



Universidad de Valladolid

Escuela de Ingeniería Informática

TRABAJO FIN DE GRADO

Grado en Ingeniería Informática
Mención en Tecnologías de la Información

Diseño de un modelo neuronal para la detección y la clasificación de intrusiones en redes informáticas

Alumno:
Hugo López Álvarez

Tutor:
Diego García Álvarez

...

Agradecimientos

...

Resumen

Resumen

Abstract

Abstract

Índice general

Agradecimientos	III
Resumen	V
Abstract	VII
Lista de figuras	XIII
Lista de tablas	XV
1. Introducción	1
1.1. Contexto	1
1.2. Motivación	2
1.3. Objetivos del proyecto	3
1.4. Objetivos académicos	3
1.5. Estructura de la memoria	3
2. Metodología	5
2.1. CRISP-DM	5
2.2. SCRUM	8
3. Planificación	11
3.1. Planificación temporal	11

3.2. Gestión de riesgos	14
3.3. Estimación de costes	26
3.3.1. Costes materiales	26
3.3.2. Costes humanos	28
4. Entendimiento del problema	31
4.1. ¿Qué es un ataque a un sistema informático?	31
4.2. Tipos de ataque a sistemas informáticos	31
4.3. ¿Qué es TCP?	33
4.3.1. ¿Qué es un segmento TCP?	33
4.4. Importancia de protegerse frente a un ataque	35
4.5. Importancia de detectar los ataques rápidamente	36
4.6. Soluciones comerciales o actuales a estos problemas	36
4.7. Requisitos	37
4.7.1. Requisitos Funcionales	37
4.7.2. Requisitos No Funcionales	37
4.8. Objetivos del proyecto ¿HACE FALTA TENIENDO EN LA INTRODUCCIÓN?	38
4.9. ¿DEBERÍA AÑADIR ALGUNA SECCIÓN NUEVA?	38
5. Entendimiennto de los datos	39
5.1. Origen de los datos	39
5.2. Tipos de ataques registrados en los datos	39
5.3. Parámetros de los datos	40
5.4. Patrones preliminares, valores atípicos y sesgos	43
5.5. Preparación de los datos ¿SEPARAR EN OTRO CAPÍTULO?	43
6. Modelos	47
6.1. ¿Qué es un modelo neuronal?	47

6.1.1.	¿Qué tipos de modelos neuronales existen?	48
6.1.2.	Función de perdida	50
6.1.3.	Algoritmo de optimización	51
6.2.	Parámetros e hiperparámetros	53
6.3.	Modelo neuronal de clasificación binaria	54
6.3.1.	Proposito del modelo de clasificación binaria	54
6.3.2.	Diseño del modelo e hiperparámetros seleccionados	55
6.3.3.	Arquitecturas desarrolladas	56
6.4.	CREAR E INSERTAR IMAGENES DE LAS ARQUITECTURAS	59
6.4.1.	Matriz de confusión para clasificación binaria	59
6.4.2.	Métricas útiles en clasificación binaria	60
6.4.3.	Mejores hiperparámetros para cada arquitectura	63
6.4.4.	Comparación de las arquitecturas	65
6.5.	Modelo neuronal de clasificación multiclase	66
6.5.1.	Proposito del modelo de clasificación multiclase	67
6.5.2.	Diseño del modelo e hiperparámetros seleccionados	67
6.5.3.	Diseños del modelo de clasificación multiclase descartados	70
6.5.4.	Arquitecturas desarrolladas	71
6.5.5.	Matriz de confusión para clasificación multiclase	74
6.5.6.	Mettricas útiles en clasificación multiclase	75
6.5.7.	Mejores hiperparámetros para cada arquitectura	78
6.5.8.	Comparación de las arquitecturas	80
7.	Test	85
8.	Despliegue	87
9.	Tecnologías usadas	89

10.Seguimiento del proyecto	91
11.Conclusiones	93
A. Manuales	95
A.1. Manual de despliegue e instalación	95
A.2. Manual de mantenimiento	95
A.3. Manual de usuario	95
B. Resumen de enlaces adicionales	97
Bibliografía	99

Lista de Figuras

2.1. Esquema del ciclo CRISP-DM estándar [1]. 8

2.2. Esquema del ciclo de vida de scrum [2]. 9

3.1. Diagrama de Gantt para la planificación del proyecto. 13

4.1. Esquema del *three way handshake* de TCP [3]. 33

4.2. Esquema segmento TCP [4]. 34

4.3. Formato de la cabecera en IPv4 [5]. 35

5.1. Función de transformación para los parámetros IPv4 44

6.1. Esquema de redes neuronales. [6] 48

6.2. Esquema de redes neuronal convolucional. [7] 49

6.3. Esquema de algoritmo de optimización usando los gradiente calculados por la retropropagación. [8] 52

6.4. Estructura del modelo de clasificación binaria diseñada. 55

6.5. Mejores cinco configuraciones de hiperparámetros del modelo de clasificación binaria con una capa oculta de 25 neuronas. 64

6.6. Mejores cinco configuraciones de hiperparámetros del modelo de clasificación binaria con una capa oculta de 49 neuronas. 64

6.7. Mejores cinco configuraciones de hiperparámetros del modelo de clasificación binaria con una capa oculta de 98 neuronas. 65

6.8. Mejores cinco configuraciones de hiperparámetros del modelo de clasificación binaria. 65

6.9. Estructura del modelo de clasificación multiclase diseñada. 68

6.10. Resultados de los diferentes diseños probados para el modelo de clasificación multiclase. 70

6.11. Mejores cinco configuraciones de hiperparámetros del modelo de clasificación multiclase con una capa oculta de 25 neuronas. 79

6.12. Mejores cinco configuraciones de hiperparámetros del modelo de clasificación multiclase con una capa oculta de 49 neuronas. 80

6.13. Mejores cinco configuraciones de hiperparámetros del modelo de clasificación multiclase con una capa oculta de 98 neuronas. 80

6.14. Mejores diez configuraciones de hiperparámetros del modelo de clasificación multiclase. 81

6.15. Matriz de confusión de la mejor configuración de hiperparámetros encontrada para el modelo de clasificación multiclase. 83

Lista de Tablas

3.1. Cronograma de hitos y horas de trabajo. 14

3.2. Matriz Probabilidad-Impacto. 16

3.3. R01: Conflicto con periodos de exámenes y entregas de otras asignaturas. . . 18

3.4. R02: Fallos hardware en equipos de desarrollo. 19

3.5. R03: Limitaciones de capacidad de procesamiento para el entrenamiento de los modelos. 20

3.6. R04: Pérdida o corrupción del código. 21

3.7. R05: Disponibilidad limitada del tutor académico. 22

3.8. R06: Desviaciones en la planificación temporal inicial. 23

3.9. R07: Cambios en los requisitos técnicos. 24

3.10. R08: Dependencia de tecnologías inestables o no documentadas. 25

3.11. R10: Problemas de licencia de software. 26

3.12. Costes de Software. 28

3.13. Costes de Profesionales - Fase 1. 29

3.14. Costes de Profesionales - Fase 2. 29

3.15. Costes de Profesionales - Fase 3. 30

3.16. Coste Total por Profesional 30

5.1. Clasificación de amenazas de seguridad. 40

6.1. Matriz de confusión para clasificación binaria. 60

6.2. Matriz de confusión para clasificación con 9 clases. 75

Capítulo 1

Introducción

Este documento corresponde con la memoria del Trabajo de Fin de Grado (TFG) del Grado en Ingeniería Informática de la Escuela de Ingeniería Informática (EII) de la Universidad de Valladolid (UVa). Este trabajo se centra en el diseño e implementación de un modelo neuronal capaz de detectar intrusiones en una red informática. La principal ventaja de utilizar un modelo neuronal para la detección de intrusiones en una red, frente a los algoritmos tradicionales (como firmas basadas en reglas o análisis estadísticos), radica en su capacidad para aprender patrones complejos y no lineales en los datos, lo que le permite identificar amenazas desconocidas o variantes de ataques existentes (zero-day attacks). Mientras que los métodos tradicionales dependen de reglas predefinidas y actualizaciones manuales para detectar intrusiones (limitándose a ataques conocidos), las redes neuronales pueden analizar grandes volúmenes de tráfico de red, detectando anomalías sutiles y correlaciones ocultas mediante capas de abstracción.

1.1. Contexto

En la actualidad, los sistemas informáticos reciben muchos más ataques de denegación de servicio y de intrusión que hace unos años, esto se debe en parte a los avances en los modelos de Inteligencia Artificial (IA). Los sistemas informáticos enfrentan actualmente graves amenazas debido al uso malintencionado de la IA por parte de ciberdelincuentes. Una de las principales problemáticas es la automatización de ataques, donde herramientas basadas en IA permiten ejecutar campañas de ataques informáticos con mayor precisión y escala. Estas IAs pueden generar mensajes convincentes, imitar patrones de comportamiento legítimos y evadir medidas de seguridad tradicionales, lo que incrementa la frecuencia y sofisticación de los ataques.

Otro desafío crítico es la explotación de vulnerabilidades mediante IA, que acelera la identificación de fallos en sistemas sin intervención humana. Existen algoritmos de aprendizaje automático (*Machine Learning*, ML) que analizan grandes volúmenes de datos para

descubrir brechas de seguridad en tiempo récord, facilitando ataques dirigidos incluso contra infraestructuras críticas como hospitales. Por estos motivos, la ciberseguridad es un tema de actualidad del que cada vez más se habla con más frecuencia incluso en la prensa general o económica [9].

La IA también complica la defensa, ya que los sistemas de detección tradicionales no siempre pueden anticipar tácticas adaptativas generadas por algoritmos hostiles. Esto obliga a las organizaciones y empresas a invertir en soluciones de IA defensiva, como sistemas de respuesta autónoma. Sin embargo, esto genera una carrera tecnológica desigual donde actores maliciosos aprovechan herramientas accesibles y de bajo costo. La falta de regulación global agrava este escenario, dificultando la mitigación de riesgos asociados.

Además, los modelos neuronales son adaptativos: mejoran su precisión con el tiempo al entrenarse con nuevos datos, lo que es crucial en entornos dinámicos donde los ciberataques evolucionan rápidamente. Por ejemplo, pueden distinguir entre comportamientos legítimos inusuales (como un empleado accediendo a recursos fuera de horario) y actividades maliciosas (como filtración de datos), reduciendo falsos positivos. En cambio, los enfoques tradicionales suelen ser rígidos y requieren ajustes manuales frecuentes para mantener su eficacia.

Sin embargo, el uso de modelos neuronales para la defensa de los sistemas conlleva grandes desafíos, como la necesidad de grandes conjuntos de datos etiquetados y recursos computacionales intensivos. Aun así, en escenarios donde la sofisticación de los ataques supera las capacidades de detección convencionales, los modelos neuronales representan un salto cualitativo en proactividad y escalabilidad.

1.2. Motivación

La principal motivación, que impulsó el desarrollo de este proyecto es facilitar la detección de ataques en redes informáticas, que tantas complicaciones está generando a los encargados de la administración de estos sistemas. Para cumplir con esta motivación, se decidió implementar un modelo neuronal que cumpliera con estos requisitos.

Durante la formación universitaria en el Grado en Ingeniería Informática, los alumnos de la mención de tecnologías de la información, aprenden a administrar grandes sistemas de computación en aspectos como: la seguridad, la garantía de la información, la evaluación de dichos sistemas y el almacenamiento de los datos. Además de contener formación sobre ciertos componentes de desarrollo de software. La falta de formación en algunos aspectos de la IA en dicha mención, desemboca en una de las grandes motivaciones académicas para desarrollar este proyecto, como es la familiarización con las técnicas de ML, en concreto de las redes neuronales, como herramientas útiles.

1.3. Objetivos del proyecto

En esta sección se listan los objetivos del proyecto, que constituyen las metas específicas, medibles, alcanzables, relevantes y con plazos definidos que se persiguen con la ejecución del mismo. Dichos objetivos describen los resultados concretos que se espera lograr al finalizar el proyecto y proporcionan un marco de referencia para la planificación, la ejecución, el seguimiento y la evaluación de su progreso.

1. Diseñar e implementar un modelo capaz de detectar intrusiones en redes informáticas y proporcionar una clasificación previa de la intrusión.
2. Los modelos de detección han de ser modelos neuronales.
3. Evaluar y comparar los modelos generados con un dataset real y complejo .

1.4. Objetivos académicos

En esta sección se enumeran los objetivos académicos del presente estudio, los cuales representan las metas específicas, susceptibles de evaluación, realizables, pertinentes para el ámbito del conocimiento, que se pretenden alcanzar a través del desarrollo de este proyecto.

1. Comprender el funcionamiento de los modelos neuronales através de pytorch y las métricas de evaluación.
2. Aprender las características de varios de los tipo de modelos neruonales que existen.
3. Descubrir el potencial de las redes neuronales para optimizar y mejorar las tecnologías de la información, incluyendo la ciberseguridad de los sistemas.

1.5. Estrucutra de la memoria

Este trabajo fin de grado se estructura de la siguiente manera:

Capítulo 2 Metodología: En este capítulo se definen cuales son las fases de la metodología CRISP-DM que se utiliza como metodología y modelo de proceso para el diseño y evaluación de los modelos neuronales para la detección de intrusiones. Se describe la aplicación de cada una de las seis fases al contexto específico del desarrollo de modelos neuronales.

Capítulo 3 Planificación: Este capítulo presenta la planificación del proyecto. Se definen los recursos necesarios, se identifican las tareas principales, se estiman los plazos y se establece el cronograma. También se aborda la gestión de riesgos inicial y la asignación de roles.

Capítulo 4 Entendimiento del problema: En este capítulo se describe el problema que el proyecto busca abordar. Se presenta el contexto, la relevancia y los objetivos generales.

Capítulo 5 Entendimiento de los datos: Este capítulo se dedica a la exploración y comprensión del conjunto de datos utilizado para el entrenamiento y evaluación de los modelos desarrollados. Se describe la fuente, el formato, el tamaño y las variables de los datos. Se presenta un análisis exploratorio para identificar patrones, problemas de calidad y la distribución de las variables.

Capítulo 6 Modelos: En este capítulo se detallan los modelos de redes neuronales desarrollados y entrenados. Se describe la arquitectura, la justificación de su elección, los hiperparámetros, la función de pérdida y el optimizador. Se incluye la estrategia de entrenamiento y las métricas de evaluación.

Capítulo 7 Test: Este capítulo se centra en la evaluación final de los modelos entrenados. Se describe el conjunto de datos de prueba, el proceso de evaluación, la presentación de los resultados de las métricas y el análisis de las fortalezas y debilidades de los modelos.

Capítulo 8 Despliegue: Este capítulo aborda la fase de despliegue de los modelos entrenados. Se describe la integración en un entorno operativo, las consideraciones técnicas, los posibles desafíos y las estrategias de monitorización y mantenimiento. Debido a la diversidad de entornos existentes en los que se puede realizar el despliegue, se comentarán los pasos generales que habría que seguir para desplegar los modelos en un entorno de producción real, sin entrar en profundidad.

Capítulo 9 Tecnologías utilizadas: En este capítulo se listan y describen las tecnologías de software y hardware empleadas en el proyecto. Se incluyen lenguajes de programación, bibliotecas de aprendizaje automático, herramientas de visualización y plataformas de seguimiento de experimentos.

Capítulo 10 Seguimiento del proyecto: Este capítulo describe cómo se ha realizado el seguimiento del progreso del proyecto. Se definen los indicadores clave de rendimiento, las metodologías de seguimiento, las herramientas de gestión y los mecanismos para la identificación y resolución de desviaciones.

Capítulo 11 Conclusiones: En este capítulo final se presentan las conclusiones del proyecto. Se resumen los principales hallazgos, se evalúa el cumplimiento de los objetivos académicos, se discuten las implicaciones de los resultados, las limitaciones y las posibles líneas de trabajo futuro.

Capítulo 2

Metodología

En este capítulo se explica la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*), que se utiliza en el desarrollo del resto del proyecto para alcanzar los objetivos propuestos.

La adopción de metodologías estructuradas es fundamental en el desarrollo de proyectos informáticos, puesto que proporcionan un marco sistemático para garantizar la calidad, eficiencia y trazabilidad del proyecto. En particular, metodologías como CRISP-DM, permiten: alinear objetivos técnicos con necesidades de negocio, reducir riesgos mediante fases iterativas y documentadas, y facilitar la colaboración entre equipos multidisciplinares.

Según algunos estudios, los proyectos que utilizan metodologías estandarizadas incrementan un 35 % su probabilidad de éxito, frente a aproximaciones *ad-hoc*, al minimizar desviaciones en costes y plazos [10]. En el ámbito de la ciberseguridad, donde los requisitos legales y técnicos son críticos, este enfoque metodológico resulta indispensable para asegurar soluciones robustas y auditables

2.1. CRISP-DM

La metodología CRISP-DM, fue diseñada para guiar proyectos de minería de datos y aprendizaje automático. Su estructura cíclica y flexible, como se puede apreciar en la Figura 2.1 que muestra su ciclo de vida, la hace aplicable en diversos dominios, desde marketing hasta ciberseguridad. Está compuesta por las siguientes fases:

1. Comprensión del negocio: La primera fase de CRISP-DM establece los cimientos estratégicos del proyecto mediante un proceso de alineación entre los objetivos técnicos y las necesidades organizacionales. Para lograr establecer los cimientos, se lleva a cabo un análisis exhaustivo del contexto empresarial para identificar los problemas clave que el proyecto debe abordar, así como las oportunidades de mejora que podrían aprovecharse. Se realiza

un proceso de recopilación y documentación de requisitos que involucra a todas las partes interesadas relevantes. El resultado de esta fase es una definición precisa del alcance del proyecto, que incluye no solo los objetivos cuantificables sino también los criterios de éxito que permitirán evaluar el impacto real de la solución propuesta. Además, se establecen las limitaciones operativas y estratégicas que condicionarán el desarrollo del proyecto, asegurando que todas las fases posteriores se ejecuten dentro de un marco bien definido y alineado con las prioridades organizacionales. Esta fase se detalla en los capítulos [3.3.2 Coste Total del Proyecto \(300 horas\)](#) y [4 Entendimiento del problema](#).

2. Comprensión de los datos: Esta fase se centra en el análisis detallado de los datos disponibles para el proyecto, con el objetivo de evaluar su idoneidad y calidad para abordar los problemas identificados en la fase anterior. Este proceso implica un examen minucioso de las diversas fuentes de información, su estructura y sus características fundamentales. Durante esta etapa, se identifican y documentan aspectos críticos como la complejidad de los datos, la presencia de posibles sesgos y la representatividad de la información en relación con los objetivos del proyecto. La comprensión profunda de los datos permite anticipar desafíos potenciales y establecer estrategias adecuadas para su tratamiento en fases posteriores. Además, esta fase proporciona perspectivas que pueden influir en decisiones técnicas importantes, como la selección de algoritmos o el diseño de características. El resultado es un conocimiento del potencial y las limitaciones de los datos disponibles, que sirve como base para las transformaciones que se realizan en la siguiente fase. En este trabajo, esta fase se documenta en el capítulo [5 Entendimiento de los datos](#).

