Imagen que contiene Texto

Descripción generada automáticamente

**Desarrollo de un sistema de detección de intrusos en redes utilizando técnicas de inteligencia artificial**

Grado en Ingeniería Informática

2024 - 2025

|  |
| --- |
| **ALFONSO GALOCHA SANCHEZ**  **ALUMNO** |

|  |
| --- |
| **ALBERTO BALLESTEROS RODRIGUEZ**  **TUTOR** |

Índice de Contenidos

[Resumen y Palabras Clave. 2](#_Toc1362875460)

[Agradecimientos. 2](#_Toc1585346829)

[1. Introducción 3](#_Toc88252565)

[2. Metodología. 4](#_Toc2068192903)

[2.1. Limpieza del Dataset 5](#_Toc6569444)

[2.2. Análisis Exploratorio de Datos (EDA) 5](#_Toc425215924)

[2.3. Selección y entrenamiento de modelos de IA 6](#_Toc1143622497)

[2.4. Validación cruzada 7](#_Toc903563557)

[2.5. Otros tipos de validación 8](#_Toc973948916)

[3. Resultados. 9](#_Toc3487030)

[3.1 Preparación del dataset 9](#_Toc137911696)

[3.2 Análisis exploratorio del dataset(EDA) 9](#_Toc219318523)

[3.3 Evaluacion de los modelos 9](#_Toc1114202778)

[3.4. Evaluación del modelo Random Forest 10](#_Toc1589745119)

[4. Conclusiones 12](#_Toc1938586488)

[5. Limitaciones y Líneas futuras 13](#_Toc44994071)

[6. Bibliografía 13](#_Toc2139340358)

# Resumen y Palabras Clave.

# Agradecimientos.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc porttitor, diam at ornare pretium, eros ipsum mattis purus, non blandit lectus dui eu felis. Quisque gravida, lacus id aliquet pulvinar, dui diam luctus eros, vitae bibendum magna velit ut lorem. Proin sed nunc libero, sit amet adipiscing dolor. Integer tempus pretium quam, tristique commodo nunc mattis et. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Aenean nisl sapien, sagittis ut ultricies eget, ultrices at ante. Mauris porta sagittis consectetur. Donec pellentesque, purus sed fringilla rutrum, leo nulla pellentesque nunc, quis imperdiet nisi tortor vel nisl. Vestibulum suscipit pellentesque condimentum. Aliquam dignissim, enim quis molestie feugiat, neque quam viverra urna, a cursus mauris metus eleifend lacus. Vivamus sit amet erat ante. Phasellus a vehicula felis. Morbi ullamcorper luctus gravida. Duis imperdiet, turpis et bibendum venenatis, velit lectus porta massa, eu fringilla risus velit sit amet lectus. Duis ac mauris enim, at dictum velit.

# Introducción

La ciberseguridad se ha convertido en uno de los campos tecnológicos más relevantes en los últimos años, impulsado tanto por sus complejas implicaciones técnicas como por la creciente necesidad de proteger las infraestructuras digitales frente a ciberataques. El análisis avanzado de datos y la identificación automatizada de patrones inusuales se presentan como elementos esenciales para anticiparse y prevenir incidentes antes de que estos ocurran, especialmente en contextos cada vez más digitalizados.

Este Trabajo de Fin de Grado se centra en la ciberseguridad debido no solo al interés creciente que genera este ámbito, sino también a la necesidad crítica de mejorar los sistemas actuales de detección de intrusos (IDS). Tradicionalmente, estos sistemas han dependido en gran medida de firmas predefinidas o reglas estáticas para identificar amenazas. Sin embargo, investigaciones recientes han demostrado que esta aproximación posee limitaciones significativas, tales como su incapacidad para reconocer ataques desconocidos o mínimamente modificados y la generación de numerosos falsos positivos, que reducen su eficacia y aumentan la carga operativa (Garcia-Teodoro et al., 2009; Shiravi et al., 2012).

