LEAD UNIVERSITY

INFORME DE PROYECTO #1

Fecha Entrega: 6 de marzo 2025

Construcción Completa de un Data Pipeline y Aplicación de un Modelo de Inteligencia Artificial con un DataSet de Cyberseguridad, Intrusión de Detección

Roles Participantes:

- **Ingeniero de Datos:**

  Samuel Salazar

- **Cientifico de Datos:**

  Nathan Ghenassia

- **Programadora Web:**

  Maria Fernanda Camacho

INTRODUCCIÓN

En el contexto de la ciberseguridad, la identificación de intrusiones, es un reto importante para asegurar la protección de los sistemas de información. La mayor complejidad de los ataques requiere el uso de métodos avanzados de análisis de datos y aprendizaje automático para reconocer patrones de conducta dañina. En este proyecto grupal, hemos desarrollado un completo Data Pipeline y un modelo avanzado de inteligencia artificial enfocado en la detección de intrusiones cibernéticas. Usamos el Cybersecurity Intrusion Detection Dataset, que nos permitió analizar el tráfico de red y el comportamiento del usuario para identificar posibles amenazas. Esta solución combina la administración de datos, gobernanza de datos, y la aplicación de un modelo de IA avanzado, integrados en una plataforma web interactiva utilizando Streamlit.

OBJECTIVO

Desarrollar un pipeline de datos seguro y automatizado que descargue, procese y almacene datos para entrenar modelos de Machine Learning con el fin de detectar intrusiones en sistemas informáticos. Además, optimizar el modelo de IA para mejorar la precisión en la clasificación de ataques, asegurando su eficacia en la detección de intrusiones. Se busca garantizar la protección de credenciales y datos sensibles a lo largo del proceso.

DISEÑO

El sistema está compuesto por cuatro módulos principales:

- **Seguridad y cifrado de credenciales:** Uso de la librería cryptography para encriptar datos y tokens de acceso.
- **Pipeline de datos**: Descarga, limpieza y almacenamiento seguro de datasets.
- **Modelo de Machine Learning:** Entrenamiento y optimización de un clasificador para la detección de intrusiones.
- **Interfaz de usuario:** Aplicación en Streamlit para ejecutar y visualizar los resultados.

METODOLOGÍA (DATA PIPELINE)

El data pipeline comienza con la autenticación y acceso a las fuentes de información. Se descargan los datasets desde Kaggle, se verifica el acceso de los usuarios a través del archivo roles.json y se desencripta el token de GitHub para permitir la interacción con el repositorio. Estas medidas garantizan que solo usuarios autorizados puedan ejecutar el pipeline y acceder a la información procesada.

Luego, los datos se procesan eliminando valores nulos y la columna session_id para preservar la privacidad. Se encriptan los identificadores de sesión y se convierten variables categóricas en valores numéricos para su análisis en modelos de Machine Learning. Finalmente, los datos procesados se almacenan en GitHub, asegurando su disponibilidad y actualización.

METODOLOGÍA (MODELO IA)

El modelo de Machine Learning sigue diferentes pasos para asegurar su efectividad en la detección de intrusiones. Primero, los datos son procesados por normalización con MinMaxScaler, eliminación de valores nulos y conversión de variables categóricas. Este paso es clave para mejorar la calidad de los datos antes de entrenar el modelo.

Para la fase de entrenamiento, se evalúan diferentes clasificadores como Support Vector Machine (SVM), Decision Tree y Multilayer Perceptron (MLP). Se ajustan los hiperparámetros usando GASearchCV para mejorar el rendimiento del modelo. Finalmente, se mide la precisión y el rendimiento del modelo mediante la matriz de confusión y métricas de clasificación antes de exportarlo en formato joblib para su uso en la aplicación.

RESULTADOS

1. **Precisión del modelo:** Se obtuvieron valores superiores al 90% en la detección de intrusiones.
2. **Eficiencia del Pipeline:** Reducción significativa en el tiempo de procesamiento y almacenamiento de datos.
3. **Seguridad**: Implementación exitosa de medidas criptográficas para proteger credenciales y datos sensibles.
4. **Automatización exitosa**: La integración del pipeline con GitHub permitió la actualización y almacenamiento seguro de los datos sin intervención manual.
5. **Mejora en la interpretabilidad:** El análisis de la matriz de confusión ayudó a identificar patrones de error y ajustar el modelo para mejorar su precisión.

ANALISIS

El modelo de IA demuestra una alta capacidad para la detección de intrusiones con un rendimiento estable. La integración del pipeline con GitHub permite una gestión eficiente de los datos y asegura su almacenamiento seguro. La implementación de medidas criptográficas minimiza riesgos de exposición de credenciales.

Además, la optimización con algoritmos genéticos ha permitido encontrar combinaciones de hiperparámetros que maximizan la precisión del modelo sin comprometer el tiempo de entrenamiento. La reducción del tiempo de procesamiento en el pipeline ha sido clave para garantizar que los datos estén disponibles en tiempo real para su análisis.

CONCLUSION

- Se logró desarrollar un sistema seguro y automatizado para la detección de intrusiones.
- El pipeline garantiza la protección de datos y credenciales.
- La optimización del modelo con algoritmos genéticos mejoró el desempeño

RECOMENDACIONES

- Implementar validaciones adicionales en roles.json para mejorar la seguridad.
- Explorar arquitecturas de Deep Learning para mejorar la detección de ataques sofisticados.
- Integrar dashboards para mejorar la visualización de resultados en la aplicación.
- Optimizar el pipeline mediante paralelización y procesamiento distribuido para reducir los tiempos de ejecución.
- Fortalecer la autenticación en la aplicación web para prevenir accesos no autorizados.