3. Preparación de los Datos: Se trata de una fase crítica donde los datos brutos se transforman en un conjunto adecuado para modelado. Esta etapa implica una serie de operaciones fundamentales que garantizan la calidad y consistencia de los datos que alimentan a los modelos analíticos. Las actividades realizadas en esta fase son cruciales para el éxito del proyecto, ya que determinan en gran medida la capacidad de los algoritmos para extraer patrones significativos y generar resultados confiables. Se aplican técnicas especializadas para abordar problemas comunes en los datos, asegurando que la información sea representativa, completa y se encuentre adecuadamente estructurada para los análisis posteriores. Cualquier deficiencia en la preparación de los datos puede comprometer significativamente la efectividad de las siguientes fases. Al finalizar este proceso, se obtiene un conjunto de datos optimizado que conserva la esencia de la información original mientras elimina ruido y distorsiones que podrían afectar negativamente a los resultados del modelado. Al igual que la fase anterior, esta fase se aborda en el capítulo [5 Entendimiento de los datos](#).

4. Modelado: Constituye el núcleo técnico del proceso CRISP-DM, donde se desarrollan y evalúan los algoritmos diseñados para extraer conocimiento de los datos preparados. Esta etapa comienza con la selección cuidadosa de las técnicas de modelado más apropiadas para los objetivos específicos del proyecto y las características de los datos disponibles. Durante el proceso de modelado, se exploran diferentes enfoques algorítmicos, ajustando meticulosamente sus parámetros para optimizar su rendimiento. En esta fase se incluyen procesos de validación diseñados para garantizar que los modelos desarrollados sean robustos y generalizables, capaces de mantener su efectividad cuando se enfrenten a datos nuevos y no vistos previamente. El modelado es un proceso iterativo que puede requerir volver a fases anteriores para refinar la preparación de datos o incluso reconsiderar algunos aspectos del planteamiento inicial del problema. El resultado de esta fase es uno o varios modelos validados que cumplen

con los criterios de calidad establecidos y están listos para su evaluación en el contexto de los objetivos empresariales definidos inicialmente. Esta fase se documenta en el capítulo 6 [Modelos](#) de este trabajo.

5. Evaluación: Esta fase representa un examen exhaustivo de los modelos desarrollados, contrastando su desempeño técnico con los objetivos empresariales establecidos en la primera fase del proyecto. Este proceso va más allá de las métricas estadísticas tradicionales para incorporar una valoración del impacto potencial de la solución propuesta. Durante la evaluación, se analiza minuciosamente la capacidad de los modelos para resolver el problema de negocio original, considerando tanto su precisión técnica como su aplicabilidad práctica en el contexto organizacional. Se identifican y documentan las limitaciones de los modelos, así como los posibles riesgos asociados a su implementación. Esta fase también incluye la validación de los resultados con las partes interesadas clave, asegurando que la solución cumpla con las expectativas y requisitos operativos. La evaluación termina con una decisión fundamentada sobre la idoneidad de los modelos para su implementación, junto con recomendaciones para su posible mejora o adaptación a escenarios futuros. También se valida su robustez en escenarios realistas. Esta fase se detalla en el capítulo 7 [Test](#) de este trabajo.

6. Despliegue: Se trata de la fase final de CRISP-DM, esta se centra en la transición del modelo analítico desde un entorno de desarrollo hasta un entorno de producción real donde pueda generar valor tangible para la organización. Este proceso implica una serie de actividades cuidadosamente planificadas que garantizan la integración efectiva de la solución en los procesos empresariales existentes. El despliegue incluye aspectos técnicos como la implementación de la infraestructura necesaria, el desarrollo de interfaces adecuadas y la creación de mecanismos de monitoreo continuo. También se ha de tener en cuenta la capacitación de los usuarios finales y la documentación exhaustiva de la solución, asegurando su adopción efectiva y su uso óptimo. La fase de despliegue también establece procesos para el mantenimiento y actualización periódica del modelo, puesto que las soluciones analíticas requieren evolución continua para mantener su relevancia y efectividad. Como en el resto de metodologías, se implementan mecanismos para medir el impacto real de la solución una vez en producción, cerrando el ciclo al proporcionar retroalimentación valiosa que puede ser la base de futuros proyectos analíticos. Esta fase no se aborda en el trabajo, pero se le dedica el capítulo 8 [Despliegue](#), en él que se comentan los pasos generales que habría que seguir para desplegar los modelos en un entorno de producción real, sin entrar en profundidad.

Como se ha explicado, CRISP-DM es una metodología iterativa, esto significa que los resultados de fases posteriores pueden revelar la necesidad de ajustes en etapas anteriores (como recolectar más datos o redefinir objetivos). Su enfoque estructurado minimiza riesgos y maximiza el valor entregado, siendo especialmente útil en proyectos complejos donde la alineación entre técnica y negocio es esencial.

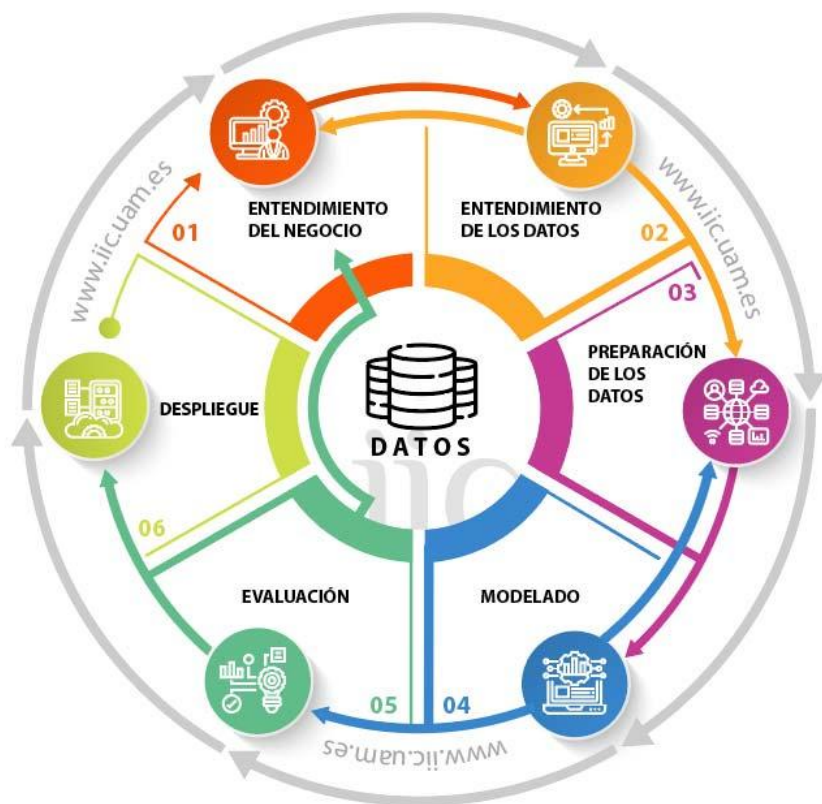


Figura 2.1: Esquema del ciclo CRISP-DM estándar [1].

2.2. SCRUM

Un marco de trabajo ágil es una estructura metodológica utilizada para gestionar proyectos de manera flexible, iterativa y centrada en la entrega continua de valor. Estos marcos se basan en los principios del Manifiesto Ágil, el cual promueve la colaboración entre equipos multidisciplinarios, la respuesta rápida al cambio y la entrega frecuente de productos funcionales al cliente [11].

Los marcos ágiles permiten a los equipos adaptarse a condiciones cambiantes y enfocarse en mejorar constantemente, promoviendo una comunicación fluida, ciclos de retroalimentación cortos y una participación activa del cliente durante todo el desarrollo del producto.

Dentro de estos marcos, uno de los más populares y ampliamente adoptados es Scrum.

Scrum es un marco de trabajo ágil diseñado para ayudar a equipos a desarrollar productos complejos de forma incremental. Se caracteriza por su estructura liviana y fácil de entender, aunque su dominio completo requiere disciplina y experiencia. Scrum se basa en ciclos de

trabajo cortos llamados *sprints*, que típicamente duran entre una y cuatro semanas, durante los cuales se entrega un incremento funcional del producto. La Figura 2.2 muestra el ciclo de vida del marco de trabajo ágil scrum.

Este marco de trabajo ágil, establece roles definidos (*Product Owner*, *Scrum Master* y Equipo de Desarrollo), eventos específicos (como el *Sprint*, la Planificación del *Sprint*, la Revisión, la Retrospectiva y el *Scrum* Diario) y artefactos clave (*Product Backlog*, *Sprint Backlog* e Incremento). Estos elementos promueven la transparencia, la inspección continua y la adaptación dentro del proceso de desarrollo.

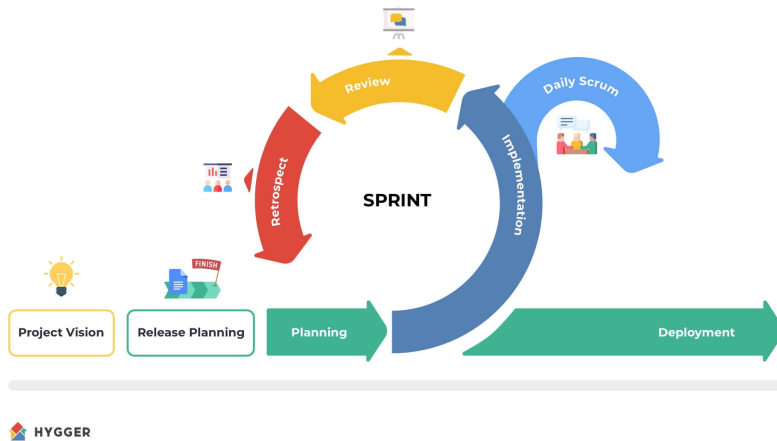


Figura 2.2: Esquema del ciclo de vida de scrum [2].

Capítulo 3

Planificación

Este capítulo aborda la organización detallada del trabajo de fin de grado, cubriendo desde su diseño inicial hasta la implementación y el seguimiento durante su desarrollo. Una planificación rigurosa resulta fundamental para sentar las bases del proyecto, ya que permite definir con claridad los objetivos, los recursos necesarios, los plazos de entrega y las actividades clave para alcanzar los resultados esperados [12].

En primer lugar, se establece una planificación temporal preliminar, donde se estiman los tiempos requeridos para cada etapa. Este cronograma se estructura en torno a las fases de la metodología CRISP-DM, complementadas con etapas específicas propias de un Trabajo de Fin de Grado. A continuación, se realiza un análisis de riesgos exhaustivo, evaluando tanto la probabilidad como el impacto de cada posible contingencia.

Además, se elabora un presupuesto detallado para las tareas del proyecto, abordado desde dos perspectivas. Por un lado, se incluye una estimación realista de los costes asociados a la ejecución del trabajo en el ámbito académico. Por otro lado, se plantea una proyección teórica de los gastos que implicaría un proyecto equivalente en un contexto profesional.

Por último, se contrasta la planificación inicial con el desarrollo real del trabajo, lo que permite evaluar posibles desviaciones y los aprendizajes obtenidos durante el proceso.

3.1. Planificación temporal

La planificación temporal constituye un elemento fundamental en la ejecución de un proyecto fin de grado, ya que permite estructurar de manera sistemática todas las actividades necesarias para alcanzar los objetivos propuestos. En el contexto de un trabajo académico que combine el desarrollo de software con una metodología de investigación, como es el caso de CRISP-DM para el modelo de proceso y SCRUM para la gestión del proyecto, una adecuada planificación garantiza la distribución equilibrada del tiempo disponible entre

las distintas fases del trabajo. Esta organización temporal resulta especialmente relevante cuando se deben coordinar aspectos teóricos, desarrollo técnico y validación de resultados, asegurando que cada componente reciba la atención necesaria sin comprometer la calidad global del proyecto.

El empleo de un diagrama de Gantt como herramienta de planificación ofrece ventajas significativas para visualizar la secuencia de actividades y su superposición temporal. Este tipo de representación gráfica facilita la identificación de hitos críticos y dependencias entre tareas, aspectos particularmente importantes cuando se combinan metodologías diferentes como CRISP-DM y SCRUM. La primera, con sus fases bien definidas, proporciona la estructura para el desarrollo del núcleo analítico del proyecto, mientras que SCRUM, con sus sprints iterativos, permite adaptar el trabajo a los descubrimientos que vayan surgiendo durante la investigación. La integración de ambas aproximaciones en un único cronograma exige una cuidadosa coordinación que el diagrama de Gantt ayuda a materializar de forma clara y comprensible. El diagrama de Gantt utilizado para la planificación de este trabajo se muestra en la Figura 3.1. Este diagrama tiene una duración de 300 horas, que corresponden con las horas de trabajo de alumno de los 12 créditos ECTS que exige el plan de estudios del Grado de Ingeniería Informática de la UVa para el TFG.



Figura 3.1: Diagrama de Gantt para la planificación del proyecto.

3.2. Gestión de riesgos

En este apartado se presentan los principales riesgos potenciales del proyecto junto con sus correspondientes planes de mitigación. Además, en la Figura 3.1 se muestran las horas dedicadas al cumplimiento de cada hito del trabajo. De acuerdo con el PMBOK (*Project Management Body of Knowledge*) [12], los riesgos en gestión de proyectos se clasifican en las categorías de la sección 3.2.

Hitos del Proyecto

Hito	Horas hasta el hito	Horas acumuladas
Finalización de formación y trabajos previos	20	20
Finalización de la planificación inicial	30	50
Finalización de la comprensión de datos	40	90
Finalización del modelado	80	170
Obtención de modelos óptimos	40	210
Finalización y entrega de la memoria	90	300

Tabla 3.1: Cronograma de hitos y horas de trabajo.

Tipología de Riesgos

1. Riesgos Técnicos:

- Limitaciones en la infraestructura de hardware.
- Incompatibilidad entre sistemas o tecnologías.
- Problemas de rendimiento o escalabilidad.
- Deficiencias en el diseño o implementación de software.

2. Riesgos de Gestión:

- Deficiencias en la comunicación entre los miembros del equipo.
- Modificaciones en los requisitos del proyecto.
- Inestabilidad del equipo por conflictos internos o rotación de personal.
- Retrasos en la disponibilidad de recursos críticos.

3. Riesgos de Mercado:

- Aparición de competencia no anticipada.
- Variaciones en las condiciones del mercado que afectan la demanda.
- Cambios regulatorios que impactan la ejecución del proyecto.

4. Riesgos Financieros:

- Limitaciones en la disponibilidad de fondos.
- Excesos presupuestarios no previstos.
- Fluctuaciones en los tipos de cambio.

5. Riesgos Externos:

- Fenómenos meteorológicos adversos.
- Interrupciones en la cadena de suministro.
- Eventos naturales catastróficos.

Metodología de Evaluación

La identificación y valoración de riesgos se realiza mediante criterios cualitativos, al no disponer de métricas cuantitativas suficientemente fiables para un análisis más exhaustivo. Este enfoque permite priorizar los riesgos según su impacto potencial y probabilidad de ocurrencia [13].

		Impacto				
		Mínimo	Bajo	Medio	Alto	Extremo
Probabilidad	Extrema					
	Alta					
	Media					
	Baja					
	Muy Baja					

Nivel de Riesgo	Color
Extremo	
Alto	
Moderado	
Bajo	
Mínimo	

Tabla 3.2: Matriz Probabilidad-Impacto.

Probabilidad

Grado de posibilidad de que un riesgo se materialice. Se suele cuantificar en escala del 1 (muy improbable) al 5 (casi seguro). En la matriz que se muestra en la Tabla 3.2, determina el eje vertical y se combina con el impacto para priorizar riesgos.

Impacto

Consecuencia o efecto potencial que tendría la materialización del riesgo. Se valora del 1 (impacto mínimo) al 5 (impacto catastrófico). Representa en la matriz que se muestra en la Tabla 3.2, el eje horizontal en la matriz y mide la severidad del riesgo.

Plan de Mitigación

Acciones proactivas para reducir la probabilidad o impacto del riesgo antes de que ocurra. Incluye:

- Prevención: Eliminar las causas del riesgo.
- Reducción: Disminuir su probabilidad o impacto.
- Transferencia: Trasladar el riesgo a terceros.

Plan de Contingencia

Medidas reactivas que se implementan cuando el riesgo se materializa. Contiene:

- Activación: Criterios para ejecutar el plan.
- Respuesta: Acciones específicas de contención.
- Recuperación: Cómo volver a la normalidad.

Nivel de Riesgo

Resultado de multiplicar la probabilidad por el impacto. Clasifica riesgos en:

- Alto (15-25): Requieren acción inmediata.
- Medio (5-14): Necesitan monitoreo.
- Bajo (1-4): Aceptables con supervisión mínima.

Umbral de Riesgo

Límite máximo aceptable de riesgo para el proyecto. Determina cuándo se deben implementar planes de mitigación o contingencia.

Propietario del Riesgo

Persona o equipo responsable de monitorear cada riesgo y ejecutar los planes correspondientes.

Riesgos identificados

1. **R01:** Conflicto con periodos de exámenes y entregas de otras asignaturas (Tabla 3.3).
2. **R02:** Fallos hardware en equipos de desarrollo (Tabla 3.4).
3. **R03:** Limitaciones de capacidad de procesamiento para el entrenamiento de los modelos (Tabla 3.5).
4. **R04:** Pérdida o corrupción del código (Tabla 3.6).
5. **R05:** Disponibilidad limitada del tutor académico (Tabla 3.7).
6. **R06:** Desviaciones en la planificación temporal inicial (Tabla 3.8).
7. **R07:** Cambios en los requisitos técnicos (Tabla 3.9).
8. **R08:** Interpretación errónea de los resultados por falta de formación en técnicas de aprendizaje automático. (Tabla 3.10).
9. **R09:** Problemas de licencias de software (Tabla 3.11).

Riesgo R01	
Título	Conflicto con periodos de exámenes y entregas de otras asignaturas.
Descripción	Sobrecarga académica que dificulta la dedicación y el rendimiento en todas las tareas.
Probabilidad	4 (Alta)
Impacto	3 (Media)
Matriz P/I	Alta/Media (12)
Plan Mitigación	<ul style="list-style-type: none">■ Coordinar calendario académico anticipadamente.■ Avanzar trabajo en periodos de menor carga de trabajo.
Plan Contingencia	<ul style="list-style-type: none">■ Dedicar horas extra en los asuntos académicos.■ Reorganizar prioridades temporales.

Tabla 3.3: R01: Conflicto con periodos de exámenes y entregas de otras asignaturas.