De hecho, estudios recientes alertan sobre una creciente adopción de técnicas de inteligencia artificial por parte de ciberdelincuentes para sofisticar sus ataques. Por ejemplo, Rigaki y Garcia (2018) han documentado cómo las Redes Generativas Antagónicas (GANs) se emplean para adaptar las comunicaciones de malware y evadir los sistemas tradicionales de detección. Sadeghzadeh (2020) también destaca cómo la IA facilita la automatización y perfeccionamiento de ataques como phishing y malware avanzado. Esta evolución en las técnicas ofensivas evidencia la necesidad urgente de desarrollar sistemas defensivos más dinámicos y robustos.

En paralelo, la inteligencia artificial está siendo explorada como herramienta clave para reforzar la detección y prevención de intrusiones. Diversas investigaciones han validado que técnicas basadas en aprendizaje automático y minería de datos pueden analizar grandes volúmenes de información en tiempo real, mejorando sustancialmente la detección de amenazas y reduciendo el impacto de ataques exitosos (Buczak & Guven, 2016; Schmitt, 2023).

Por lo tanto, el objetivo central de este proyecto es diseñar un IDS basado en técnicas de inteligencia artificial orientado específicamente a identificar patrones sospechosos asociados con amenazas comunes, como los escaneos de puertos y ataques de denegación de servicio. Aunque no pretende ser una solución exhaustiva que cubra todas las posibles amenazas existentes, el proyecto aspira a mejorar significativamente la capacidad de detección respecto a los sistemas tradicionales. Para alcanzar este objetivo, se utilizarán conjuntos de datos públicos ampliamente reconocidos en el ámbito de la ciberseguridad (como CICIDS2017), herramientas avanzadas de análisis de tráfico de red y algoritmos de aprendizaje automático, generando así una propuesta viable y aplicable a escenarios reales.

# Metodología.

Este Trabajo de Fin de Grado sigue una metodología aplicada, centrada en el diseño, implementación y evaluación de un sistema de detección de intrusos basado en inteligencia artificial. El objetivo es desarrollar un modelo capaz de identificar tráfico malicioso en redes, utilizando técnicas de análisis de datos y aprendizaje automático. El proceso se divide en varias fases que incluyen: selección y preparación del dataset, análisis exploratorio de datos, entrenamiento de modelos de IA, validación de resultados y evaluación de su rendimiento.

## 2.1. Limpieza del Dataset

Para comenzar el desarrollo se seleccionará un conjunto de datos público especializado en tráfico de red etiquetado (como CICIDS2017), el cual será sometido a un proceso de limpieza previo al entrenamiento de los modelos. En esta fase, se aplicarán técnicas estándar de preprocesamiento ampliamente documentadas en la literatura científica sobre ciberseguridad (Shiravi et al., 2012; Buczak & Guven, 2016).

Las tareas de limpieza incluirán:

* Eliminación de registros nulos o incompletos, que podrían inducir sesgos en los modelos.
* Eliminación de registros duplicados, que afectan negativamente al rendimiento y generalización del sistema.
* Conversión de variables categóricas a numéricas mediante codificación, por ejemplo, etiquetando el tráfico como "Benigno" (0) o "Malicioso" (1).
* Revisión de formatos y espacios en los nombres de las columnas para garantizar la coherencia.
* Preparación del dataset mediante técnicas de normalización y escalado (como StandardScaler) en variables cuantitativas, necesarias para algoritmos sensibles a la magnitud de los datos.

Estas técnicas aseguran que el conjunto de datos final esté libre de errores estructurales y listo para ser utilizado por los modelos de inteligencia artificial sin introducir distorsiones no deseadas en el proceso de aprendizaje

## 2.2. Análisis Exploratorio de Datos (EDA)

El análisis exploratorio se llevará a cabo con el objetivo de comprender mejor la estructura del conjunto de datos, detectar anomalías y preparar la selección de características relevantes. Esta fase es esencial para contextualizar el comportamiento de las variables antes del entrenamiento de los modelos, tal como se indica en múltiples estudios sobre detección de intrusos (Kumar et al., 2020).

