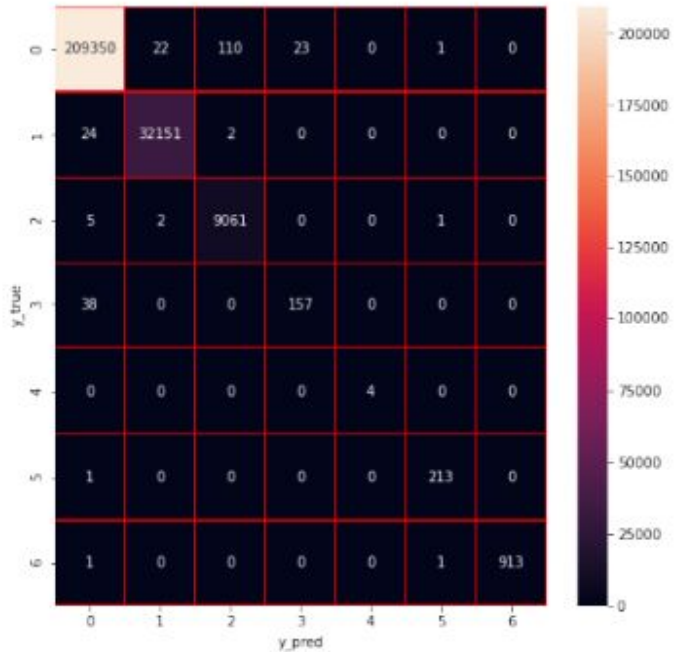
Port Scan → 2,

Bot → 3,

Infiltration → 4,

Web Attack → 5,

Brute Force → 6.



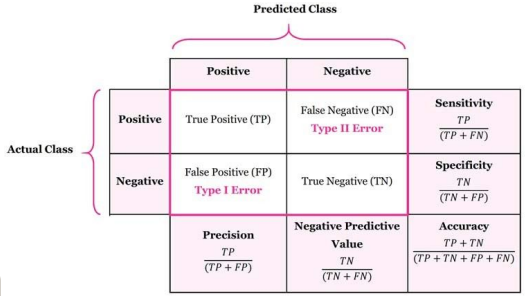
(a) Random Forest (RF)

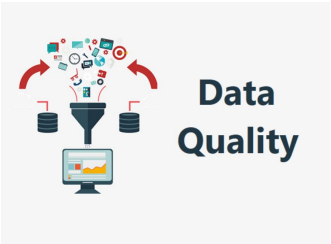
Falsos positivos (FP) son los casos donde el modelo **predijo "ataque" (clase 1–6)**, pero **en realidad era tráfico benigno (clase 0)**.

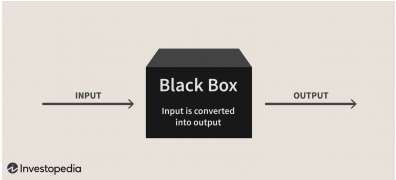
Problemas y Soluciones

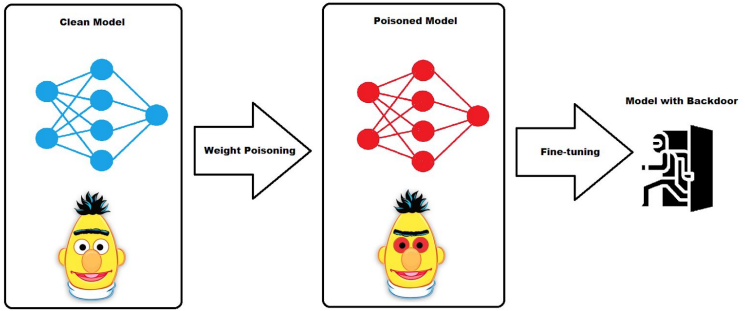
Principales Problemas

- Alta tasa de Falsos Positivos en ABIDS.
- Falta de Interpretabilidad en ABIDS.
- Calidad de los Datos.
- Ataques Adversariales.
- Envenenamiento de los datos.