Riesgo R02	
Título	Fallos hardware en equipos de desarrollo.
Descripción	Problemas causados por el mal funcionamiento de los componentes físicos de un ordenador, como la placa base, la tarjeta gráfica, la memoria RAM, el disco duro o la fuente de alimentación en equipos de desarrollo.
Probabilidad	3 (Media)
Impacto	4 (Alto)
Matriz P/I	Media/Alto (12)
Plan Mitigación	<ul style="list-style-type: none">■ Mantenimiento preventivo mensual.■ Uso de equipos redundantes.
Plan Contingencia	<ul style="list-style-type: none">■ Utilizar equipos alternativos.■ Uso de máquinas virtuales del sistema de virtualización de la Escuela de Ingeniería de Informática de Valladolid.

Tabla 3.4: R02: Fallos hardware en equipos de desarrollo.

Riesgo R03	
Título	Limitaciones de capacidad de procesamiento para el entrenamiento de los modelos.
Descripción	Restricciones de hardware (cálculo, memoria) que impactan la velocidad, viabilidad y calidad del entrenamiento, influyendo en el tamaño y complejidad de los modelos.
Probabilidad	4 (Alta)
Impacto	5 (Extremo)
Matriz P/I	Alto/Extremo (20)
Plan Mitigación	<ul style="list-style-type: none">■ Optimización temprana del código.■ Uso de técnicas de muestreo.
Plan Contingencia	<ul style="list-style-type: none">■ Utilizar sistema de virtualización de la escuela■ Reducir complejidad de modelos.

Tabla 3.5: R03: Limitaciones de capacidad de procesamiento para el entrenamiento de los modelos.

Riesgo R04	
Título	Pérdida o corrupción del código.
Descripción	Extraviación o daño en los conjuntos de datos del proyecto que se utilizan para en entrenamiento y la validación del modelo.
Probabilidad	2 (Baja)
Impacto	5 (Extremo)
Matriz P/I	Baja/Extremo (10)
Plan Mitigación	<ul style="list-style-type: none">■ Almacenamiento redundante de los datos en repositorios.■ Verificación de la integridad de los datos con checksums.
Plan Contingencia	<ul style="list-style-type: none">■ Recuperar datasets desde backups externos.■ Regenerar datos sintéticos.

Tabla 3.6: R04: Pérdida o corrupción del código.

3.2. GESTIÓN DE RIESGOS

Riesgo R05	
Título	Disponibilidad limitada del tutor académico.
Descripción	Restricciones de tiempo y acceso al tutor académico, afectando la frecuencia y profundidad de la retroalimentación y el apoyo al estudiante en su proceso de aprendizaje.
Probabilidad	3 (Media)
Impacto	3 (Media)
Matriz P/I	Media/Media (9)
Plan Mitigación	<ul style="list-style-type: none">■ Agendar reuniones con anticipación.■ Preparar preguntas concretas.
Plan Contingencia	<ul style="list-style-type: none">■ Consultar con profesores alternativos.■ Usar foros académicos.

Tabla 3.7: R05: Disponibilidad limitada del tutor académico.

Riesgo R06	
Título	Desviaciones en la planificación temporal inicial.
Descripción	Variaciones o retrasos respecto al cronograma original, impactando los plazos de entrega, la gestión del tiempo y la consecución de los objetivos previstos.
Probabilidad	4 (Alta)
Impacto	4 (Alto)
Matriz P/I	Alto/Alto (16)
Plan Mitigación	<ul style="list-style-type: none">■ Incluir días asignados a descanso como días dedicados al proyecto.■ Revisiones semanales de progreso.
Plan Contingencia	<ul style="list-style-type: none">■ Reorganizar del diagrama de Gantt.■ Eliminar funcionalidades no críticas.

Tabla 3.8: R06: Desviaciones en la planificación temporal inicial.

Riesgo R07	
Título	Cambios en los requisitos técnicos.
Descripción	Modificaciones o alteraciones en las especificaciones necesarias para un proyecto o tarea, que pueden afectar al diseño, a la implementación, a los recursos y a los plazos.
Probabilidad	4 (Alta)
Impacto	4 (Alto)
Matriz P/I	Alto/Alto (16)
Plan Mitigación	<ul style="list-style-type: none">■ Documentar requisitos iniciales con precisión.■ Establecer procedimiento de cambio formal.
Plan Contingencia	<ul style="list-style-type: none">■ Revisar el alcance con tutor.■ Asignar tiempo adicional para cambios.

Tabla 3.9: R07: Cambios en los requisitos técnicos.

Riesgo R08	
Título	Interpretación errónea de los resultados por falta de formación en técnicas de aprendizaje automático.
Descripción	El desconocimiento de las métricas utilizadas para analizar los resultados como precisión, recall, F1, puede llevar a confiar en modelos con bajo desempeño real o tomar decisiones equivocadas basadas en correlaciones espurias.
Probabilidad	3 (Media)
Impacto	5 (Extremo)
Matriz P/I	Media/Extremo (15)
Plan Mitigación	<ul style="list-style-type: none">■ Buscar alternativas para obtener una formación básica en técnicas de aprendizaje automático.■ Corroborar los resultados obtenidos con el tutor académico.
Plan Contingencia	<ul style="list-style-type: none">■ Suspender de manera inmediata las decisiones tomadas basadas en las salidas mal interpretadas del modelo.■ Reemplazar el modelo mal interpretado por una versión anterior o un modelo base.

Tabla 3.10: R08: Dependencia de tecnologías inestables o no documentadas.

Riesgo R09	
Título	Problemas de licencia de software.
Descripción	Inconvenientes o restricciones legales relacionadas con el uso, la distribución o la activación de software, pudiendo causar interrupciones, costos adicionales o incluso acciones legales.
Probabilidad	2 (Baja)
Impacto	3 (Media)
Matriz P/I	Baja/Media (6)
Plan Mitigación	<ul style="list-style-type: none">■ Verificar licencias antes de usarlas para su implementación.■ Priorizar la utilización software open-source.
Plan Contingencia	<ul style="list-style-type: none">■ Buscar alternativas equivalentes.■ Solicitar licencias académicas.

Tabla 3.11: R10: Problemas de licencia de software.

3.3. Estimación de costes

En esta sección se presenta la estimación de costes, que comprende la identificación y valoración de los recursos necesarios para el desarrollo del proyecto. Este proceso implica la cuantificación de los gastos previsible asociados a materiales y software específico utilizado. También se considera un coste el tiempo dedicado por el estudiante a la realización del trabajo.

La precisión en la estimación de costes facilita la elaboración de un presupuesto realista y la planificación financiera del proyecto. Permite anticipar las necesidades económicas, buscar posibles fuentes de financiación si fuese necesario y gestionar eficientemente los recursos disponibles. Una estimación detallada contribuye a evitar desviaciones presupuestarias y a asegurar la viabilidad económica del proyecto [14].

3.3.1. Costes materiales

A continuación, se hace un recuento de los costes materiales en hardware y software que han sido utilizados para el desarrollo de este proyecto.

Hardware

El hardware comprende el conjunto de componentes físicos y tangibles que constituyen un sistema informático. Proporciona la infraestructura física necesaria para la ejecución del software y el procesamiento de la información.

Para realizar este proyecto se ha utilizado un equipo informático con los siguientes componentes:

- **CPU:** AMD Ryzen 7 6800HS with Radeon Graphics (16) @ 4.785GHz
- **RAM:** 16GB SO-DIMM DDR5 4800MH
- **Memoria:** 512GB PCIe® 4.0 NVMe™ M.2 SSD
- **GPU1:** NVIDIA GeForce RTX 3050 Mobile
- **GPU2:** AMD ATI Radeon 680M

Debido al tamaño del conjunto de datos, el componente que más ha ralentizado el proyecto es la RAM, que en ciertas ocasiones durante el entrenamiento de los modelos se quedaba algo escasa en capacidad.

Teniendo en cuenta que el coste del ordenador en el momento de la compra fue de 829 €, que la vida útil aproximada es de 8 años y que para el desarrollo de este proyecto se ha estado utilizando durante 3,5 meses, la amortización del hardware es:

$$\text{Amortización del Hardware} = 829 \text{ €} \times \frac{1}{8} \times \frac{3,5}{12} = 30,22 \text{ €} \quad (3.1)$$

Software

El software constituye el conjunto intangible de programas, datos e instrucciones que habilitan el funcionamiento de un sistema informático. Es el responsable de definir la funcionalidad, el comportamiento y la interacción del sistema con el usuario y con otros sistemas. En la Tabla 3.12 se detalla el coste del software utilizado para desarrollar este trabajo.

Funcionalidad	Software	Coste Mensual	Duración	Coste Total
Sistema Operativo (SO)	Kubuntu 24.10 x86_64	0 €	3,5 meses	0 €
Lenguaje (memoria)	Latex	0 €	3 meses	0 €
Editor latex	TexMaker	0 €	3 meses	0 €
IDE	MS Visual Studio Code	0 €	1 mes	0 €
Lenguaje (modelos)	Python	0 €	1 mes	0 €
IDE de Python	Jupyter Notebooks	0 €	1 mes	0 €
Plataforma MLOps ¹	Weights&Biases	0 €	1 mes	0 €
Control de versiones	GitHub	0 €	1 mes	0 €
IA generativa (código)	DeepSeek	0 €	1 mes	0 €
IA generativa (memoria)	Gemini	0 €	2 mes	0 €
Comunicación 1	MS Outlook	0 € ²	3,5 mes	0 €
Comunicación 2	MS Teams	0 €	3,5 mes	0 €

Tabla 3.12: Costes de Software.

Debido a que el proyecto se realiza en un ámbito académico, se han minimizado los costes software del proyecto utilizando exclusivamente herramientas cedidas por la entidad académica o bien herramientas con licencia open-source que no suponen un coste monetario para el desarrollo del proyecto.

3.3.2. Costes humanos

Como proyecto académico, los costes humanos representan el valor del tiempo y el esfuerzo personal invertido en la planificación, investigación, redacción y presentación del trabajo. Estos costes se manifiestan en las horas dedicadas al proyecto, el esfuerzo intelectual requerido, el aplazamiento o anulación de otras actividades personales o profesionales y el estrés asociado al proceso.

En el caso de la simulación de los costes monetarios de un proyecto similar a este, sería necesario contar con personas cualificadas para los siguientes puestos:

- Ingeniero de Aprendizaje Automático.
- Científico de Datos.

- Analista de Satos.

Los costes humanos se dividen en tres fases diferentes:

1. Definición y Preparación: la Tabla 3.13 muestra los costes para esta fase.
2. Desarrollo y Entrenamiento: los costes para esta fase vienen desarrollados en la Tabla 3.14.
3. Definición y Preparación: la Tabla 3.15 detalla los costes de esta fase del proyecto.

Para obtener una visión de conjunto de los costes humanos del proyecto, la Tabla 3.16 refleja la suma de las fases comentadas.

Fase 1: Definición y Preparación (75 horas)

Profesional	€/hora	Horas	Total (€)
Ingeniero ML	42	12	504
Científico Datos	34,5	38	1 311
Analista Datos	18,5	25	462,5
Total Fase 1		75	2 277,5

Tabla 3.13: Costes de Profesionales - Fase 1.

Fase 2: Desarrollo y Entrenamiento (150 horas)

Profesional	€/hora	Horas	Total (€)
Ingeniero ML	42	95	3 990
Científico Datos	34,5	38	1 311
Analista Datos	18,5	17	314,5
Total Fase 2		150	5 615,5

Tabla 3.14: Costes de Profesionales - Fase 2.

Fase 3: Validación y Evaluación (75 horas)

Profesional	€/hora	Horas	Total (€)
Ingeniero ML	42	30	1 260
Científico Datos	34,5	37	1 276,5
Analista Datos	18,5	8	148
Total Fase 3		75	2 684,5

Tabla 3.15: Costes de Profesionales - Fase 3.

Coste Total del Proyecto (300 horas)

Profesional	Coste Total (€)
Ingeniero ML	5 754
Científico Datos	3 898,5
Analista Datos	925
Coste Total del Proyecto	10 577,5

Tabla 3.16: Coste Total por Profesional

Las estimaciones salariales proporcionadas se basan en los salarios en el sector tecnológico y de análisis de datos en España. Esta información se encuentra publicada en portales de empleo (Talent.com³), escuelas de negocio (Aicad Business School⁴, KSchool⁵, Esden Business School⁶) y noticias del sector (Tokio School⁷).

³<https://es.talent.com/>

⁴<https://www.aicad.es/>

⁵<https://kschool.com/>

⁶<https://esden.es/>

⁷<https://www.tokioschool.com/>

Capítulo 4

Entendimiento del problema

En este capítulo se trata el entendimiento del problema. Tal y como se comenta en el capítulo dos, es la fase inicial de la metodología CRISP-DM. A continuación, se alinean los objetivos técnicos con las necesidades del negocio y con el problema a resolver. Se definen requisitos, se identifican métricas de éxito y se trata de dar comprensión sobre el contexto organizacional.

4.1. ¿Qué es un ataque a un sistema informático?

Un ataque a un sistema informático constituye una acción deliberada y no autorizada que explota vulnerabilidades con el objetivo de comprometer la confidencialidad, integridad o disponibilidad de los datos y recursos del sistema. Esta actividad maliciosa puede manifestarse a través de diversas técnicas, incluyendo la inyección de código malicioso, la denegación de servicio, el acceso no autorizado y la ingeniería social. Su ejecución busca obtener beneficios ilícitos, interrumpir operaciones o dañar la infraestructura tecnológica.

La consecuencia de un ataque puede variar desde la pérdida o alteración de información sensible hasta la paralización completa de los servicios ofrecidos por el sistema. La identificación, análisis y mitigación de estas amenazas representan un aspecto fundamental en la seguridad informática, requiriendo la implementación de medidas preventivas y reactivas para proteger los activos digitales de una organización o individuo.

4.2. Tipos de ataque a sistemas informáticos

En esta sección se mencionan algunos de los diversos tipos de ataques informáticos que existen. Los ataques a sistemas informáticos se pueden clasificar de distintas maneras según su naturaleza. A continuación, se detallan los más comunes y peligrosos:

Ataques de Denegación de Servicio (DoS) Un ataque de Denegación de Servicio tiene como objetivo hacer que un sistema o red no esté disponible para sus usuarios legítimos, interrumpiendo su funcionamiento normal. Estos ataques se llevan a cabo sobrecargando los recursos de un servidor o red con una cantidad excesiva de solicitudes. Un ejemplo de este tipo de ataque es el *Distributed Denial of Service* (DDoS), en el cual múltiples dispositivos comprometidos atacan al mismo tiempo un servidor, sobrecargándolo hasta que se vuelve inoperativo [15].

Ataques de Inyección Los ataques de inyección ocurren cuando un atacante introduce código malicioso en un sistema que no valida correctamente las entradas de los usuarios. Un ejemplo común es la inyección SQL, donde un atacante puede insertar comandos SQL en una aplicación web para acceder, modificar o eliminar datos en una base de datos. Este tipo de ataque explota la falta de medidas de validación y filtrado en las entradas del sistema [16].

Ataques de *Phishing* El *phishing* es un tipo de ataque en el que un atacante engaña a una persona para que proporcione información confidencial, como contraseñas o detalles bancarios. Generalmente, esto se realiza mediante correos electrónicos o sitios web fraudulentos que simulan ser una entidad de confianza, como bancos o servicios en línea. La principal característica del *phishing* es la manipulación psicológica de la víctima [17].

Ataques de *Malware* El *malware* es un software malicioso diseñado para dañar o robar información de un sistema informático. Entre los tipos más comunes de *malware* se encuentran los virus, gusanos, troyanos y *ransomware*. El *ransomware*, por ejemplo, cifra los archivos de un sistema y exige un pago para liberar el acceso a los mismos. Los ataques de *malware* pueden ser ejecutados a través de archivos adjuntos de correo electrónico, descargas de software malicioso o vulnerabilidades de seguridad.

Ataques *Man-in-the-Middle* (MitM) En un ataque *Man-in-the-Middle*, el atacante intercepta y, a veces, modifica la comunicación entre dos partes sin que ellas lo sepan. Estos ataques son comunes en redes no seguras, como las redes Wi-Fi públicas. El atacante puede obtener información confidencial, como contraseñas y datos de tarjetas de crédito, si no se utilizan medidas de seguridad adecuadas, como cifrado de las comunicaciones [15].

***Exploits* de Vulnerabilidades** Un *exploit* es un ataque que se aprovecha de una vulnerabilidad en el software o hardware de un sistema. Estos ataques se basan en el aprovechamiento de fallos conocidos en programas o dispositivos sin que el fabricante o el usuario los haya corregido. Los *exploits* pueden permitir a los atacantes ejecutar código malicioso en un sistema, obtener acceso a información confidencial o incluso tomar control total de la máquina afectada [17].

4.3. ¿Qué es TCP?

El Protocolo de Control de Transmisión (TCP) constituye uno de los protocolos fundamentales de la capa de transporte del modelo TCP/IP, sobre el cual se sustenta gran parte de la comunicación en redes IP, incluyendo Internet. Su diseño se orienta a proporcionar un servicio de transferencia de datos fiable, ordenado y con detección de errores entre aplicaciones que se ejecutan en sistemas finales diferentes. Para lograr esta fiabilidad, TCP establece una conexión virtual punto a punto entre las aplicaciones comunicantes mediante un proceso de *three-way handshake* que se muestra en la Figura 4.1, lo que permite la negociación de parámetros de la conexión y la sincronización de los números de secuencia iniciales.

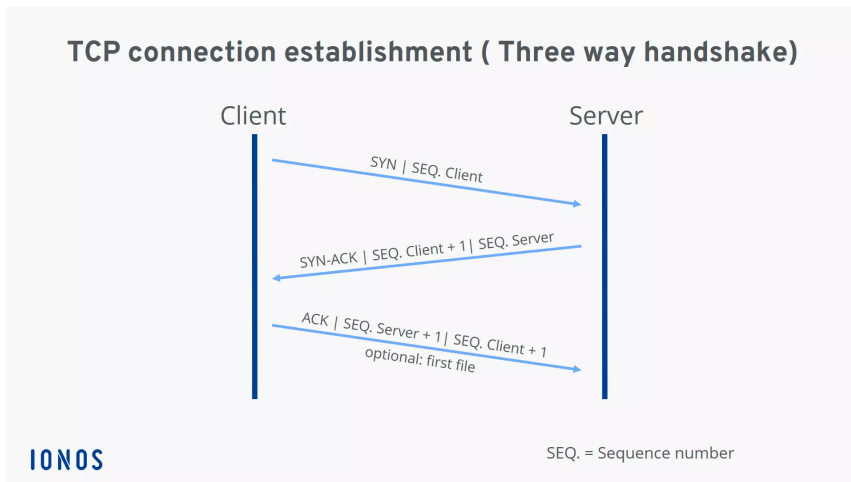


Figura 4.1: Esquema del *three way handshake* de TCP [3].