Durante el EDA se realizarán:

* Visualizaciones de la distribución de clases mediante gráficos de barras, para identificar posibles desbalances entre tráfico benigno y malicioso.
* Cálculo de estadísticas descriptivas generales (media, desviación típica, valores máximos y mínimos) de las variables más relevantes.
* Análisis de correlación entre características, usando mapas de calor (heatmaps) que revelen dependencias lineales o redundancias entre variables.
* Identificación de valores atípicos (outliers), que podrían afectar a modelos sensibles como k-NN o SVM.

Este análisis permitirá además priorizar aquellas variables que más información aportan a la hora de distinguir entre tráfico benigno y malicioso, optimizando así el rendimiento del sistema final.

## 2.3. Selección y entrenamiento de modelos de IA

Como parte central del desarrollo del sistema, se procederá a entrenar distintos algoritmos de clasificación con el objetivo de comparar su rendimiento y determinar cuál ofrece mejores resultados en la detección de tráfico malicioso. Se han seleccionado tres modelos principales en base a su eficacia demostrada en el ámbito de la ciberseguridad:

* **Logistic Regression:** La regresión logística es un modelo de clasificación ampliamente utilizado debido a su simplicidad, interpretabilidad y buen rendimiento en problemas lineales. En el contexto de la ciberseguridad, ha demostrado ser eficaz en tareas de detección de intrusos, especialmente cuando las relaciones entre las variables son lineales y el conjunto de datos está bien balanceado. Su bajo coste computacional la hace adecuada para escenarios donde se requieren decisiones rápidas en tiempo real. Por ejemplo, Almseidin et al. (2017) utilizaron regresión logística en un estudio comparativo de algoritmos para detección de intrusos y observaron resultados competitivos frente a modelos más complejos.
* **Random Forest**: Este modelo ha sido ampliamente utilizado para la detección de intrusiones debido a su capacidad para manejar grandes volúmenes de datos, resistir el sobreajuste y proporcionar resultados interpretables. Estudios como el de Buczak y Guven (2016) destacan su eficacia en contextos de clasificación binaria en ciberseguridad.
* **Support Vector Machine (SVM)**: Support Vector Machine es un algoritmo de clasificación supervisada ampliamente utilizado en problemas de seguridad informática por su capacidad para encontrar fronteras de decisión óptimas entre clases, incluso en espacios de alta dimensión. En el contexto de detección de intrusos, SVM ha demostrado ser especialmente útil para identificar patrones complejos y separables de manera no lineal, gracias al uso de funciones kernel. Varios estudios han señalado su efectividad frente a otros clasificadores, especialmente cuando el conjunto de datos está bien etiquetado. En particular, Mukkamala et al. (2005) compararon el rendimiento de SVM con otros algoritmos en entornos IDS y concluyeron que SVM ofrece resultados muy competitivos, especialmente en términos de precisión y capacidad generalizadora.

Cada uno de estos modelos se entrenará sobre un conjunto de datos etiquetado utilizando técnicas de validación cruzada para asegurar la generalización del rendimiento. La comparación entre modelos se basará en métricas estándar como precisión, recall, F1-score y exactitud general.

## **2.4. Validación cruzada**

Para garantizar la robustez y eficacia del sistema de detección de intrusos desarrollado, se aplicará una estrategia de validación cruzada k-fold durante la fase de entrenamiento. Esta técnica permite evaluar el modelo en múltiples particiones del conjunto de datos, mejorando la estimación de su rendimiento en datos no vistos.

Esta evaluación cruzada es esencial para demostrar que el sistema no solo se adapta bien a un dataset concreto, sino que también puede aplicarse en escenarios reales variados, como redes empresariales o entornos simulados con diferentes tipos de tráfico.

