Aprendizaje Automático

Modelos de clasificación

Integrantes:

- Bryan Ismael Alvarado Quimiz
- Diego André Arámbulo Vítores
- Jesús Antonio Zambrano Parrales

Entregable 1: Repositorio

https://github.com/diegoarambulo/2025_301_MIAR0525_A01_SEM2_GRUPAL/tree/main

Entregable 2: Video presentación

https://youtu.be/lWwAwwNrHdU

Resumen

Entrenamiento de modelos de machine learning supervisados usando Regresión logística, SVM(Suport Virtual Machine) y Árboles de Decisión, aplicados a un caso técnico de detección de patrones.

Utilizando un dataset enfocado en la detección de anomalías en ataques informáticos reales **BETH**

Problemática general

Identificación temprana y proactiva de posibles vulnerabilidades en el sector de la ciberseguridad mediante la implementación de clasificadores supervisados usando Regresión logística, SVM(Suport Virtual Machine) y Árboles de decisión para encontrar el mejor resultado y rendimiento al evaluar sus métricas.

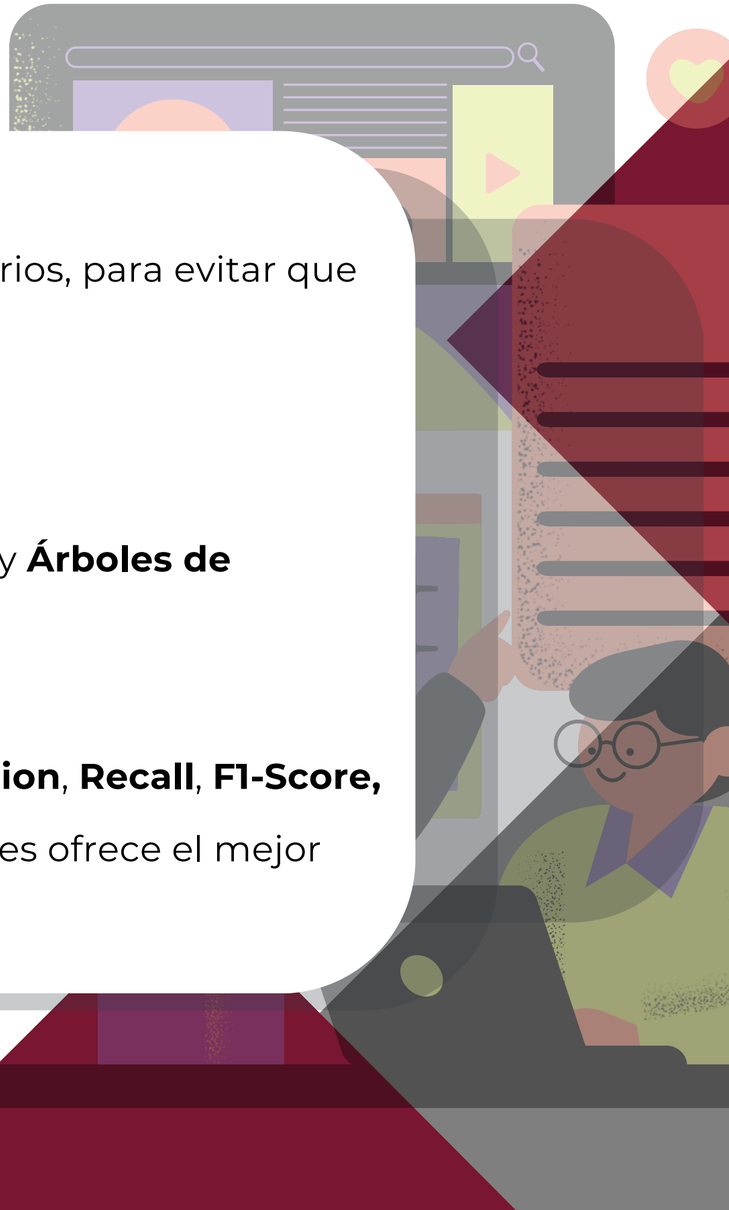
Metodología

La metodología implementada para abordar la identificación temprana de vulnerabilidades se estructura en seis fases clave, asegurando la robustez y el rendimiento de los modelos.

1. **Adquisición y comprensión del dato:**
   - Se utilizó el dataset BETH, para tener una fuente de datos de auditoría detallada.
   - Se definieron las clases objetivo **sus** y **evil** como los indicadores a predecir.
2. **Preprocesamiento y codificación:**
   - Dentro de la limpieza se eliminaron columnas irrelevantes o redundantes como processId, argsNum, etc.
   - Las variables categóricas de texto processName, eventName se transformaron en un formato numérico binario.
3. **Análisis de correlación y selección de variables:**
   - Se calculó y visualizó la Matriz de Correlación para identificar qué variables procesos o eventos tienen la relación más fuerte con las clases objetivo sus y evil.
   - Esto permitió una selección de características Feature Selection enfocada en las variables más predictivas

Metodología

4. **Balanceo de clases:**
   - Se abordó el problema del desequilibrio de clases donde los ataques son eventos minoritarios, para evitar que los modelos se sesgaran hacia la predicción de "no ataque".
5. **Entrenamiento de clasificadores supervisados:**
   - Los datos balanceados se dividieron en conjuntos de entrenamiento y prueba.
   - Se entrenaron y ajustaron los tres clasificadores especificados: **Regresión Logística, SVM,** y **Árboles de Decisión**.
6. **Evaluación de modelos:**
   - Se validó el rendimiento de cada modelo utilizando métricas clave como **Accuracy**, **Precision**, **Recall**, **F1-Score,** y la curva ROC (Receiver Operating Characteristic) para determinar cuál de los clasificadores ofrece el mejor rendimiento y la mayor capacidad de detección proactiva de ciberataques.



Conclusión final

Basado en los resultados obtenidos después de la comparativa entre los modelos usados y sus diferentes métricas, podemos inferir que para estos escenarios de detección temprana son mas eficientes aquellos modelos que tienen un menor consumo computacional ya que al ser en tiempo real podremos observar que tendrán un uso intensivo por la misma naturaleza de ayudar a una estrategia proactiva que priorice una alta disponibilidad en sectores estratégicos donde las TI's sean la principal fuerza operativa