El protocolo garantiza la entrega ordenada de la información al receptor mediante la asignación de números de secuencia a cada byte transmitido, permitiendo así la reordenación en caso de que la información no llegue al receptor en el orden correcto. La fiabilidad se logra a través de un mecanismo de acuse de recibo (*acknowledgment*, ACK) positivo con retransmisión, donde el receptor confirma la recepción correcta de los paquetes de información, y el emisor retransmite aquellos partes de la información para los que no recibe confirmación dentro de un tiempo límite (*timeout*).

4.3.1. ¿Qué es un segmento TCP?

Una vez establecida la conexión, TCP divide los datos de la aplicación en unidades más pequeñas denominadas segmentos cuya estructura se detalla en la Figura 4.2. Un segmento o paquete TCP constituye la unidad de datos fundamental que se intercambia a través de una red utilizando el mencionado protocolo TCP. Este segmento encapsula una porción de los datos de la capa de aplicación, precedida por una cabecera TCP.

La cabecera TCP contiene información de control esencial para la funcionalidad del pro-

4.3. ¿QUÉ ES TCP?

toloco, incluyendo los números de puerto de origen y destino que identifican las aplicaciones comunicantes, los números de secuencia y de acuse de recibo (ACK) que garantizan la entrega ordenada y fiable, las banderas de control que indican el propósito del segmento (establecimiento de conexión, finalización, ACK, entre otros muchos), y otros campos como la ventana de recepción para el control de flujo y la suma de verificación para la detección de errores.

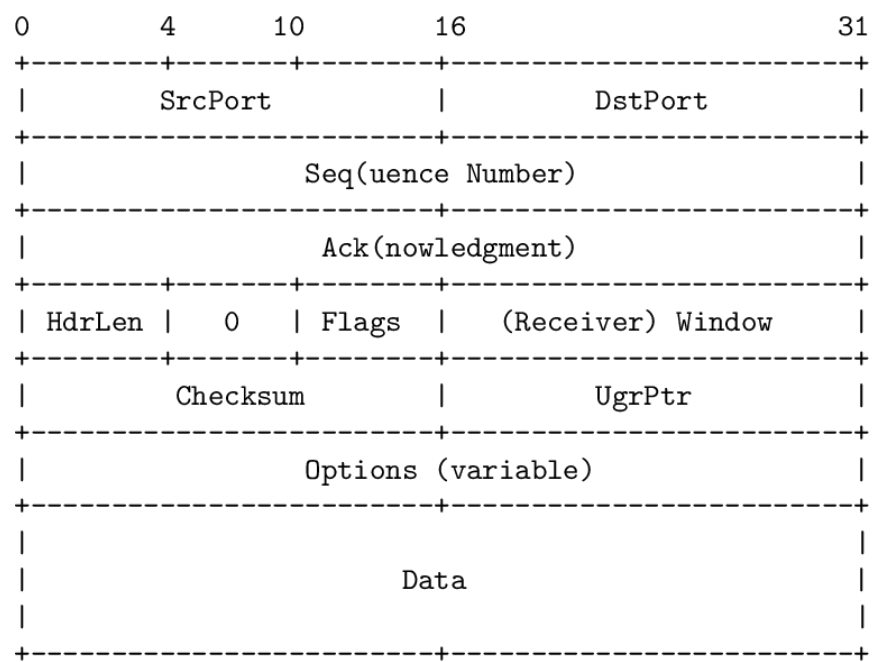


Figura 4.2: Esquema segmento TCP [4].

En el proceso de transmisión, el segmento TCP se encapsula a su vez dentro de un paquete IP (Protocolo de Internet) para su enrutamiento a través de la red. El paquete IP añade su propia cabecera con las direcciones IP de origen y destino, entre otra información necesaria para el transporte a nivel de red. Toda esta estructura se puede observar con detalle en la Figura 4.3.

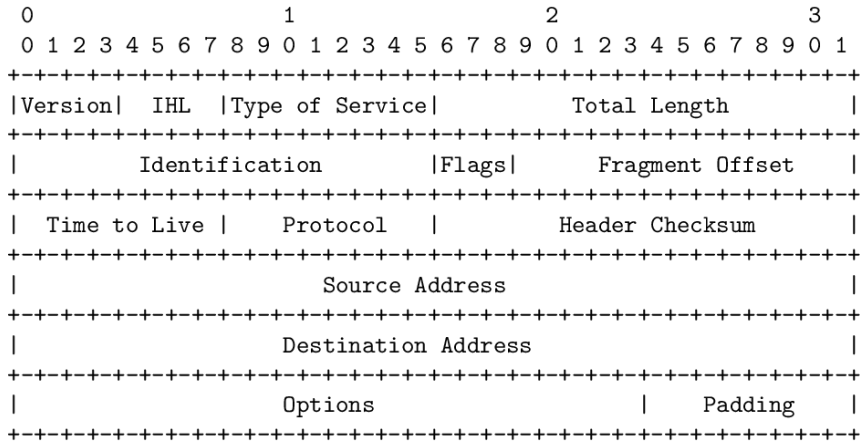


Figura 4.3: Formato de la cabecera en IPv4 [5].

4.4. Importancia de protegerse frente a un ataque

La importancia de protegerse frente a ataques informáticos radica en la salvaguarda de activos digitales críticos, la garantía de la continuidad operativa y la preservación de la confianza y la reputación. En un entorno digital cada vez más interconectado, los ataques informáticos representan una amenaza significativa para individuos, organizaciones y la sociedad en su conjunto, pudiendo acarrear consecuencias devastadoras.

Para las organizaciones, las implicaciones de un ataque informático pueden ser aún más costosas. Estas implicaciones incluyen pérdidas financieras directas debido al robo de fondos, la interrupción de las operaciones comerciales, los costes de recuperación y las posibles sanciones regulatorias. Además, se puede producir un daño significativo a la reputación y la pérdida de la confianza de los clientes, lo que a largo plazo afecta la viabilidad del negocio. Los ataques también pueden resultar en el robo de propiedad intelectual, secretos comerciales e información estratégica, otorgando ventajas competitivas a adversarios.

Por otra parte, la interrupción de servicios críticos, como energía, comunicaciones o salud, puede tener consecuencias graves para la sociedad en su conjunto.

La protección frente a ataques informáticos no es solo una cuestión de seguridad tecnológica, sino una necesidad imperante en la actualidad para proteger activos valiosos, asegurar la continuidad de las actividades, mantener la confianza de los usuarios y garantizar la estabilidad y el bienestar en el mundo digital actual. La implementación de prácticas de seguridad robustas y la concienciación sobre las amenazas cibernéticas son fundamentales en la hora de defenderse de estos ataques.

4.5. Importancia de detectar los ataques rápidamente

La detección temprana de ataques informáticos constituye un pilar fundamental en la ciberseguridad moderna debido a su capacidad para mitigar consecuencias críticas. Cuando un sistema logra identificar intrusiones o actividades maliciosas en sus fases iniciales, se reducen significativamente los daños operativos y económicos. Esta rapidez de respuesta permite contener amenazas antes de que comprometan infraestructuras completas, preservando tanto la integridad de los datos como la continuidad del negocio.

Desde una perspectiva técnica, la identificación inmediata limita la superficie de ataque, impidiendo que los actores maliciosos escalen privilegios o se propaguen lateralmente por la red. En el ámbito regulatorio, cumple con los estrictos plazos que exigen normativas como el Reglamento General de Protección de Datos (RGPD), que obliga a notificar violaciones de seguridad en un máximo de 72 horas. Además, desde el punto de vista económico, reduce los costes asociados a las reparaciones, que suelen multiplicarse exponencialmente cuando los ataques permanecen indetectados durante largos períodos.

La capacidad de detectar rápidamente anomalías en el tráfico de red, accesos no autorizados o patrones de comportamiento sospechosos no solo protege los activos digitales, sino que también salvaguarda la reputación institucional. Organizaciones con sistemas de detección temprana robustos demuestran proactividad ante clientes y socios comerciales, generando confianza en su capacidad para manejar información sensible. Esta anticipación resulta especialmente crítica en entornos donde la disponibilidad del servicio es primordial, como en las infraestructuras críticas anteriormente mencionadas.

4.6. Soluciones comerciales o actuales a estos problemas

En esta sección se comentan algunas de las soluciones y software que se utilizan en la actualidad para detectar y neutralizar posibles ataques informáticos. Estas herramientas protegen los sistemas informáticos analizando y controlando el tráfico de la red.

Los *firewalls* de próxima generación (NGFW) como Palo Alto Networks, Check Point o Cisco Firepower, inspeccionan el tráfico de red a un nivel profundo (Deep Packet Inspection - DPI), analizando el contenido de los paquetes más allá de los puertos y protocolos tradicionales. Esto permite identificar y bloquear amenazas sofisticadas, *malware*, y tráfico de aplicaciones maliciosas, además de ofrecer funcionalidades como prevención de intrusiones (IPS) y control de aplicaciones [18].

Los sistemas de detección y prevención de intrusiones o IDS e IPS, como: Snort, Suricata o Trend Micro TippingPoint, monitorean el tráfico de red en tiempo real en busca de patrones sospechosos o firmas de ataques conocidos. Los IDS alertan sobre posibles intrusiones, mientras que los IPS tienen la capacidad de bloquear o mitigar activamente el tráfico malicioso detectado, interrumpiendo los ataques en curso [19].

La microsegmentación de red con herramientas como VMware NSX, Cisco ACI o Illumio, divide la red en segmentos más pequeños y aislados, aplicando políticas de seguridad granular

a cada segmento. Esto limita el movimiento lateral de los atacantes dentro de la red una vez que han comprometido un punto inicial. Al controlar el tráfico entre estos segmentos, se reduce la superficie de ataque y se contiene la propagación de las amenazas [20].

4.7. Requisitos

Como se ha comentado en la sección [1.3 Objetivos del proyecto](#), el principal objetivo del proyecto es desarrollar un modelo neuronal que detecte la presencia de ataques en una red informática y los clasifique según su tipo. Para cumplir con dicho objetivo, se considera imprescindible cumplir con los requisitos que se listan a continuación.

4.7.1. Requisitos Funcionales

Primera versión de requisitos, no me convencen mucho

- **RF-1:** El sistema deberá detectar cuales de las conexiones podrían ser potenciales intrusiones en la red.
- **RF-2:** El sistema deberá clasificará las conexiones en 10 categorías predefinidas en [Clasificación de amenazas de seguridad](#).
- **RF-3:** El sistem deberá ser capaz de procesar formatos estándar de logs como son Syslog, NetFlow y PCAP.
- **RF-4:** El sistema deberá diferenciar entre ataques conocidos (basados en firmas) y desconocidos (basados en anomalías).
- **RF-5:** El sistema deberá ofrecer API REST para conexión con SIEMs (Splunk, IBM QRadar).
- **RF-6:** Generar alertas automatizadas con nivel de criticidad (bajo/medio/alto).
- **RF-7:** Proveer recomendaciones de mitigación básicas (ej. bloquear IPs maliciosas).
¿Debería integrar el modelo en algún sistema o crear un script o alguna forma para comunicarme con él?

4.7.2. Requisitos No Funcionales

- **RNF-1:** Latencia ¡50 ms en redes de 10Gbps (requisito crítico para SOC [21]).
- **RNF-2:** Interfaz accesible para usuarios no técnicos (evaluado con test SUS [22]).

4.8. **Objetivos del proyecto ¿HACE FALTA TENIENDO EN LA INTRODUCCIÓN?**

4.9. **¿DEBERÍA AÑADIR ALGÚNA SECCIÓN NUEVA?**

Capítulo 5

Entendimiennto de los datos

Este capítulo se corresponde con la segunda etapa de la metodología CRISP-DM, En el se explicará la naturaleza de los datos y sus características, así como los valores atípicos que presentan y sus sesgos.

5.1. Origen de los datos

Los datos que se han utilizado para desarrollar este trabajo, se han obtenido de conjuntos de datos diseñados para entrenar Sistemas de Detección de Intrusión de Red (NIDS) basados en el aprendizaje automático. El *dataset* en cuenstión forma parte de un análisis realizado en la Universidad de Queensland, Australia [23].

El *dataset* utilizado es NF-UNSW-NB15-v3, este es una versión basada en NetFlow del conocido conjunto de datos UNSW-NB15, mejorada con características adicionales de NetFlow y etiquetada de acuerdo con sus respectivas categorías de ataque.

5.2. Tipos de ataques registrados en los datos

En esta sección se explican los tipos de datos presentes en el *dataset* que se utiliza para entrenar a los modelos del proyecto. Se explica en que consiste cada tipo de ataque registrado así como el número exacto de ataques de cada tipo presente.

El conjunto de datos consiste en un total de 2 365 424 flujos de datos, donde 127 639 (5,4%) son muestras de ataque y 2 237 731 (94,6%) son benignos. Los flujos de ataque se clasifican en nueve clases, cada una representando una amenaza a la red distinta. La Tabla 5.1 proporciona una distribución detallada del conjunto de datos:

Clase	Cantidad	Descripción
Benigno	2 237 731	Flujos normales no maliciosos.
<i>Fuzzers</i>	33 816	Tipo de ataque en el que el atacante envía grandes cantidades de datos aleatorios que hacen que un sistema se bloquee y también apuntan a descubrir vulnerabilidades de seguridad en un sistema.
<i>Analysis</i>	2 381	Un grupo que presenta una variedad de amenazas que se dirigen a aplicaciones web a través de puertos, correos electrónicos y <i>scripts</i> .
<i>Backdoor</i>	1 226	Una técnica que tiene como objetivo eludir los mecanismos de seguridad respondiendo a aplicaciones específicas de clientes contruidos.
DoS	5 980	La denegación de servicio es un intento de sobrecargar los recursos de un sistema informático con el objetivo de evitar el acceso o la disponibilidad de sus datos.
<i>Exploits</i>	42 748	Son secuencias de comandos que controlan el comportamiento de un <i>host</i> a través de una vulnerabilidad conocida.
<i>Generic</i>	19 651	Un método que se dirige a la criptografía y causa una colisión con cada cifrado de bloques.
<i>Reconnaissance</i>	17 074	Una técnica para recopilar información sobre un <i>host</i> de red, también se conoce como sonda.
<i>Shellcode</i>	4 659	Un <i>malware</i> que penetra en un código para controlar el <i>host</i> de una víctima.
<i>Worms</i>	158	Ataques que se replican y se extienden a otros sistemas.

Tabla 5.1: Clasificación de amenazas de seguridad.

5.3. Parámetros de los datos

En esta sección se explicarán el significado de cada uno de los 55 parámetros que componen cada fila del *dataset* seleccionado.

<https://arxiv.org/pdf/2503.04404>

- **IPV4_SRC_ADDR**: Dirección IPv4 de origen.
- **IPV4_DST_ADDR**: Dirección IPv4 de destino.

- **L4_SRC_PORT**: Número de puerto de origen de la capa 4.
- **L4_DST_PORT**: Número de puerto de destino de la capa 4.
- **PROTOCOL**: Byte identificador del protocolo IP.
- **L7_PROTO**: Protocolo de aplicación (numérico) de la capa 7.
- **IN_BYTES**: Número de bytes entrantes.
- **OUT_BYTES**: Número de bytes salientes.
- **IN_PKTS**: Número de paquetes entrantes.
- **OUT_PKTS**: Número de paquetes salientes.
- **FLOW_DURATION_MILLISECONDS**: Duración del flujo en milisegundos.
- **TCP_FLAGS**: Acumulativo de todos los *flags* TCP.
- **CLIENT_TCP_FLAGS**: Acumulativo de todos los *flags* TCP del cliente.
- **SERVER_TCP_FLAGS**: Acumulativo de todos los *flags* TCP del servidor.
- **DURATION_IN**: Duración del flujo Cliente a Servidor (mseg).
- **DURATION_OUT**: Duración del flujo Cliente a Servidor (mseg).
- **MIN_TTL**: TTL mínimo del flujo.
- **MAX_TTL**: TTL máximo del flujo.
- **LONGEST_FLOW_PKT**: Paquete más largo (bytes) del flujo.
- **SHORTEST_FLOW_PKT**: Paquete más corto (bytes) del flujo.
- **MIN_IP_PKT_LEN**: Longitud del paquete IP más pequeño del flujo observado.
- **MAX_IP_PKT_LEN**: Longitud del paquete IP más grande del flujo observado.
- **SRC_TO_DST_SECOND_BYTES**: *Bytes*/segundo de origen a destino.
- **DST_TO_SRC_SECOND_BYTES**: *Bytes*/segundo de destino a origen.
- **RETRANSMITTED_IN_BYTES**: Número de bytes TCP retransmitidos del flujo (src a dst).
- **RETRANSMITTED_IN_PKTS**: Número de paquetes TCP retransmitidos del flujo (src a dst).
- **RETRANSMITTED_OUT_BYTES**: Número de bytes TCP retransmitidos del flujo (dst a src).
- **RETRANSMITTED_OUT_PKTS**: Número de paquetes TCP retransmitidos del flujo (dst a src).

- **SRC_TO_DST_AVG_THROUGHPUT**: Tasa de transferencia promedio de origen a destino (bps).
- **DST_TO_SRC_AVG_THROUGHPUT**: Tasa de transferencia promedio de destino a origen (bps).
- **NUM_PKTS_UP_TO_128_BYTES**: Paquetes cuyo tamaño IP es ≤ 128 .
- **NUM_PKTS_128_TO_256_BYTES**: Paquetes cuyo tamaño IP es > 128 y ≤ 256 .
- **NUM_PKTS_256_TO_512_BYTES**: Paquetes cuyo tamaño IP es > 256 y ≤ 512 .
- **NUM_PKTS_512_TO_1024_BYTES**: Paquetes cuyo tamaño IP es > 512 y ≤ 1024 .
- **NUM_PKTS_1024_TO_1514_BYTES**: Paquetes cuyo tamaño IP es > 1024 y ≤ 1514 .
- **TCP_WIN_MAX_IN**: Ventana TCP máxima (src a dst).
- **TCP_WIN_MAX_OUT**: Ventana TCP máxima (dst a src).
- **ICMP_TYPE**: Tipo ICMP * 256 + Código ICMP.
- **ICMP_IPV4_TYPE**: Tipo ICMP.
- **DNS_QUERY_ID**: ID de transacción de la consulta DNS.
- **DNS_QUERY_TYPE**: Tipo de consulta DNS (ej., 1=A, 2=NS..).
- **DNS_TTL_ANSWER**: TTL del primer registro A (si existe).
- **FTP_COMMAND_RET_CODE**: Código de retorno del comando del cliente FTP.
- **FLOW_START_MILLISECONDS**: Marca de tiempo de inicio del flujo en milisegundos.
- **FLOW_END_MILLISECONDS**: Marca de tiempo de fin del flujo en milisegundos.
- **SRC_TO_DST_IAT_MIN**: IAT mínimo (src a dst).
- **SRC_TO_DST_IAT_MAX**: IAT máximo (src a dst).
- **SRC_TO_DST_IAT_AVG**: IAT promedio (src a dst).
- **SRC_TO_DST_IAT_STDDEV**: Desviación estándar del IAT (src a dst).
- **DST_TO_SRC_IAT_MIN**: IAT mínimo (dst a src).
- **DST_TO_SRC_IAT_MAX**: IAT máximo (dst a src).
- **DST_TO_SRC_IAT_AVG**: IAT promedio (dst a src).
- **DST_TO_SRC_IAT_STDDEV**: Desviación estándar del IAT (dst a src).
- **Label**: Toma valor 0 si el flujo es benigno y 1 si se trata de un ataque.
- **Attack**: Clase de flujo, perteneciente a las clases mencionadas en [5.1 Clasificación de amenazas de seguridad](#)

5.4. Patrones preliminares, valores atípicos y sesgos

En esta sección se comentan los patrones preliminares, así como los valores y los sesgos presentes en el *dataset*. La identificación de estas irregularidades es fundamental para el correcto desarrollo del proyecto.