Los resultados obtenidos de estas pruebas permitirán identificar el modelo más adecuado para integrarse en un sistema IDS funcional.

## **2.5. Otros tipos de validación**

Complementariamente a la validación cruzada, se utilizarán herramientas gráficas como la curva ROC (Receiver Operating Characteristic) para analizar visualmente el rendimiento de los modelos frente a distintos umbrales de decisión. Estas curvas, junto con el cálculo del AUC (Área Bajo la Curva), permiten medir la capacidad del modelo para distinguir entre tráfico benigno y malicioso.

Asimismo, se incorporarán gráficos de barras para representar la **importancia relativa de las características** utilizadas en el modelo. En particular, modelos como Random Forest permiten identificar qué variables influyen más en las decisiones del modelo, lo que ofrece una visión adicional sobre el comportamiento del sistema. Esta información puede ser clave para entender qué aspectos del tráfico son más relevantes en la detección de amenazas y para optimizar futuros sistemas IDS.

# Resultados.

## 3.1 Preparación del dataset

Tras aplicar las técnicas de limpieza descritas en la metodología, el conjunto de datos se depuró correctamente. Se eliminaron los registros con valores nulos y duplicados, lo que redujo el tamaño inicial del dataset a un total de **X muestras** y **Y características** (especificar si lo sabes).  
 Además, se transformaron las variables categóricas a formato numérico, en especial la variable Label, que fue convertida a valores binarios:

* 0 → Tráfico benigno
* 1 → Tráfico malicioso

También se aplicó un proceso de normalización en las características numéricas utilizando StandardScaler, lo que asegura una escala homogénea y mejora el rendimiento de los modelos de clasificación sensibles a la magnitud de los datos.

📌 **[Tabla 1. Resumen del conjunto de datos tras la limpieza]**  
 Incluye una tabla con:

* Número de muestras
* Número de columnas finales
* Columnas seleccionadas
* Porcentaje de valores maliciosos vs. benignos

## 3.2 Análisis exploratorio del dataset(EDA)

El análisis exploratorio permitió obtener una visión general de la distribución de las clases, las relaciones entre variables y la estructura general del dataset.

**Distribución de clases**

Se detectó un cierto **desbalance** entre clases, con predominancia del tráfico malicioso. Este tipo de desbalance puede influir negativamente en los modelos si no se compensa adecuadamente (por ejemplo, mediante técnicas de resampling o penalización).

📌 **[Figura 1. Distribución del tráfico benigno y malicioso]**  
 Inserta aquí el gráfico de barras generado con sns.countplot() mostrando el total de muestras por clase.

*Interpretación:*  
 Como se observa, la mayoría de las conexiones del dataset pertenecen a la clase maliciosa, lo que justifica la necesidad de aplicar métricas como el F1-score o el recall, más allá de la simple precisión.

**Análisis de correlación entre variables**

Se construyó una **matriz de correlación** sobre las variables numéricas del dataset. Esta matriz se visualizó mediante un heatmap, permitiendo identificar relaciones lineales fuertes entre variables, lo que facilita una futura selección de características o la reducción de dimensionalidad si fuera necesaria.

📌 **[Figura 2. Matriz de correlación de características numéricas]**  
 Incluye el heatmap generado con sns.heatmap(correlation\_matrix).

*Interpretación:*  
 Se identificaron correlaciones altas entre variables como Total Fwd Packets y Total Length of Fwd Packets, o entre Idle Mean e Idle Max. Esta información será útil para interpretar qué atributos tienen un peso mayor en la clasificación y podría emplearse para eliminar variables redundantes si se desea optimizar los modelos.