Alta Tasa Falsos Positivos en ABIDS

Alta Tasa Falsos Positivos en ABIDS

Optimización de Umbrales y Puntuación de Riesgo

↓

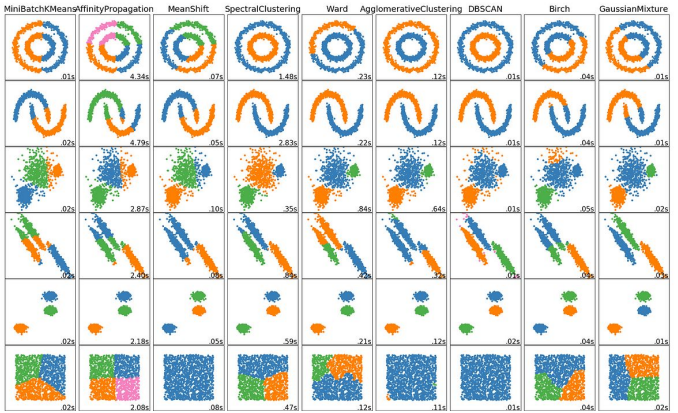
Sistemas Híbridos

↓

Aprendizaje Activo y Supervisión Humana

↓

Técnicas de Agrupamiento y Correlación de Eventos



Uso en el Mercado & Futuro

Adopción en la Industria

- Security Operations Center.
- Plantas de Energía.
- Sistemas Hospitalarios.
- Bancos.

DARKTRACE

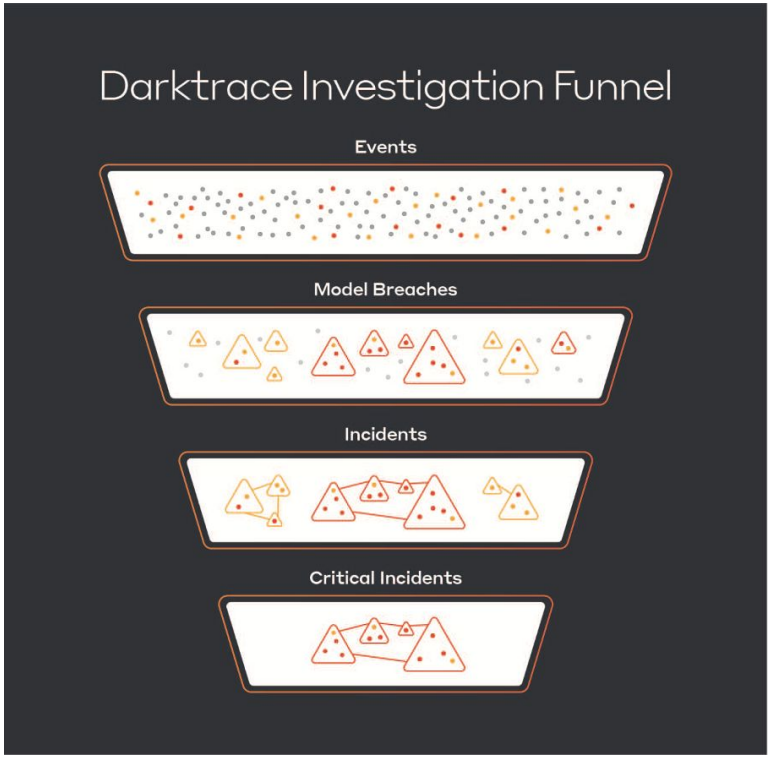
splunk>

paloalto NETWORKS

Hacia un SOC más inteligente

- Integración con **aprendizaje** por **refuerzo**.
- Análisis por **multicapas**
- **Aprendizaje Federativo.**
- Automatización con **revisión humana adaptativa**.
- **Modelos** específicos por **contexto**.

Ejemplo: Alta Tasa Falsos Positivos



¿Qué hicieron?