Al analizar los datos originales del dataset, se encuentran características que afectan de forma negativa al entrenamiento del modelo y por lo tanto, a su correcto funcionamiento. A continuación, se mencionan cuales son las características problemáticas de los datos que se utilizan.

Tras estudiar los datos, se descubre que algunos parámetros presentan valores infinitos, estos valores alteran la distribución inherente de las variables, introduciendo sesgos significativos en el proceso de aprendizaje. Los algoritmos de entrenamiento, diseñados para operar con valores numéricos finitos, pueden comportarse de manera impredecible o inestable ante la presencia de infinitos, dificultando la convergencia hacia una solución óptima. Asimismo, la interpretación de las características con valores infinitos se vuelve ambigua, comprometiéndola la capacidad del modelo para establecer relaciones significativas con otras variables y para generalizar correctamente a datos futuros que no contengan tales valores extremos. En consecuencia, la presencia de infinitos puede degradar sustancialmente el rendimiento y la fiabilidad del modelo entrenado.

Como se comenta en la sección anterior [5.3 Parámetros de los datos](#), existen dos parámetros que registran la IP origen y la IP destino de la conexión. En principio, si los datos no fuesen sintéticos, la dirección IP de la máquina que origina la comunicación sería aleatoria. En este conjunto de datos, se puede observar como todos los ataques provienen de IP con máscara 175.45.176.255, lo que no encaja con la realidad. Independientemente de este patrón, que no se manifiesta de forma natural en las conexiones entre sistemas, el uso de las IPs de las máquinas en el entrenamiento de los modelos provoca que el algoritmo encargado de este entrenamiento asigne pesos de manera incorrecta a un parámetro que no influye en la clasificación que se propone para este trabajo.

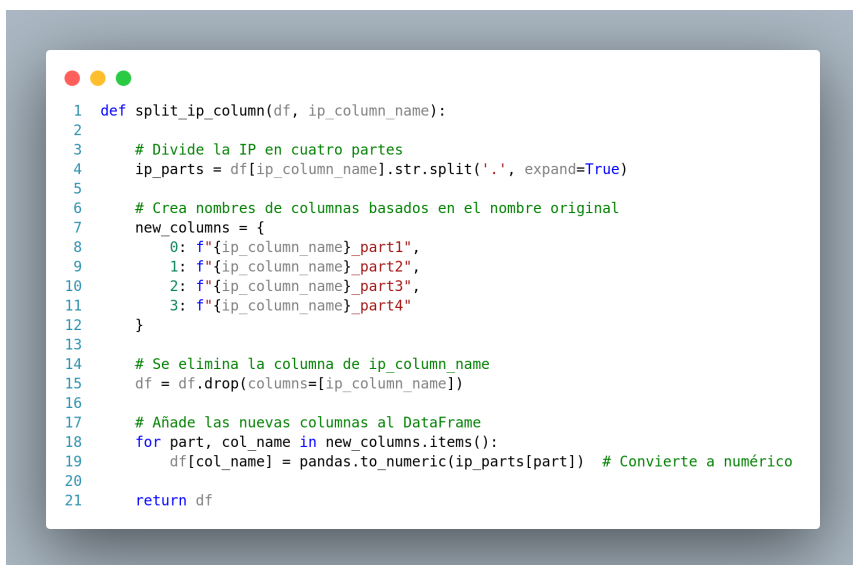
5.5. Preparación de los datos ¿SEPARAR EN OTRO CAPÍTULO?

En esta sección se modificarán los datos que presentan patrones, valores atípicos o sesgos que se comentan en la sección anterior, [5.4 Patrones preliminares, valores atípicos y sesgos](#). Así como el tratamiento necesario de los datos para poder entrenar los modelos.

Una de las características problemáticas, es la existencia de parámetros con valores infinito. Para tratar con esta problemática existen dos enfoques, o bien se sustituyen los valores infinitos por el máximo valor posible para ese parámetro, o bien, se elimina el flujo al que pertenece cada valor infinito. Para este trabajo, se opta por eliminar los datos con valores infinitos, ya que asignarles un valor falso, podría crear una relación sintética entre los datos y provocar que el modelo no converja en una solución real.

Para entrenar un modelo es necesario que en primer lugar todos los parámetros sean numéricos. En los datos originales del *dataset*, se encuentran tres parámetros que no presentan un formato numérico: IPV4_SRC_ADDR, IPV4_DST_ADDR y Attack.

- Para dar un formato correcto al parámetro Attack, se utiliza el codificador LabelEncoder de la biblioteca sklearn.preprocessing y el método fit_transform de la biblioteca sklearn. La primera función, LabelEncoder codifica las etiquetas de características categóricas en valores numéricos entre 0 y el número de clases menos 1. Una vez se instancian las categorías del parámetro, el método fit_transform nos permite entrenar el codificador y transformar el conjunto de datos en un único paso.
- Las direcciones IPv4 están formadas por 4 octetos separados por puntos. Obviamente este formato no es numérico, por ese motivo se opta por dividir las direcciones IPv4 en 4 parámetros diferentes, uno por cada octeto. De esta manera, si el valor de un flujo para el parámetro IPV4_SRC_ADDR es 175.45.176.23, se separa en cuatro nuevos valores que son: 175, 45, 176 y 23. Para realizar esta transformación se utiliza la función [5.1 Función de transformación para los parámetros IPv4](#).



```
1 def split_ip_column(df, ip_column_name):
2
3     # Divide la IP en cuatro partes
4     ip_parts = df[ip_column_name].str.split('.', expand=True)
5
6     # Crea nombres de columnas basados en el nombre original
7     new_columns = {
8         0: f"{ip_column_name}_part1",
9         1: f"{ip_column_name}_part2",
10        2: f"{ip_column_name}_part3",
11        3: f"{ip_column_name}_part4"
12    }
13
14    # Se elimina la columna de ip_column_name
15    df = df.drop(columns=[ip_column_name])
16
17    # Añade las nuevas columnas al DataFrame
18    for part, col_name in new_columns.items():
19        df[col_name] = pandas.to_numeric(ip_parts[part]) # Convierte a numérico
20
21    return df
```

Figura 5.1: Función de transformación para los parámetros IPv4

Una vez que todos los parámetros presentan valores numéricos, se separan los parámetros entre los que formarán parte de la entrada del modelo neuronal (a partir de ahora se les menciona X) y los que contienen el valor de la salida que debe proporcionar el modelo neuronal (a partir de ahora se les menciona Y).

Los parámetros que pertenecen a Y son Label y Attack. Sin embargo, en función del modelo que se entrena se utiliza o bien Label, o bien Attack, pero nunca los dos al mismo tiempo.

Por su parte, X la conforman el total de los parámetros del Dataset a excepción de:

- **Label y Attack:** Se tratan de los parámetros que conforman Y , si estos parámetros formasen parte de los datos de entrada del modelo, el algoritmo encargado de entrenarlo, les asignaría a estos parámetros unos pesos muy elevados y tendría una tasa teórica de éxito del 100 %, puesto que, se estaría introduciendo la respuesta al problema que se aborda directamente como entrada al modelo. En un sistema informático real esta práctica es imposible.
- **IPV4_SRC_ADDR y IPV4_DST_ADDR:** Tal y como se menciona en la sección [5.3 Parámetros de los datos](#), estos parámetros presentan un patrón irregular y sintético. Al introducir estos parámetros como entradas para el entrenamiento del modelo, de forma análoga a como sucede con los parámetros Label y Attack se estarían introduciendo valores que no se pueden obtener en una situación realista y condicionando el comportamiento del modelo. Esto implica como se menciona en los siguientes capítulos que el modelo resultante obtendría muy buenas métricas en validación pero muy malas al someterlo a datos que nunca ha visto durante su entrenamiento.
- **FLOW_START_MILLISECONDS y FLOW_END_MILLISECONDS:** A diferencia de los anteriores parámetros, estos muestran valores que puedan condicionar el entrenamiento de los modelos ni sus resultados, pero se trata de parámetros que ofrecen información redundante, pues el parámetro FLOW_DURATION_MILLISECONDS es combinación lineal de estos dos. Al eliminar estos parámetros se agiliza el entrenamiento del modelo y su tiempo de respuesta una vez entrenado sin perjudicar su precisión.

Para finalizar con el tratamiento de los datos, se normalizan los valores de X para que no haya tanta varianza entre los parámetros que lo conforman. Para normalizar los valores se utiliza el escalador `MinMaxScaler` que forma parte de la biblioteca `sklearn.preprocessing` junto con el método `fit_transform` que se menciona en la explicación de la transformación del parámetro `Attack`. Llamando a `MinMaxScaler` con un rango (0,1), tras haber tratado los datos con el método `fit_transform`, se obtienen valores escalados en un rango entre 0 y 1, lo que facilita la comparación y el procesamiento por parte de los algoritmos de aprendizaje automático sin alterar las relaciones proporcionales inherentes en los datos originales.

Capítulo 6

Modelos

6.1. ¿Qué es un modelo neuronal?

En esta sección, se explica qué es un modelo neuronal, sus características y cómo funciona en el contexto de la inteligencia artificial.

Los modelos neuronales son estructuras computacionales inspiradas en el cerebro humano, diseñadas para procesar información mediante una red de unidades interconectadas que imitan, de manera simplificada, la forma en que las neuronas biológicas se comunican entre sí [citegoodfellow2016deep](#).

Los descubrimientos del Premio Nobel de biología español, Santiago Ramón y Cajal sobre la estructura y funcionamiento de las neuronas, proporcionaron las bases para la comprensión del sistema nervioso en el que se basan los modelos de inteligencia artificial. A finales del siglo XIX, Cajal formuló la teoría de que las neuronas son células individuales conectadas por sinapsis, una idea que revolucionó la neurociencia y sirvió de inspiración para los modelos computacionales de redes neuronales. Su trabajo permitió comprender cómo las señales eléctricas viajan entre las neuronas y cómo se pueden formar conexiones adaptativas, conceptos que más tarde serían adoptados en el diseño de redes neuronales artificiales.

Un modelo neuronal es una arquitectura matemática que busca resolver problemas complejos de aprendizaje mediante el procesamiento de datos. Está compuesto por una serie de unidades de procesamiento, conocidas como neuronas artificiales, organizadas en capas. Cada capa recibe la salida de la capa anterior, aplica una función matemática sobre los datos y transmite el resultado a la siguiente capa. Este proceso se repite de manera secuencial hasta que se alcanza la capa de salida, que proporciona el resultado final del modelo [\[24\]](#).

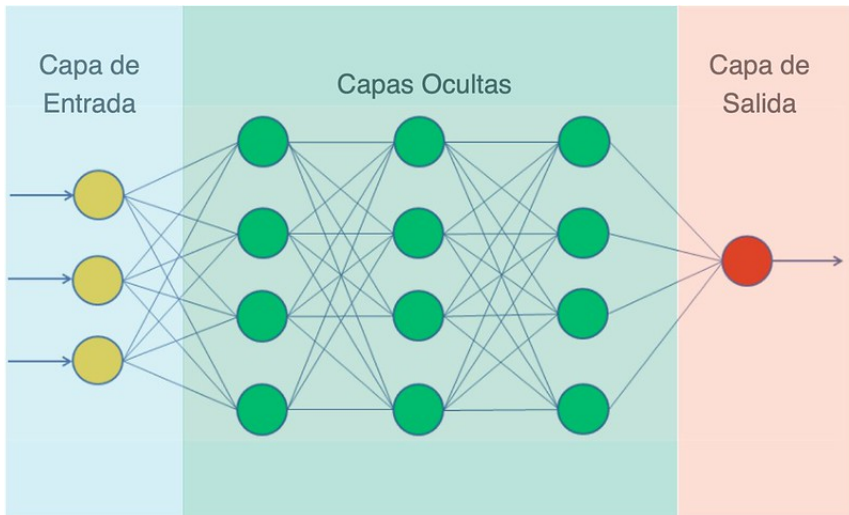


Figura 6.1: Esquema de redes neuronales. [6]

Las neuronas dentro de un modelo neuronal no operan de forma aislada, sino que están conectadas entre sí a través de enlaces denominados "pesos". Estos pesos determinan la importancia de la señal que se transmite de una neurona a otra. Durante el proceso de entrenamiento, los pesos se ajustan con el objetivo de minimizar el error en la salida del modelo, lo que permite que el modelo "aprenda" de los datos y mejore su capacidad para predecir o clasificar nueva información [25].

El proceso de entrenamiento de un modelo neuronal implica la retroalimentación o back-propagation, donde el error de la predicción se calcula y se distribuye hacia atrás a través de la red para ajustar los pesos de manera que el modelo se optimice progresivamente. En este sentido, el modelo neuronal tiene la capacidad de adaptarse a distintos tipos de datos y mejorar su rendimiento con el tiempo [26].

A grandes rasgos, un modelo neuronal es una estructura computacional que imita el funcionamiento del cerebro humano para procesar y aprender de datos, y se utiliza en tareas como clasificación, predicción y reconocimiento de patrones.

6.1.1. ¿Qué tipos de modelos neuronales existen?

En esta sección se explican cuales son tipos de modelos neuronales existen y sus características principales.

Los modelos neuronales pueden clasificarse según su estructura, la forma en que procesan la información y la aplicación específica a la que están destinados. Estos modelos neuronales son fundamentales para resolver una variedad de problemas en áreas relacionadas con la inteligencia artificial como la visión por computadora, el procesamiento del lenguaje natural y el reconocimiento de patrones.

Uno de los tipos más comunes de modelos neuronales es el perceptrón multicapa (MLP), que está formado por varias capas de neuronas organizadas en una estructura jerárquica. Cada capa en un MLP recibe la salida de la capa anterior y la procesa mediante una función de activación antes de pasar el resultado a la siguiente capa. Este tipo de red se utiliza principalmente para tareas de clasificación y regresión, y su entrenamiento se realiza utilizando algoritmos como el de retropropagación. Es el tipo de modelo que se utiliza en este proyecto para obtener un modelo capaz de detectar intrusiones en una red informática. [27]

Otro tipo importante de modelo neuronal es la red neuronal convolucional (CNN), que está diseñada específicamente para procesar datos con una estructura de cuadrícula, como las imágenes. En una CNN, las neuronas están organizadas en capas convolucionales que aplican filtros a los datos de entrada para extraer características relevantes, como bordes, texturas o formas. Esta estructura permite que las CNN sean altamente eficaces para tareas como el reconocimiento de imágenes y la visión por computadora [27].

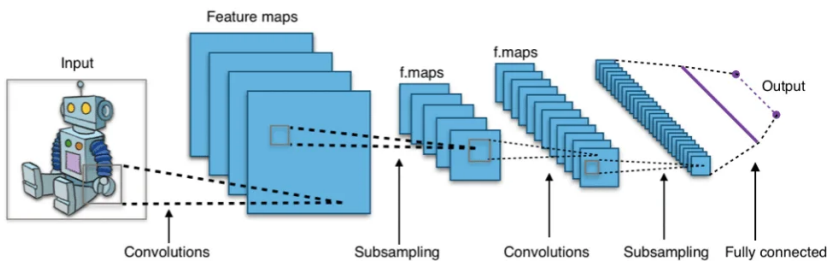


Figura 6.2: Esquema de redes neuronal convolucional. [7]

Las redes neuronales recurrentes (RNN), por otro lado, son adecuadas para procesar secuencias de datos, como el texto o el audio. Las RNN son únicas porque sus neuronas tienen conexiones que permiten que la información fluya hacia atrás a lo largo del tiempo, lo que las hace útiles para modelar dependencias temporales en los datos. Este tipo de red es comúnmente utilizado en tareas de procesamiento de lenguaje natural y en modelos de predicción de series temporales [25].

En un enfoque más avanzado, existen las redes generativas antagónicas (GAN), que se componen de dos redes neuronales: un generador y un discriminador. El generador crea datos falsos a partir de ruido, mientras que el discriminador intenta diferenciar entre los datos reales y los generados. A través de un proceso de entrenamiento competitivo, ambas redes mejoran en sus respectivas tareas. Las GANs se utilizan principalmente en la generación de imágenes, música y otros tipos de contenido artístico [26].

6.1.2. Función de pérdida

En esta sección se explica que es una función de pérdida y su importancia en el proceso de entrenamiento de un modelo neuronal. También se justifica el uso de las funciones de pérdida que se emplean para el entrenamiento de los modelos de este trabajo.

Una función de pérdida es un componente fundamental en el entrenamiento de modelos de aprendizaje automático, cuya finalidad consiste en cuantificar la discrepancia entre las predicciones generadas por el modelo y los valores reales esperados. Esta medida permite guiar el proceso de optimización, ya que el objetivo durante el entrenamiento es minimizar dicha pérdida para mejorar la precisión del modelo.[28]

Las funciones de pérdida desempeñan un papel central en la formulación matemática del aprendizaje supervisado, al establecer una métrica que penaliza el error cometido por el modelo. Esta penalización permite que los algoritmos de optimización, como el descenso por gradiente, ajusten iterativamente los parámetros del modelo en la dirección que reduce la pérdida total. Para que esta estrategia sea efectiva, la función de pérdida debe poseer ciertas propiedades fundamentales:

- **Diferenciabilidad:** Permite el cálculo de gradientes necesarios para optimizar los parámetros mediante técnicas basadas en derivadas, como el descenso por gradiente.
- **Convexidad (deseable):** Favorece la convergencia hacia un mínimo global, aunque en la práctica, muchos modelos no lineales presentan funciones de pérdida no convexas.
- **Estabilidad numérica:** Previene errores computacionales al manejar valores extremos o transformaciones exponenciales, manteniendo precisión durante el entrenamiento.
- **Sensibilidad al error:** Garantiza que los errores mayores se penalicen de forma más significativa, orientando el modelo hacia mejores predicciones.
- **Escalabilidad:** Permite su aplicación eficiente en contextos con grandes volúmenes de datos y arquitecturas de red complejas.