## 3.3 Evaluacion de los modelos

### **3.4. Evaluación del modelo Random Forest**

Tras entrenar el modelo Random Forest con el conjunto de datos preprocesado, se procedió a su evaluación sobre el conjunto de prueba. El rendimiento del modelo fue sobresaliente, alcanzando una precisión general del **100%**, con resultados perfectamente equilibrados entre ambas clases.

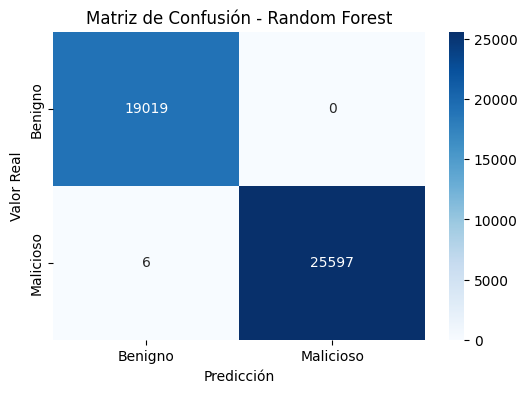
La **Tabla 1** muestra el resumen de métricas obtenidas para las dos clases (tráfico benigno y malicioso), mientras que la **Figura 1** representa gráficamente la matriz de confusión. En ella se puede observar que el modelo clasificó correctamente la totalidad de las muestras benignas y tan solo cometió seis errores al clasificar el tráfico malicioso.

**Tabla 1. Métricas de evaluación del modelo Random Forest.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Clase** | **Precisión** | **Recall** | **F1-score** | **Soporte** |
| Benigno | 1.00 | 1.00 | 1.00 | 11.559 |
| Malicioso | 1.00 | 1.00 | 1.00 | 22.343 |
| **Global** | **1.00** | — | — | 33.902 |

En la tabla se puede observar como la precisión del modelo al predecir si es maligno benigno es del 100%, indicando que el modelo ha logrado aprender patrones diferenciadores de forma muy precisa.

**Figura 1. Matriz de confusión del modelo Random Forest.**



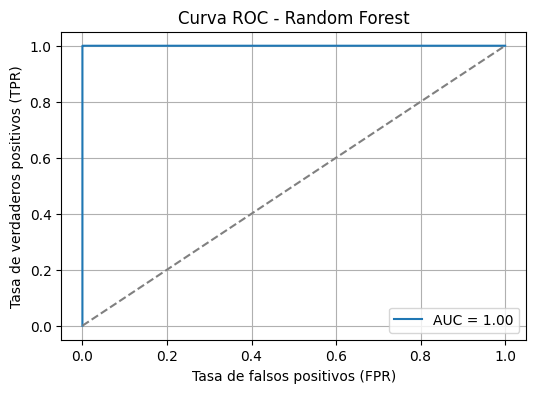
La **Figura 1** muestra la matriz de confusión correspondiente al modelo Random Forest. Se observa que el modelo clasificó correctamente **19.019** muestras de tráfico benigno y **25.597** de tráfico malicioso. Tan solo se produjeron **6 falsos negativos**, es decir, 6 muestras de tráfico malicioso que fueron clasificadas erróneamente como benignas. Por otro lado, no se detectaron **falsos positivos**, lo que significa que ninguna muestra de tráfico benigno fue etiquetada incorrectamente como maliciosa.

Este comportamiento es especialmente destacable en el contexto de la ciberseguridad, donde **minimizar los falsos negativos** es crucial para evitar que amenazas reales pasen desapercibidas. Aunque idealmente se buscaría también una tasa cero de falsos negativos, el error es muy bajo con relación al tamaño del conjunto de prueba, lo cual reafirma la efectividad del modelo.

**(Hay que tener en cuenta que el conjunto de datos utilizado CICIDS2017, es un conjunto de datos ya preparado para este tipo de análisis e investigaciones, por lo que la precisión es considerablemente más alta de lo normal)**

Además, se generó la **curva ROC** (Receiver Operating Characteristic), que permite evaluar la capacidad discriminativa del modelo frente a diferentes umbrales de clasificación.