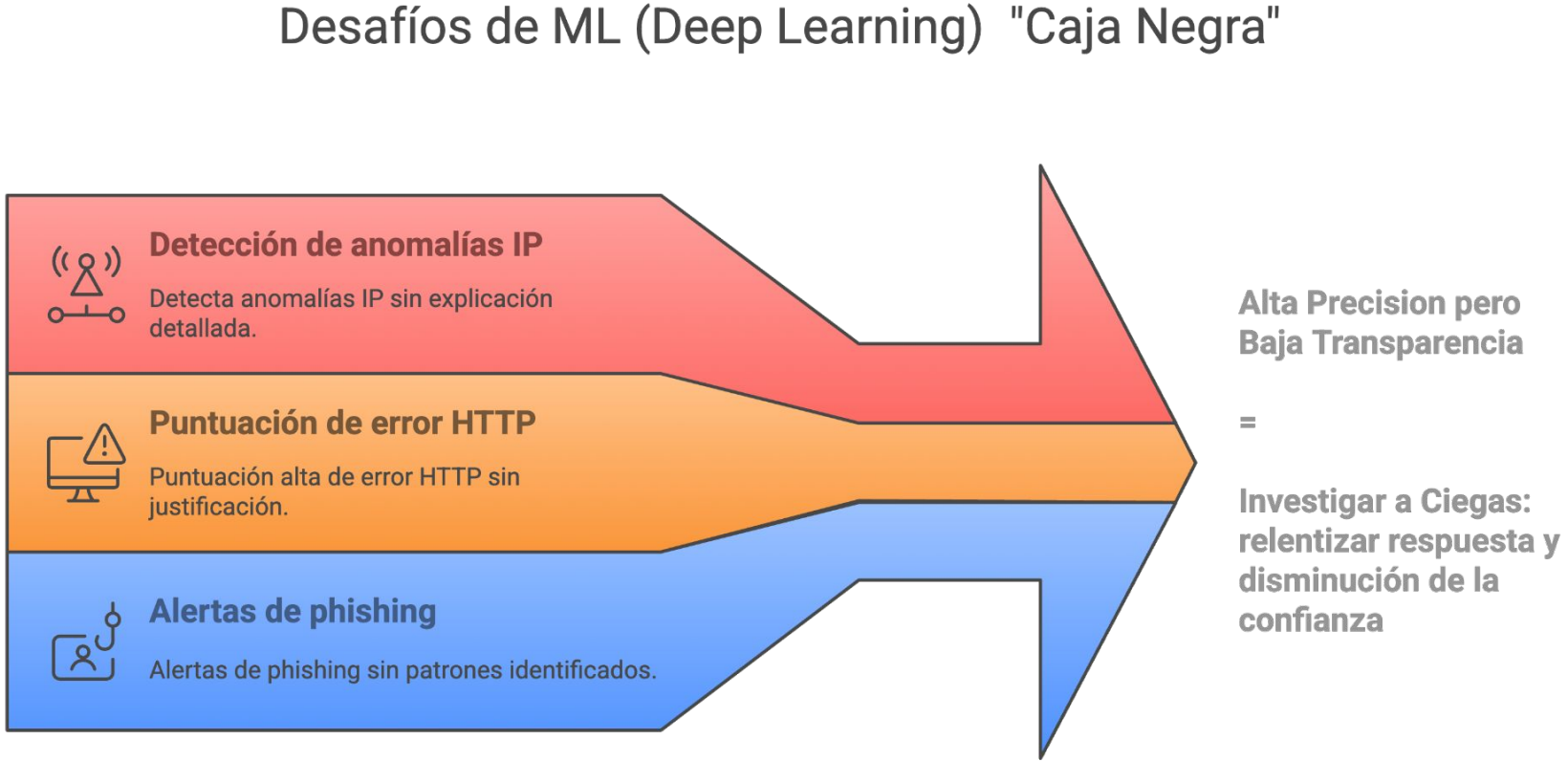
- Desplegaron Darktrace para **agrupar** y **priorizar eventos**.
- Configuraron **thresholds** basados en aprendizaje continuo.
- Integraron **feedback** de **analistas** para ajustar baseline.

Resultados

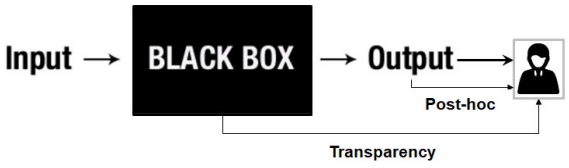
- **Reducción exponencial** de alertas (miles a decenas).
- Tasa de FP cayó debajo de 20%.
- Mayor **enfoque** en **incidentes reales**.

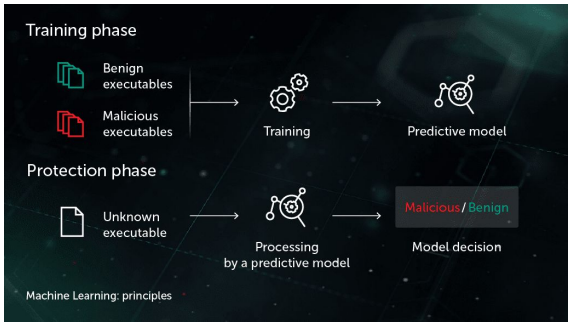
Falta de Interpretabilidad en ABIDS

Falta de Interpretabilidad en ABIDS - Problemas



Falta de Interpretabilidad en ABIDS - Soluciones





Modelos Explicables por Diseño

Entendibles pero con Menor Precisión: árboles, regresión, etc.