Estas propiedades aseguran que la función de pérdida cumpla su objetivo central, actuar como un mecanismo confiable y eficiente para guiar la actualización de los parámetros del modelo. [29]

En el caso específico del modelo de clasificación binaria, se opta por la función BCEWithLogitsLoss (Binary Cross Entropy with Logits Loss), que está especialmente recomendada para utilizarse en entornos de aprendizaje profundo como PyTorch, entorno utilizado en el desarrollo práctico de este proyecto. Esta función combina en una sola operación dos pasos fundamentales: la aplicación de la función sigmoide y el cálculo de la entropía cruzada binaria. Al integrar ambos procedimientos en una única función, se obtienen varias ventajas prácticas.[30]

En primer lugar, se mejora la estabilidad numérica, ya que evita operaciones redundantes que podrían dar lugar a pérdidas de precisión, especialmente al manejar las salidas no normalizadas del modelo (logits) con valores extremos.

En segundo lugar, se optimiza el rendimiento computacional, al reducir el número de transformaciones necesarias antes del cálculo de la pérdida.

Finalmente, permite trabajar directamente con logits, lo cual simplifica la implementación y reduce errores potenciales derivados de transformaciones incorrectas. [31]

6.1.3. Algoritmo de optimización

En esta sección se explica qué es un algoritmos de optimización, su funcionamiento, sus principales características, su importancia dentro del proceso de entrenamiento de modelos neuronales y algunos ejemplo de los algoritmos de optimización más utilizados.

Un algoritmo de optimización es un conjunto de procedimientos matemáticos y computacionales cuyo propósito es encontrar el mejor valor posible (o el más cercano a este) para un conjunto de parámetros o variables dentro de un problema específico. Estos algoritmos buscan minimizar o maximizar una función objetivo, que representa el rendimiento del sistema o modelo a optimizar. En el contexto de entrenamiento de redes neuronales, la función objetivo comúnmente se refiere a una función de pérdida, que cuantifica la diferencia entre las predicciones del modelo y las etiquetas reales de los datos. Los algoritmos de optimización ajustan los parámetros (pesos y sesgos) de la red neuronal durante el proceso de entrenamiento con el fin de reducir dicha diferencia [27].

El funcionamiento de un algoritmo de optimización se basa en un proceso iterativo en el que se ajustan los parámetros de la red neuronal utilizando información acerca del gradiente de la función de pérdida con respecto a estos parámetros. En este contexto, la retropropagación juega un papel crucial, ya que es el método utilizado para calcular esos gradientes. A través de la retropropagación, el error de la red se propaga hacia atrás desde la capa de salida hasta las capas anteriores, permitiendo calcular cómo cada parámetro contribuye al error total. Posteriormente, el algoritmo de optimización utiliza estos gradientes calculados por la retropropagación para actualizar los parámetros de la red en la dirección opuesta al gradiente, con el objetivo de reducir progresivamente la pérdida. Generalmente, se utilizan técnicas de descenso de gradiente para realizar estos ajustes. Durante cada iteración, el algoritmo calcula el gradiente de la función de pérdida a través de la retropropagación y lo utiliza para modificar los parámetros de manera eficiente, buscando alcanzar el mínimo global o local de la función de pérdida [32].

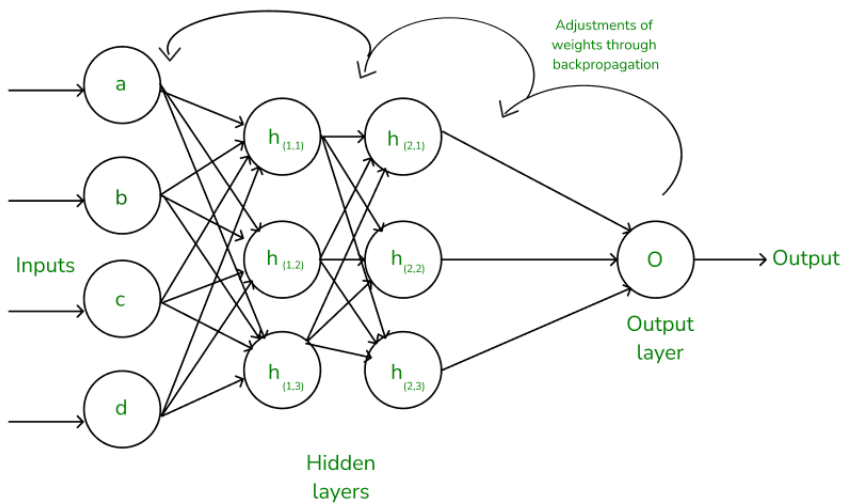


Figura 6.3: Esquema de algoritmo de optimización usando los gradiente calculados por la retropropagación. [8]

Las características clave de los algoritmos de optimización incluyen la tasa de aprendizaje, la capacidad para escapar de mínimos locales, y la convergencia en un tiempo razonable. La tasa de aprendizaje (learning rate o lr) es un hiperparámetro crítico que determina la magnitud de los ajustes que se realizan en los parámetros del modelo durante cada iteración. Si el learning rate es muy alto, el algoritmo puede saltar el mínimo global, por otro lado, si es muy baja, el proceso puede volverse muy lento. Los métodos de optimización varían en cuanto a la forma en que gestionan estas características, algunos métodos mejoran la eficiencia utilizando técnicas como el momento, la normalización del gradiente, o el uso de segunda orden [33].

La importancia de los algoritmos de optimización en el entrenamiento de redes neuronales es fundamental, ya que la capacidad de un modelo para aprender de los datos y generalizar de manera efectiva depende en gran medida de cómo se optimizan los parámetros de la red. Un algoritmo de optimización bien elegido y configurado puede significar la diferencia entre un modelo con un rendimiento subóptimo y uno que logra altos niveles de precisión y eficiencia. De esta manera, la correcta selección y ajuste de estos algoritmos se considera un aspecto crucial para el desarrollo exitoso de modelos de aprendizaje profundo y redes neuronales [27].

Existen varios algoritmos de optimización ampliamente utilizados en el entrenamiento de redes neuronales. Entre los más comunes se encuentran el Descenso de Gradiente Estocástico (SGD), Momentum, Adagrad, RMSprop y Adam. El SGD es uno de los más simples y fundacionales, pero sufre de inestabilidad debido a su naturaleza estocástica y la falta de adaptación de la tasa de aprendizaje. Momentum ayuda a mejorar la convergencia al incorporar la información de los gradientes pasados, lo que permite que el modelo escape de los mínimos locales más fácilmente. Adagrad adapta la tasa de aprendizaje para cada parámetro, mientras que RMSprop ajusta la tasa de aprendizaje utilizando una media móvil del

gradiente, lo que mejora la estabilidad [32].

Sin embargo, el algoritmo más recomendado para entrenar un modelo de clasificación binaria es Adam (Adaptive Moment Estimation). Adam combina las ventajas del descenso de gradiente estocástico y el momento, adaptando la tasa de aprendizaje de manera eficiente para cada parámetro del modelo. Esto lo hace particularmente adecuado para tareas de clasificación binaria, ya que garantiza una convergencia más rápida y estable, minimizando el riesgo de sobreajuste (overfitting) mientras se optimizan los parámetros del modelo. Sin embargo, una variante reciente de Adam, conocida como AdamW, introduce un cambio importante en la forma en que se maneja la regularización de los parámetros. AdamW separa explícitamente la actualización de los pesos y la regularización (típicamente L2 o weight decay), lo que mejora la eficacia de la regularización sin interferir con la adaptación de la tasa de aprendizaje.

Aunque ambos algoritmos, Adam y AdamW, tienen un rendimiento destacado en tareas de clasificación binaria, AdamW se ha demostrado más eficaz en la prevención de sobreajuste en redes neuronales complejas. Esto se debe a su tratamiento explícito del weight decay, lo que permite una mejor generalización en problemas con grandes volúmenes de datos como el que se trata en este proyecto. Sin embargo, Adam sigue siendo una opción popular y efectiva, especialmente cuando se trabaja con redes neuronales menos complejas o cuando se desea una mayor estabilidad y rapidez de convergencia. En general, AdamW es preferido en situaciones donde la regularización adecuada es crucial, pero Adam sigue siendo una opción sólida y eficiente. [33].

6.2. Parámetros e hiperparámetros

En este apartado se explica que son los parámetros y los hiperparámetros de un modelo neuronal. También se listan y describen los principales hiperparámetros que se deben tener en cuenta al entrenar un modelo neuronal.

Los parámetros de un modelo neuronal son los valores ajustables que el modelo aprende a partir de los datos durante el proceso de entrenamiento. Estos parámetros incluyen los pesos y sesgos de las conexiones entre las neuronas de las diferentes capas de la red neuronal. Los pesos determinan la importancia de las entradas de cada neurona, mientras que los sesgos permiten ajustar la salida de cada neurona antes de la activación. Como se explica en los apartados anteriores, el aprendizaje de estos parámetros se realiza mediante algoritmos de optimización como el gradiente descendente, los cuales buscan minimizar la función de pérdida. [27]

Los hiperparámetros son los valores que se configuran antes del entrenamiento y no se ajustan directamente durante el proceso de aprendizaje. Este tipo de parámetros controlan el comportamiento del proceso de entrenamiento y afectan la capacidad del modelo para aprender patrones complejos de los datos. A diferencia de los parámetros, los hiperparámetros deben ser establecidos manualmente o mediante técnicas de optimización específicas, como la búsqueda en cuadrícula o la búsqueda aleatoria, técnicas que se aplican en las siguientes secciones de este documento. Los hiperparámetros influyen en aspectos clave como la velo-

cidad de aprendizaje, la regularización y la estructura de la red [34]. Los más importantes son:

- **Tasa de aprendizaje (Learning rate):** Controla la magnitud de los ajustes realizados en los pesos durante cada iteración del entrenamiento. Como se comenta en [6.1.3 Algoritmo de optimización](#), una tasa demasiado alta puede hacer que el modelo no converja, mientras que una tasa demasiado baja puede llevar a una convergencia muy lenta.
- **Número de épocas (Epochs):** El número de épocas se refiere a cuántas veces el algoritmo de entrenamiento recorre todo el conjunto de datos. Un número adecuado de épocas permite que el modelo aprenda de manera efectiva, sin embargo, demasiadas épocas pueden llevar al sobreajuste.
- **Tamaño del lote (Batch size):** El tamaño del lote determina la cantidad de ejemplos utilizados en cada actualización de los pesos. Un tamaño de lote pequeño puede hacer que el modelo sea más ruidoso y difícil de entrenar, mientras que un tamaño de lote grande puede hacer que el entrenamiento sea más estable, pero menos eficiente.
- **Número de capas ocultas (Hidden layers):** Este hiperparámetro define la profundidad de la red neuronal, es decir, el número de capas intermedias entre la capa de entrada y la capa de salida. Un mayor número de capas permite modelar funciones más complejas, pero también aumenta el riesgo de sobreajuste.
- **Número de neuronas por capa (Hidden factor):** Este parámetro establece la cantidad de neuronas en cada capa oculta. Un mayor número de neuronas puede mejorar la capacidad del modelo para capturar patrones complejos, aunque también aumenta el tiempo de entrenamiento y la complejidad computacional.

6.3. Modelo neuronal de clasificación binaria

En esta sección se describe tanto el propósito como los procedimientos empleados para el desarrollo del modelo de clasificación binaria. Se detalla el proceso de selección de hiperparámetros, así como las distintas arquitecturas diseñadas con el fin de identificar la combinación que optimiza el rendimiento del modelo. Además, se presenta una comparación entre las arquitecturas evaluadas para determinar la más adecuada.

6.3.1. Propósito del modelo de clasificación binaria

El modelo de clasificación binaria se ha diseñado con el objetivo de identificar de manera precisa y eficiente, si una instancia de tráfico de red corresponde a una actividad legítima o a un comportamiento malicioso. Su función principal consiste en distinguir entre accesos normales y posibles intentos de intrusión, permitiendo así la detección temprana de ataques y contribuyendo a la protección proactiva de los sistemas informáticos.

6.3.2. Diseño del modelo e hiperparámetros seleccionados

En esta sección se detalla cual es el diseño del modelo de clasificación binaria escogido, así como los valores de los hiperparámetros utilizados durante los experimentos en el entrenamiento del modelo y sus correspondientes justificaciones.

Para diseñar la arquitectura del modelo, se ha desarrollado la clase `Modelo`. Esta clase está conformada por dos funciones, la primera corresponde con la inicialización de la clase, y la segunda con la arquitectura utilizada para este modelo.

Como se puede observar en la figura 6.4, el modelo está compuesto por dos capas. La `capa1` tiene un número de entradas igual al número de parámetros que poseen los datos utilizados para entrenar al modelo y un número de salidas que depende del parámetro `ventanaOculta`, con el que se llama a la función de inicialización del modelo. El número de salidas de la `capa1` será precisamente el que determine la arquitectura específica que se utiliza. Por su parte, la `capa2` recibe como entrada el número de salidas de la `capa1` y al tratarse de la capa final de un modelo de clasificación binaria, esta capa tendrá una única salida.

Debido a que la función de pérdida utilizada para entrenar el modelo es `BCEWithLogitsLoss()`, que ya aplica la función sigmoidea, sería contraproducente aplicar la función sigmoidea sobre la salida del modelo para su entrenamiento.

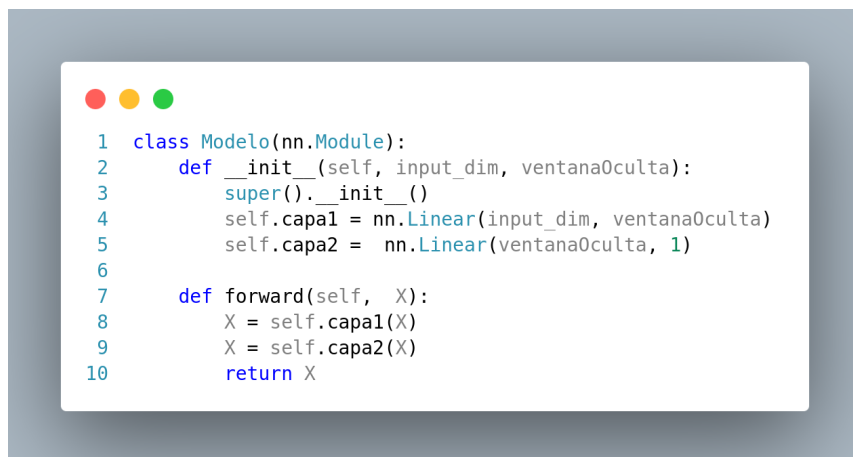
The image shows a code editor window with a light blue background. At the top left, there are three colored circles (red, yellow, green). The code is written in a dark blue font on a white background. It defines a class named 'Modelo' that inherits from 'nn.Module'. The class has two methods: '__init__' and 'forward'. The '__init__' method takes 'input_dim' and 'ventanaOculta' as arguments and initializes two layers: 'capa1' and 'capa2'. 'capa1' is a 'nn.Linear' layer with 'input_dim' inputs and 'ventanaOculta' outputs. 'capa2' is a 'nn.Linear' layer with 'ventanaOculta' inputs and 1 output. The 'forward' method takes 'X' as input, passes it through 'capa1' and then 'capa2', and returns the final output 'X'.

Figura 6.4: Estructura del modelo de clasificación binaria diseñada.

Para evitar el sobreajuste del modelo y obtener valores de las predicciones más confiables que se utilizarán para la evaluación del rendimiento del modelo, se utiliza validación cruzada *k-fold*. La validación cruzada consiste en dividir los datos de entrenamiento en *n* folds que se utilizan para entrenar al modelo con *n-1* folds y después validar el modelo entrenado con el fold que no se ha utilizado para entrenarlo. Este proceso se repite para los *n* folds y en cada iteración es un fold distinto el que no se utiliza para el entrenamiento. Una vez completadas las iteraciones se extraen las medias de las métricas obtenidas con cada fold durante la validación. Para entrenar el modelo de clasificación binaria, se ha utilizado `StratifiedKFold()`

que como su nombre indica, estratifica los folds para que las proporciones de las clases de los datos utilizados para el entrenamiento se mantengan similares en cada fold y no se creen folds con un único tipo de clase. Estratificar los datos es especialmente crítico para aquellos datasets cuyos datos están muy desbalanceados como es el caso de los datos utilizados para entrenar los modelos de este proyecto. Para el entrenamiento del modelo de clasificación binaria, se ha optado por utilizar validación cruzada StratifiedKFold con 5 folds o divisiones.

Como algoritmo de optimización para el entrenamiento del modelo de clasificación binaria, se ha optado por utilizar AdamW(). Como se comenta en la sección [6.1.3 Algoritmo de optimización](#), se ha demostrado que AdamW es más eficaz en la prevención de sobreajuste en redes neuronales que otros algoritmos de optimización. Además, permite una mejor generalización en problemas con grandes volúmenes de datos como el que se trata en este proyecto.

Los hiperparámetros que han sido seleccionados para encontrar la mejor configuración para el modelo de clasificación binaria diseñado son: el batch size, la tasa de aprendizaje (learning rate) y el número de épocas. Estos hiperparámetros resultan imprescindibles en el entrenamiento de modelos de clasificación binaria, dado que influyen directamente en la eficiencia del proceso de optimización y en la calidad del modelo resultante.

Como se comenta en la sección [6.2 Referenciassec:paramhiper](#), los hiperparámetros controlan el comportamiento del proceso de entrenamiento y afectan la capacidad del modelo para aprender patrones complejos de los datos. Teniendo en cuenta las características de cada hiperparámetro, se han escogido los siguientes valores para los experimentos del entrenamiento del modelo binario con el objetivo de encontrar la mejor combinación de ellos.

- **Batch size:** Teniendo en cuenta el tamaño de los datos que se utilizan para la fase de entrenamiento del modelo binario, se han considerado que los valores [2000, 10000, 15000, 20000], eran apropiados y suficientemente dispares como para notar diferencias significativas en los resultados de la ejecución de los experimentos.
- **Learning rate:** Siguiendo los avisos de la sección anterior en la que se explicaba el propósito y características de cada hiperparámetro, se ha optado por elegir unos valores con los que se espera que el modelo converja sin sobreajustarse ni se exceda en el tiempo de obtención de una solución válida. Los valores elegidos han sido [0.01, 0.001, 0.0001].
- **Épocas:** De forma similar a como sucede con la tasa de aprendizaje, un valor excesivamente bajo en el número de épocas puede provocar que el modelo no converja. Sin embargo, un número demasiado alto de épocas provoca sobreajuste en el modelo. Teniendo en cuenta las consideraciones mencionadas, los valores que se han probado para este hiperparámetro son [10, 20, 30].

6.3.3. Arquitecturas desarrolladas

En esta sección se abordan cuales son las arquitecturas a desarrollar para el modelo de clasificación binaria. Como se comenta en la sección anterior [6.3.2 Diseño del modelo e](#)

[hiperparámetros seleccionados](#), la diferencia entre las arquitecturas radica en el tamaño de la capa oculta del modelo.