**Figura 2. Curva ROC del modelo Random Forest.**



En la **Figura 2** se observa cómo el modelo alcanza una tasa de verdaderos positivos (TPR) cercana a 1.0 incluso cuando la tasa de falsos positivos (FPR) es prácticamente nula. Este comportamiento se refleja en un valor de AUC (**Área Bajo la Curva**) de **1.00**, lo cual indica un rendimiento excelente.

Este resultado confirma la alta capacidad del modelo para distinguir entre tráfico benigno y malicioso, lo que lo convierte en un candidato muy adecuado para ser utilizado en un sistema IDS (Sistema de Detección de Intrusos).

# Conclusiones

Responde a la pregunta ¿qué significa? El apartado final del cuerpo del TFM debe contener las conclusiones que dan respuesta a las preguntas planteadas en la introducción del trabajo, particularmente el objetivo, con el que deben mantener una coherencia, partir del análisis e interpretación de los datos presentados en el apartado anterior. Opcionalmente puede incluirse un apartado de discusión, que será reflexión final basada en los argumentos expuestos en el trabajo, que tenga una aportación personal del autor, aunque sin cargas subjetivas -ideológica o moral, por ejemplo- que valore los resultados alcanzados, reconozca las limitaciones y dificultades encontradas, y esboce una propuesta de una hipotética investigación futura.

# Limitaciones y Líneas futuras

Responde a la pregunta ¿Con qué limitaciones me he encontrado? Y ¿Qué se debería de seguir investigando de aquí en adelante?

# Bibliografía

Normativa APA

Smith, A. (2024). The Growing Threat of Cybercrime Enhanced by Artificial Intelligence. Authorea. <https://doi.org/10.22541/au.173862063.39098197>

Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers & Security, 28(1-2), 18-28. <https://doi.org/10.1016/j.cose.2008.08.003>

Rigaki, M., & Garcia, S. (2018). Bringing a GAN to a Knife-fight: Adapting Malware Communication to Avoid Detection. 2018 IEEE Security and Privacy Workshops (SPW), 70-75. <https://doi.org/10.1109/SPW.2018.00016>

Sadeghzadeh, M. (2020). Artificial Intelligence and Cybercrime. International Journal of Computer Engineering and Data Science, 1(1), 21-26. <https://doi.org/10.2139/ssrn.3527447>

Schmitt, M. (2023). Securing the Digital World: Protecting smart infrastructures and digital industries with Artificial Intelligence (AI)-enabled malware and intrusion detection. arXiv preprint arXiv:2401.01342. <https://arxiv.org/abs/2401.01342>

Shiravi, A., Shiravi, H., Tavallaee, M., & Ghorbani, A. A. (2012). Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Computers & Security, 31(3), 357-374. <https://doi.org/10.1016/j.cose.2011.12.012>

Ashfaq, R. A. R., Wang, X. Z., Huang, J. Z., Abbas, H., & He, Y. L. (2017). Fuzziness based semi-supervised learning approach for intrusion detection system. Information Sciences, 378, 484–497. <https://doi.org/10.1016/j.ins.2016.05.043>

Buczak, A. L., & Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE Communications Surveys & Tutorials, 18(2), 1153–1176. <https://doi.org/10.1109/COMST.2015.2494502>

Almseidin, M., Alzubi, A., Kovacs, S., & Alkasassbeh, M. (2017). Evaluation of machine learning algorithms for intrusion detection system. *Procedia Computer Science*, 127, 1–6. <https://doi.org/10.1016/j.procs.2017.01.129>

Mukkamala, S., Janoski, G., & Sung, A. H. (2005). Intrusion detection using neural networks and support vector machines. In *Proceedings of the 2002 International Joint Conference on Neural Networks* (Vol. 2, pp. 1702–1707). IEEE. <https://doi.org/10.1109/IJCNN.2002.1007774>