Técnicas de Explainable AI (XAI)

LIME: modelos simples locales
SHAP: contribución por característica
Grad-CAM y atención visual: zonas de activación de red

Sistemas Híbridos

Modelos complejos pero con capa de explicación (red neuronal + SHAP)

Técnicas de Reportes Automáticos

Traducción de decisiones del modelo a lenguaje natural

Uso en el Mercado & Futuro

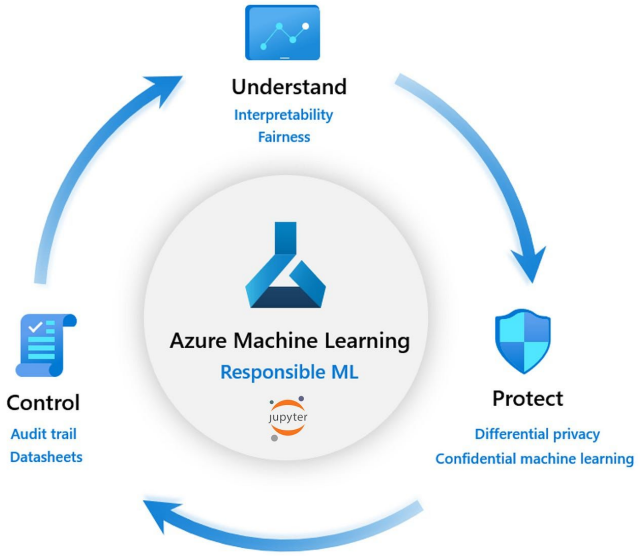
Falta de Interpretabilidad y Adopción en la Industria

- Security Operations Center (justificar acciones)
- Sectores Regulados: salud, finanzas, etc. (auditorías y responsabilidades)
- Defensa y Gobierno (confianza y trazabilidad en misiones críticas)

Tendencias

- **Modelos con capas interpretables** o explicable por diseño
- **Integración** al **ciclo** de **vida** del **sistema**: de métricas de precisión a transparencia
- Sistemas **colaboración Humano-IA** con interfaces de exploración y ajuste
- **Auditoría automatizada** de **decisiones** (trazabilidad)

Caso SAS: Falta de Interpretabilidad



¡Problema! fraudes en su programa de fidelización, reclamaciones y transacciones atípicas

¡Solución! cada alerta generaba un reporte de las variables más influyentes y una explicación local de la predicción

¿Cómo lo hicieron?

- Integraron **InterpretML** toolkit (SHAP & LIME) en su pipeline de **detección** de **fraude**.
- Generaron **explicaciones globales** y locales para cada alerta de **transacción sospechosa**.
- Permitieron al **equipo** de **fraude ver** qué **señales** (por ejemplo, pooling de puntos, patrones de claims) **dispararon** la **alerta**.

Resultados

- **80 %** de las **alertas explicadas** en < 1 min por **analista**.
- **Confianza** del **equipo** en el **sistema** subió **+30 %**, **acelerando** la adopción de **alertas**.
- **Tiempo** de **investigación reducido** en **25 %**, enfocando a los analistas en los casos realmente críticos.

Discusión

Machine Learning aplicado a Sistemas de Detección de Intrusos basados en Anomalías (ABIDS)

¿Cómo equilibrar la sensibilidad del sistema sin generar fatiga por falsas alertas?

Contexto:

Los ABIDS tienen gran potencial para detectar amenazas desconocidas, pero suelen generar un número alto de falsos positivos, es decir, alertas sobre actividades legítimas que son mal clasificadas como amenazas.

Temas para discusión:

- ¿Deben priorizarse sistemas conservadores (que alerten más) o más precisos (que alerten menos)?
- ¿Qué papel juegan los analistas humanos en este equilibrio?
- ¿Qué técnicas consideran más útiles: validación manual, umbrales dinámicos, aprendizaje activo, reglas híbridas, etc.?

Caja Negra vs. Transparencia

¿Deberíamos sacrificar explicabilidad por precisión en modelos de seguridad?

Contexto: Los modelos más precisos (como deep learning) suelen ser opacos. En cambio, los modelos más interpretables (como árboles de decisión) tienden a ser menos precisos.

Temas para discusión

- ¿Es ético usar modelos que no se pueden explicar completamente en entornos críticos?
- ¿Cómo afecta la confianza del usuario y del personal de seguridad?
- ¿Qué rol pueden jugar las técnicas de Explainable AI (XAI)?
- ¿Existen contextos donde la precisión absoluta justifica la opacidad?

¡Gracias!