Para el desarrollo de las diferentes arquitecturas, se ha optado por realizar experimentos de entrenamiento del modelo de clasificación binaria con un tamaño de ventana oculta igual a: la mitad del número de parámetros que recibe el modelo, el mismo número que parámetros que tiene el modelo y doble del número de parámetros. Esto equivale a un tamaño de la capa de venta oculta de 25, 49 y 98 respectivamente. Estos valores se obtienen de forma dinámica después de tratar los datos del csv utilizado para los entrenamientos y las pruebas de este proyecto.

A continuación, se comentan las posibles ventajas e inconvenientes de cada arquitectura que se esperan antes de realizar los experimentos.

Capa oculta con la mitad de neuronas que parámetros de entrada

En este apartado se comentan cuales son las principales ventajas e inconvenientes, a priori, de un modelo de clasificación binaria con un número de neuronas en la capa oculta igual a la mitad del número de parámetros que recibe el modelo.

Ventajas ($n/2$)

- El riesgo de sobreajuste es menor, esto es especialmente importante si no se utiliza validación cruzada o el conjunto de datos de entrenamiento es pequeño o simple.
- Al tener un número de neuronas menor, tanto la función de pérdida como el algoritmo de optimización, tardan menos en realizar los cálculos de los ajustes de los pesos y sesgos del modelo. Este menor coste de cálculo se refleja directamente en el tiempo necesario para entrenar el modelo y en el tiempo de respuesta del mismo una vez entrenado.
- Un número bajo de neuronas es especialmente útil cuando los datos son linealmente separables o poco complejos.

Inconvenientes($n/2$)

- Si las relaciones entre los datos son demasiado complejas, el moelo puede sufrir underfitting o subajuste.
- Sensibilidad alta a la inicialización de los pesos, lo que puede provocar que un porcentaje alto de neuronas no se utilicen por tener pesos iguales o muy cercanos a 0. Esto implica que parte de la capacidad de la red quede inutilizada, empeorando el rendimiento incluso en problemas simples.

Tras comentar cuales son las principales ventajas e inconvenientes de utilizar una capa oculta con un número de neuronas igual a la mitad de los parámetros del modelo, aparentemente, las propiedades del conjunto de datos que se utiliza para entrenar al modelo no son las idóneas para esta arquitectura y no se esperan resultados especialmente destacables.

Capa oculta con el mismo número de neuronas que parámetros de entrada

En este apartado se comentan cuales son las principales ventajas e inconvenientes, a priori, de un modelo de clasificación binaria con un número de neuronas en la capa oculta igual al número de parámetros que recibe el modelo.

Ventajas (n)

- La capacidad y la generalización están balanceadas, esto significa que el modelo evita tanto el sobreajuste como el subajuste.
- Es capaz de modelar relaciones más complejas entre conjuntos de datos que arquitecturas con un número de neuronas menor, lo que es especialmente útil en problemas con relaciones no lineales⁴ entre sus datos.
- Son computacionalmente más eficientes que arquitecturas con un número elevado de neuronas en la capa oculta.

Inconvenientes(n)

- Puede resultar insuficiente en problemas demasiado complejos o con un gran volumen de datos.
- Si el conjunto de datos utilizado para entrenar el modelo es pequeño puede sobreajustar.

Tras comentar cuales son las principales ventajas e inconvenientes de utilizar una capa oculta con el mismo número de neuronas que parámetros recibe el modelo, aparentemente, esta arquitectura ofreciera resultados significativamente positivos gracias a su equilibrio y balanceo.

Capa oculta con el doble de neuronas que parámetros de entrada

En este apartado se comentan cuales son las principales ventajas e inconvenientes, a priori, de un modelo de clasificación binaria con el doble de neuronas en la capa oculta que parámetros recibe el modelo.

Ventajas (n*2)

- Alta capacidad de aprendizaje, lo que se traduce como la posibilidad de capturar patrones de datos muy complejos que no son lineales ni perceptibles en un análisis superficial del problema.
- Cuanto mayor es el conjunto de datos mejor se ajustan los pesos.
- Es menos propenso a subajustar que otras arquitecturas.

Inconvenientes(n*2)

- El riesgo de sobreajuste es elevado, memorizando de esta manera ruido en vez de aprender los patrones generalizables.
- Nivel computacional muy elevado, lo que se refleja en el tiempo necesario de entrenamiento de esta arquitectura y en su posterior tiempo de respuesta.
- Si el número de muestras del conjunto de datos es bajo tiende a sobreajustar.

Tras comentar cuales son las principales ventajas e inconvenientes de utilizar una capa oculta con un número de neuronas igual al doble de los parámetros del modelo, aparentemente, esta arquitectura obtendrá mejores resultados si los parámetros no tiene relaciones lineales. En cambio, si los datos presentan relaciones lineales tiene más probabilidades de sobreajustar los pesos del modelo, lo que implicaría que el modelo no ha sido capaz de converger a una solución genérica.

6.4. CREAR E INSERTAR IMAGENES DE LAS ARQUITECTURAS

6.4.1. Matriz de confusión para clasificación binaria

En esta sección se explica en que consiste una matriz de confusión y las variaciones que existen de las matrices de confusión en función del tipo de modelo neuronal de clasificación que se entrene. Son una herramienta fundamental utilizada en el campo del aprendizaje automático y la clasificación, especialmente cuando se evalúan modelos de clasificación como el que se propone en este proyecto.

La matriz de confusión es una representación tabular que permite evaluar el rendimiento de un modelo de clasificación. Esta matriz compara las predicciones del modelo con los valores reales (verdaderos) y proporciona una visión detallada sobre los errores cometidos. Esta matriz permite calcular diversas métricas de evaluación del modelo, que son esenciales para entender la efectividad del modelo en tareas de clasificación.

En el caso de los modelos neuronales de clasificación binaria, la matriz tiene una estructura 2x2, donde cada celda en la matriz representa la cantidad de veces que una combinación específica de clase real y clase predicha ocurrió. Los valores posibles para la clasificación binaria son:

- Verdaderos positivos (VP): son las instancias que pertenecen a la clase positiva y que el modelo ha clasificado correctamente como positivas.
- Falsos positivos (FP): corresponden a las instancias que no pertenecen a la clase positiva, pero que el modelo ha etiquetado incorrectamente como positivas.
- Falsos negativos (FN): se refieren a las instancias que deberían ser clasificadas como positivas, pero que el modelo ha predicho como negativas.
- Verdaderos negativos (VN): son las instancias que pertenecen a la clase negativa y que el modelo ha clasificado correctamente como negativas.

	Predicción Positiva	Predicción Negativa
Real Positivo	Verdaderos Positivos (VP)	Falsos Negativos (FN)
Real Negativo	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Tabla 6.1: Matriz de confusión para clasificación binaria.

6.4.2. Métricas útiles en clasificación binaria

En esta sección se define el concepto de métrica, se explica cual es su importancia en el proceso de obtención de un modelo neuronal bien entrenado y se enumeran las métricas que se utilizan para el estudio del modelo de clasificación binaria desarrollado en este proyecto explicando la importancia de cada una de ellas.

Las métricas en el contexto de los modelos de aprendizaje automático son herramientas cuantitativas utilizadas para evaluar el rendimiento de un modelo entrenado. Estas métricas permiten medir si el modelo está realizando la tarea para la cual fue diseñado de manera correcta, comparando sus predicciones con los valores reales de los datos de prueba. Son fundamentales para comprender la efectividad del modelo y para identificar áreas de mejora, ya que proporcionan una evaluación objetiva y reproducible del comportamiento del modelo.

El uso adecuado de las métricas es crucial, ya que guían las decisiones sobre el ajuste y la optimización del modelo. Sin métricas claras, el proceso de evaluación se vuelve subjetivo y difícil de manejar. Además, las métricas permiten realizar comparaciones entre diferentes modelos o configuraciones, facilitando la selección del modelo más adecuado para una tarea específica. Una interpretación correcta de las métricas ayuda a evitar problemas como el sobreajuste o el subajuste, y garantiza que el modelo generalice bien a nuevos datos [27].

Para evaluar el desempeño del modelo de detección y clasificación de ataques, se utilizan las siguientes métricas derivadas de la matriz de confusión.

■ **Exactitud (*Accuracy*):**

$$\text{Accuracy} = \frac{VP + VN}{VP + FP + VN + FN} \quad (6.1)$$

En el entrenamiento de modelos neuronales para la detección de intrusiones, esta métrica representa la proporción del total de las clasificaciones realizadas correctamente. Indica la capacidad general del modelo para distinguir entre tráfico normal e intrusivo. Si bien ofrece una visión global del rendimiento, su valor disminuye en escenarios donde la cantidad de tráfico normal supera significativamente al tráfico malicioso, ya que el modelo puede obtener una alta exactitud simplemente clasificando la mayoría de las instancias como normales.

■ **Precisión (*Precision*):**

$$\text{Precision} = \frac{VP}{VP + FP} \quad (6.2)$$

Esta métrica evalúa la capacidad del modelo neuronal para evitar la identificación incorrecta de tráfico normal como intrusivo. En la detección de intrusiones, una alta precisión es crucial para minimizar las falsas alarmas, las cuales pueden generar una sobrecarga operativa en los equipos de seguridad, requiriendo la revisión de eventos benignos y distrayendo la atención de amenazas reales. Un modelo preciso reduce la fatiga de alertas y permite una respuesta más eficiente a incidentes genuinos.

■ **Sensibilidad (*Recall*):**

$$\text{Recall} = \frac{VP}{VP + FN} \quad (6.3)$$

La sensibilidad mide la habilidad del modelo neuronal para detectar todas las instancias de intrusión presentes en el tráfico de red. En el contexto de la seguridad, un alto recall es de suma importancia, ya que implica una menor probabilidad de que ataques reales pasen desapercibidos. Un modelo con baja sensibilidad puede tener consecuencias graves, permitiendo que actividades maliciosas se infiltren y comprometan la integridad y la confidencialidad de los sistemas.

■ **Puntuación F1 (*F1-Score*):**

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6.4)$$

Esta métrica proporciona una evaluación equilibrada del rendimiento del modelo neuronal al calcular la media armónica entre la precisión y el recall. En la detección de intrusiones, donde a menudo existe un desequilibrio entre el tráfico normal y el malicioso, el F1-score ofrece una métrica más robusta que la exactitud, ya que considera tanto la capacidad de evitar falsas alarmas como la de detectar todas las intrusiones. Un valor alto de F1-score indica un buen compromiso entre ambas capacidades.

■ **Puntuación F2 (*F2-Score*):**

$$F2 = 5 \times \frac{\text{Precision} \times \text{Recall}}{4 \times \text{Precision} + \text{Recall}} \quad (6.5)$$

Esta variante de la puntuación F pondera la sensibilidad más que la precisión. En el ámbito de la detección de intrusiones, el F2-score resulta útil cuando las consecuencias de no detectar un ataque (falso negativo) se consideran significativamente más perjudiciales que generar una falsa alarma (falso positivo). Al asignar un mayor peso al recall, se prioriza la identificación de la mayor cantidad posible de actividades maliciosas, incluso a expensas de un posible aumento en las falsas alertas.

■ **Área bajo la curva (*ROC AUC*):**

$$\text{FPR} = \frac{FP}{FP + VN} \quad (6.6)$$

$$\text{Recall}(\text{FPR}(t)) = \text{Recall}(t) \quad (6.7)$$

$$\text{ROC AUC} = \int_0^1 \text{Recall}(\text{FPR}) d(\text{FPR}) \quad (6.8)$$

El área bajo la curva ROC (Receiver Operating Characteristic) mide la capacidad del modelo neuronal para discriminar entre clases positivas y negativas a lo largo de todos los posibles umbrales de decisión. En la curva ROC, la FPR (False Positive Rate) actúa como la variable del eje x, mientras que el Recall (o TPR, True Positive Rate) corresponde a la variable del eje y. En el contexto de la seguridad, un alto valor de AUC indica que el modelo asigna puntuaciones más altas a las intrusiones que al tráfico normal de manera consistente, lo que sugiere una buena capacidad de priorización de amenazas. Esta métrica es especialmente útil cuando se desea evaluar el rendimiento global del modelo sin depender de un umbral fijo. Un modelo con bajo AUC podría fallar al separar el tráfico malicioso, lo que aumentaría la probabilidad de errores en la clasificación y podría comprometer la eficacia general del sistema de detección.

REPASAR, NO TERMINA DE CONVENCERME Aplicación en Seguridad

En el contexto de detección de intrusiones:

- Un recall alto (¿95 %) asegura que pocos ataques pasan desapercibidos.
- La precisión debe optimizarse para reducir la carga operativa de analistas (falsos positivos ¿10 %).
- El F2-Score es preferible al F1 cuando la prioridad es minimizar riesgos de ataques no detectados.

6.4.3. Mejores hiperparámetros para cada arquitectura

En esta sección se pueden observar los resultados de los experimentos realizados con los posibles valores de los hiperparámetros comentados en la sección [6.3.2 Diseño del modelo e hiperparámetros seleccionados](#).

Para medir la eficiencia y eficacia del modelo, se han obtenido las métricas derivadas de la matriz de confusión comentadas en la sección [6.4.2 Métricas útiles en clasificación binaria](#). El objetivo del modelo es detectar intrusiones en redes informáticas, por este motivo, los falsos negativos, es decir cuando se produce un ataque o intrusión y no se detecta, pueden tener consecuencias muy catastróficas para el sistema. Teniendo esto en cuenta, la métrica más importante es recall, puesto que es la que se ve más afectada por la presencia de falsos negativos.

Otras métricas relevantes que se muestran en las siguientes figuras son:

- **F1 score:** Equilibra entre precisión y recall. Ideal cuando se necesita rendimiento general en un contexto con mucho desbalanceo como el que se propone en este proyecto.
- **Precisión:** Si es bajo implica que el modelo detectará muchos falsos positivos, lo que saturaría a los administradores de falsas alarmas y provocaría deconfianza en el modelo.
- **ROC AUC:** Compara el rendimiento global del modelo, independientemente del umbral. Es especialmente útil para afinar la sensibilidad o recall.

En las siguientes figuras se hace referencia a los posibles combinaciones de la matriz de confusión en inglés, es decir:

- **tp:** Verdaderos positivos (VP), intrusiones que se han clasificado como intrusiones.
- **tn:** Verdaderos negativos (VN), conexiones legítimas identificadas como tal.
- **fp:** Falsos positivos (FP), conexiones legítimas clasificadas como intrusiones.
- **fn:** Falsos negativos (FN), intrusiones que se han clasificado como conexiones legítimas.

En la figura [6.5](#), se encuentran representadas las cinco mejores configuraciones de hiperparámetros para el modelo de clasificación binaria con una arquitectura de 25 neuronas ($n/2$) en la capa oculta. Los resultados obtenidos superan lo esperado tras el análisis de la sección [6.3.3](#) para esta arquitectura. La mejor configuración para la arquitectura con la mitad de neuronas en la capa oculta que de parámetros de entrada tiene el modelo, ha sido:

- **Batch size:** 15 000
- **Epochs:** 10
- **Learning rate:** 0,01

6.4. CREAR E INSERTAR IMAGENES DE LAS ARQUITECTURAS

batch_size	epochs	learning_rate	avg_f1	avg_fn	avg_fp	avg_precision	avg_recall	avg_roc_auc	avg_tn	avg_tp
15000	10	0.01	0.9979136687	20.4	41	0.9972157482	0.9986126862	0.9997672909	344123.2	14684.2
10000	30	0.001	0.9977168211	21.8	45.4	0.9969175197	0.9985174741	0.9997787627	344118.8	14682.8
20000	10	0.01	0.9977913769	22	43	0.9970799519	0.9985038742	0.9997532199	344121.2	14682.6
20000	20	0.01	0.9980218885	22.6	35.6	0.9975812368	0.9984630745	0.9997762523	344128.6	14682
2000	10	0.001	0.9977911968	23	42	0.9971474464	0.9984358636	0.9997862838	344122.2	14681.6

Figura 6.5: Mejores cinco configuraciones de hiperparámetros del modelo de clasificación binaria con una capa oculta de 25 neuronas.

La figura 6.6, muestra la representación de las cinco mejores configuraciones de hiperparámetros para el modelo de clasificación binaria con una arquitectura de 49 neuronas (n) en la capa oculta. Los resultados medios obtenidos concuerdan con lo comentado en el análisis de la sección 6.3.3 para esta arquitectura. La mejor configuración para la arquitectura con la el mismo número de neuronas en la capa oculta que de parámetros de entrada tiene el modelo, ha sido:

- **Batch size:** 20 000
- **Epochs:** 10
- **Learning rate:** 0,01

batch_size	epochs	learning_rate	avg_f1	avg_fn	avg_fp	avg_precision	avg_recall	avg_roc_auc	avg_tn	avg_tp
20000	10	0.01	0.9979002139	19.6	42.2	0.9971345937	0.9986670886	0.9997654591	344122	14685
10000	30	0.001	0.9979069887	19.8	41.8	0.9971616627	0.9986534868	0.9997799135	344122.4	14684.8
15000	10	0.01	0.9980356649	21	36.8	0.9975001039	0.9985718792	0.9997817739	344127.4	14683.6
2000	10	0.001	0.9980220715	21.2	37	0.9974864885	0.9985582793	0.9997690767	344127.2	14683.4
2000	30	0.0001	0.9976488198	23.2	46	0.9968766609	0.9984222683	0.9998293302	344118.2	14681.4

Figura 6.6: Mejores cinco configuraciones de hiperparámetros del modelo de clasificación binaria con una capa oculta de 49 neuronas.

En la figura 6.7, se encuentran representadas las cinco mejores configuraciones de hiperparámetros para el modelo de clasificación binaria con una arquitectura de 98 neuronas (n*2) en la capa oculta. Los resultados obtenidos coinciden con lo esperado tras el análisis de la sección 6.3.3 para esta arquitectura. La mejor configuración para la arquitectura con el doble de neuronas en la capa oculta que de parámetros de entrada tiene el modelo, ha sido:

- **Batch size:** 20 000
- **Epochs:** 10
- **Learning rate:** 0,01

batch_size	epochs	learning_rate	avg_f1	avg_fn	avg_fp	avg_precision	avg_recall	avg_roc_auc	avg_tn	avg_tp
20000	10	0.01	0.9981242141	18.4	36.8	0.9975006059	0.9987486972	0.9997812469	344127.4	14686.2
10000	30	0.001	0.9979406463	21.6	39	0.9973509999	0.9985310776	0.9997768743	344125.2	14683
15000	30	0.001	0.9977710383	22	43.6	0.9970393055	0.9985038723	0.9997758709	344120.6	14682.6
15000	10	0.01	0.9980151456	22.4	36	0.9975542032	0.9984766744	0.9997726471	344128.2	14682.2
10000	20	0.001	0.9977302451	22.6	44.2	0.9969985793	0.9984630726	0.9997767639	344120	14682

Figura 6.7: Mejores cinco configuraciones de hiperparámetros del modelo de clasificación binaria con una capa oculta de 98 neuronas.

Estos resultados no son definitivos, puesto que no se está evaluando la posibilidad de que el modelo haya sufrido sobreajuste incluso utilizando validación cruzada. Para obtener resultados reales del desempeño del modelo, se realizan pruebas, también conocidas como test, en las que se obtienen las métricas del modelo al recibir datos que no ha visto nunca. Este proceso y sus resultados se comentan en el capítulo 7.

6.4.4. Comparación de las arquitecturas

En esta sección se comparan los resultados obtenidos entre las tres arquitecturas propuestas. Una vez comparadas, se desarrolla una posible explicación para comprender los resultados de los experimentos.

Para comprender mejor los resultados globales de los experimentos, se han recogido en la figura 6.8, las 5 configuraciones de hiperparámetros con mayor recall independientemente de su arquitectura.

batch_size	epochs	hidden_size	learning_rate	avg_recall
20000	10	98	0.01	0.9987486972
20000	10	49	0.01	0.9986670886
10000	30	49	0.001	0.9986534868
15000	10	25	0.01	0.9986126862
15000	10	49	0.01	0.9985718792

Figura 6.8: Mejores cinco configuraciones de hiperparámetros del modelo de clasificación binaria.

El problema para el que ha sido diseñado el modelo no presenta una relación lineal tras transformar los datos como se explica en el capítulo 5 [Entendimiento de los datos](#). Por este motivo, se esperaba que el modelo con un tamaño de capa oculta mayor fuese capaz de converger mejor, tal y como se comenta en 6.3.3.

Al observar los datos de la figura 6.8, se puede apreciar que puntualmente, la mejor

arquitectura ha sido la que tenía un tamaño de ventana oculta igual a 98. Cabe destacar que, la diferencia entre los valores de recall de los cinco mejores experimentos es muy pequeña y puede que se deba a factores externos, como el estado de la máquina en el momento en el que se entreno al modelo, o la inicialización aleatoria de los pesos. Por este motivo, los resultados pueden variar si se realizan los mismos experimentos.

Ente los mejores resultados de los experimentos, destaca la presencia de la arquitectura con un tamaño de capa oculta igual al número de parámetros de entrada del modelo. Tal y como se comenta en las secciones anteriores, esta es la arquitectura más equilibrada. Gracias a este equilibrio ha sido capaz de ajustar sus pesos mejor que el resto de arquitecturas para configuraciones de hiperparámetros diferentes.

Al aumentar el número de épocas del entrenamiento de los modelos, estos tienden a sobreajustar los pesos. Este sobreajuste implica que el modelo obtiene unas métricas excepcionalmente positivas durante los experimentos. Sin embargo, aunque los valores de las métricas son muy positivos en los experimentos realizados, sorprende que el número de épocas del 80 % de las 5 mejores configuraciones, es el menor de los valores con los que se han realizado los experimentos para el hiperparámetro epochs. Estos resultados reflejan que la probabilidad de que el modelo haya sobreajustado sus pesos es muy baja. Posiblemente, el uso de la validación cruzada que penaliza el sobreajuste de los pesos de los modelos, como se explica en secciones anteriores, sea el responsable de que los experimentos con un número de épocas más elevado hayan obtenido unos resultados peores.

Finalmente, es destacable que el learning rate de la mayoría de los 5 mejores experimentos del global de las arquitecturas tenga el valor más alto de los que se han probado. Esto puede interpretarse de manera positiva o negativa. En el mejor de los casos, un learning rate alto muestra que los datos están bien estructurados y que el optimizador y las arquitecturas son robustos frente a valores altos de learning rate. En el peor de los casos, la convergencia se realiza rápidamente pero de manera inestable, los resultados dependen en gran medida de los valores iniciales de los pesos o se este sobreajustando el modelo. Este último caso es el más improbable debido a las técnicas utilizadas que se han descrito y a los resultados analizados en el párrafo anterior.

Para conocer el verdadero desempeño del modelo y determinar si ha sufrido sobreajuste de los pesos durante el entrenamiento, se realiza un test con datos que no ha visto el modelo durante el entrenamiento. Estas pruebas y análisis se describen en el capítulo [7 Test](#). Si los resultados obtenidos de estos test son similares a los resultados obtenidos de los experimentos, significa que el modelo converge correctamente hacia una solución real. En caso de que los resultados obtenidos de los test sean significativamente peores que los obtenidos durante los experimentos, el modelo habrá sobreajustado sus pesos.

6.5. Modelo neuronal de clasificación multiclase

Esta sección aborda los objetivos y procedimientos aplicados en el desarrollo del modelo de clasificación multiclase. Se describe el proceso llevado a cabo para la selección de hiperparámetros, junto con las diversas arquitecturas propuestas, con el propósito de identificar la

configuración que maximiza el desempeño del modelo. Asimismo, se realiza una comparación exhaustiva entre las arquitecturas implementadas para determinar la opción más eficiente.

Por último, se detallan las métricas empleadas para la evaluación y la elección de las combinaciones óptimas de hiperparámetros, de forma análoga a como se hace en la sección [6.3 Modelo neuronal de clasificación binaria](#).

6.5.1. Proposito del modelo de clasificación multiclase

El modelo de clasificación multiclase se ha diseñado con el objetivo de identificar de manera precisa y eficiente, a que tipo de ataque corresponde una conexión maliciosa que ya se ha clasificado previamente como intrusiva. Su función principal consiste en distinguir entre 9 clases diferentes de ataques que se detallan en la sección [5.2 Tipos de ataques registrados en los datos](#), contribuyendo a la protección proactiva de los sistemas informáticos.

6.5.2. Diseño del modelo e hiperparámetros seleccionados

En esta sección se detalla cual es el diseño del modelo de clasificación multiclase escogido, así como los valores de los hiperparámetros utilizados durante los experimentos en el entrenamiento del modelo y sus correspondientes justificaciones.

Para diseñar la arquitectura del modelo, se ha desarrollado la clase `ModeloMulticlase`. Esta clase está conformada por dos funciones, la primera corresponde con la inicialización de la clase, y la segunda con la arquitectura utilizada para este modelo.

Como se puede observar en la figura [6.9](#), el modelo está compuesto por dos capas. La capa1 tiene un número de entradas igual al número de parámetros que poseen los datos utilizados para entrenar al modelo y un número de salidas que depende del parámetro *ventaOculta*, con el que se llama a la función de inicialización del modelo. El número de salidas de la capa1 será precisamente el que determine la arquitectura específica que se utiliza. Una vez que los datos pasan por la capa1, se utiliza la técnica Batch Normalization.

Batch Normalization es una técnica que normaliza las activaciones de una capa para que tengan media cercana a 0 y varianza cercana a 1, calculadas a partir del batch actual del entrenamiento del modelo. Aplicar BatchNorm tras la primera capa en un modelo de clasificación multiclase ayuda a estabilizar, acelerar y regularizar el entrenamiento, mejorando la capacidad del modelo para aprender buenas representaciones desde las primeras fases del entrenamiento.

Las redes neuronales necesitan funciones no lineales para aprender relaciones complejas entre las entradas y salidas, por este motivo, se aplica la función de activación Rectified Linear Unit (ReLU) después del batch normalization. Sin activaciones como ReLU, toda la red sería equivalente a una simple transformación lineal, sin capacidad de aprendizaje profundo. Que el modelo presente una alta capacidad de aprendizaje profundo es esencial para los modelos de clasificación multiclase cuya complejidad es mucho mayor que la de los modelos de clasificación binaria.

Para finalizar, la capa2 recibe como entrada el número de salidas de la capa1 y al tratarse de la capa final de un modelo de clasificación multiclase, esta capa tendrá una salida por cada clase que es capaz de identificar el modelo.

Por lo general, en clasificación multiclase, la salida del modelo suele ser un vector de logits (valores sin escalar), uno por cada clase. La función de activación Softmax transforma esos logits en probabilidades, es decir, cada valor representa lo probable que es que los datos que han entrado al modelo, pertenezca a esa clase. Debido a que durante el entrenamiento, se utiliza el frameworks PyTorch, no es necesario aplicar softmax a la salida del modelo cuando se utiliza la función de pérdida CrossEntropyLoss. La razón es que función CrossEntropyLoss, ya incluye softmax internamente por razones de eficiencia y estabilidad numérica.



```
1 class ModeloMulticlase(nn.Module):
2     def __init__(self, input_dim, ventanaOculta, numClases):
3         super().__init__()
4         self.capa1 = nn.Linear(input_dim, ventanaOculta)
5         self.bn1 = nn.BatchNorm1d(ventanaOculta, momentum=0.01)
6         self.capa2 = nn.Linear(ventanaOculta, numClases)
7
8     def forward(self, X):
9         X = torch.relu(self.bn1(self.capa1(X)))
10        X = self.capa2(X)
11        return X
```

Figura 6.9: Estructura del modelo de clasificación multiclase diseñada.

Para evitar el sobreajuste del modelo y obtener valores de las predicciones más confiables que se utilizarán para la evaluación del rendimiento del modelo, se utiliza validación cruzada k-fold. Como se explica en la sección anterior 6.3.2, la validación cruzada consiste en dividir los datos de entrenamiento en n folds que se utilizan para entrenar al modelo con $n-1$ folds y después validar el modelo entrenado con el fold que no se ha utilizado para entrenarlo. Este proceso se repite para los n folds y en cada iteración es un fold distinto el que no se utiliza para el entrenamiento. Una vez completadas las iteraciones se extraen las medias de las métricas obtenidas con cada fold durante la validación. Para entrenar el modelo de clasificación multiclase, se ha utilizado StratifiedKFold() que como su nombre indica, estratifica los folds para que las proporciones de las clases de los datos utilizados para el entrenamiento se mantengan similares en cada fold y no se creen folds con un único tipo de clase. Estratificar los datos es especialmente crítico para aquellos datasets cuyos datos están muy desbalanceados como es el caso de los datos utilizados para entrenar los modelos de este proyecto. Concretamente, tras el tratamiento de los datos, existen 284 veces más muestras de la clase con mayor presencia en el dataset, que de la clase con menor presencias. Para el entrenamiento del modelo de clasificación multiclase, se ha optado por utilizar validación cruzada StratifiedKFold con 5 folds o divisiones.

Como algoritmo de optimización para el entrenamiento del modelo de clasificación multi-clase, se ha optado por utilizar AdamW(). Como se comenta en la sección [6.1.3 Algoritmo de optimización](#), se ha demostrado que AdamW es más eficaz en la prevención de sobreajuste en redes neuronales que otros algoritmos de optimización.

Los hiperparámetros que han sido seleccionados para encontrar la mejor configuración para el modelo de clasificación multiclase diseñado son: el batch size, la tasa de aprendizaje (learning rate) y el número de épocas. Estos hiperparámetros resultan imprescindibles en el entrenamiento de modelos de clasificación multiclase, dado que influyen directamente en la eficiencia del proceso de optimización y en la calidad del modelo resultante.

Como se comenta en la sección [6.2 Referencias](#), los hiperparámetros controlan el comportamiento del proceso de entrenamiento y afectan la capacidad del modelo para aprender patrones complejos de los datos. Teniendo en cuenta las características de cada hiperparámetro, se han escogido los siguientes valores para los experimentos del entrenamiento del modelo multiclase con el objetivo de encontrar la mejor combinación de ellos.

- **Batch size:** Teniendo en cuenta el tamaño de los datos que se utilizan para la fase de entrenamiento del modelo de clasificación multiclase, se han considerado que los valores [32, 64, 128, 256, 512], eran apropiados y suficientemente dispares como para notar diferencias significativas en los resultados de la ejecución de los experimentos.
- **Learning rate:** Siguiendo los avisos de la sección anterior en la que se explicaba el propósito y características de cada hiperparámetro, se ha optado por elegir unos valores con los que se espera que el modelo converja sin sobreajustarse ni se exceda en el tiempo de obtención de una solución válida. Los valores elegidos han sido [0.01, 0.001, 0.0001, 0.00001].
- **Épocas:** De forma similar a como sucede con la tasa de aprendizaje, un valor excesivamente bajo en el número de épocas puede provocar que el modelo no converja. Sin embargo, un número demasiado alto de épocas provoca sobreajuste en el modelo. Teniendo en cuenta las consideraciones mencionadas, los valores que se han probado para este hiperparámetro son [30, 50, 80, 100].

La convergencia de los modelos de clasificación multiclase es mucho más complicada que en los modelos de clasificación binaria. Esto se debe a que en los modelos de clasificación binaria hay 4 posibles respuestas (VP, VN, FP, FN), mientras que en el modelo de clasificación multiclase que se propone en este proyecto hay 9 clases por 9 posibilidades de respuesta por clase, lo que suma un total de 81 posibles respuestas. Para encontrar una configuración de los hiperparámetros que se ajuste mejor a la complejidad de este problema, se han probado más combinaciones de valores de los hiperparámetros en el modelo de clasificación multiclase que en el modelo de clasificación binaria. El aumento del número de experimentos se ha realizado teniendo en cuenta que el número de muestras de conexiones maliciosas es mucho menor que el número total de conexiones que se ha utilizado para entrenar el modelo de clasificación binaria. Esto se traduce en que el tiempo de ejecución de los experimentos de ambos modelos ha sido similar compensando el número de muestras con el número de experimentos.

6.5.3. Diseños del modelo de clasificación multiclase descartados

Debido a la complejidad del problema al resolver, se probaron otros diseños que finalmente fueron descartados por no obtener los resultados esperados. A continuación, se detalla cuales fueron estos diseños descartados, la motivación que llevo a probarlo y cuales fueron sus resultados.

Durante las pruebas realizadas antes de ejecutar los experimentos para encontrar la combinación de hiperparámetros que mejor convergían para este modelo, se detectó que el desbalanceo de las clases afectaba de manera significativa a los resultados que se obtenían. Con el objetivo de reducir el impacto del desbalanceo en el entrenamiento del modelo, se optó por probar diferentes combinaciones que se muestran en la figura 6.10, junto con los resultados que obtuvieron estas combinaciones de funciones en la métrica F1 weighted.

Notes	batch_size	epochs	hidden_size	learning_rate	avg_f1_weighted
Capa bn1 e hiperparámetros con valores altos	256	80	512	0.01	0.5444417415
Con capa bn1	32	30	128	0.00001	0.4673102794
Con capa bn1 y dropout	32	30	128	0.00001	0.4614142168
Con SMOTE y capa bn1	32	30	128	0.00001	0.3694149376
Con SMOTE	32	30	128	0.00001	0.2828888103

Figura 6.10: Resultados de los diferentes diseños probados para el modelo de clasificación multiclase.

SMOTE (Synthetic Minority Over-sampling Technique) es una técnica de sobremuestreo utilizada para abordar el problema del desbalance de clases en conjuntos de datos de clasificación. Su funcionamiento se basa en generar ejemplos sintéticos de las clases minoritarias, en lugar de replicar instancias existentes. En tareas de clasificación multiclase, SMOTE puede aplicarse a cada clase minoritaria de forma independiente, generando ejemplos sintéticos para aquellas clases con menor representación. La aplicación de esta técnica en modelos de clasificación multiclase muy desbalanceados:

- Contribuye a reducir el sesgo del modelo hacia las clases mayoritarias.
- Mejora el rendimiento en métricas como recall y F1-score macro, al equilibrar el conjunto de entrenamiento.
- Favorece una mejor generalización en la predicción de clases minoritarias, al ampliar su representación en el espacio de características.

Sin embargo, tras realizar una serie de pruebas, el modelo diseñado que implementaba SMOTE fue el que peores resultados dió. Por eso motivo, se combinó con la función Batch Normalization explicada en la sección 6.5.2 [Diseño del modelo e hiperparámetros seleccionados](#).

Los resultados de Batch Normalization (bn1) y SMOTE fueron algo mejores que los valores obtenidos de las pruebas realizadas solo con SMOTE. Como se noto una mejora en

el desempeño del modelo al utilizar bn1, pero los resultados obtenidos seguían sin ser los esperados, se probó con combinaciones de diseño utilizando bn1 con otras técnicas.

Dropout es una técnica de regularización utilizada en redes neuronales para reducir el sobreajuste durante el entrenamiento. La función `Dropout()`, implementada en bibliotecas como TensorFlow y PyTorch, actúa introduciendo aleatoriamente una probabilidad de desactivación sobre las neuronas de una capa determinada. Durante cada iteración de entrenamiento, un subconjunto aleatorio de neuronas es temporalmente desactivado (es decir, sus salidas se anulan) con una probabilidad definida, denominada tasa de dropout (dropout rate), sin afectar la arquitectura del modelo ni sus pesos de forma permanente. Los principales beneficios de utilizar dropout en un modelo de clasificación multiclase con unos datos de entrenamiento muy desbalanceados son:

- Reduce el sobreajuste hacia las clases mayoritarias, evitando que el modelo memorice solo los patrones frecuentes y mejorando su comportamiento general.
- Fomenta la generalización en clases minoritarias, al impedir dependencias excesivas entre neuronas y promover representaciones más robustas y diversas.
- Mejora métricas sensibles al desbalance como el weighted F1-score o macro recall, al favorecer que el modelo aprenda también de clases con baja representación.

Con la implementación del dropout, los resultados mejoraron significativamente. Sin embargo, el diseño del modelo que solo utilizaba bn1 obtuvo mejores resultados de media que aquel que utilizaba bn1 y dropout. Por este motivo, el diseño utilizado en los experimentos solo implementa Batch Normalization. Además, tras realizar pruebas con valores altos para los hiperparámetros, se obtuvieron unos resultados con una convergencia mayor que la de las pruebas realizadas con valores menores, lo que incentivó a aumentar los valores de los hiperparámetros probados, tal y como se puede observar en la sección anterior.

6.5.4. Arquitecturas desarrolladas

En esta sección se abordan cuales son las arquitecturas a desarrollar para el modelo de clasificación multiclase. Como se comenta en la sección anterior [6.5.2 Diseño del modelo e hiperparámetros seleccionados](#), la diferencia entre las arquitecturas radica en el tamaño de la capa oculta del modelo.

Para el desarrollo de las diferentes arquitecturas, se ha optado por realizar experimentos de entrenamiento del modelo de clasificación multiclase con un tamaño de ventana oculta igual a: la mitad del número de parámetros que recibe el modelo, el mismo número que parámetros que tiene el modelo y doble del número de parámetros. Esto equivale a un tamaño de la capa de venta oculta de 25, 49 y 98 respectivamente, de forma análoga a como sucede con las arquitecturas del modelod de clasificación binaria. Estos valores se obtienen de forma dinámica después de tratar los datos del csv utilizado para los entrenamientos y las pruebas de este proyecto.

A continuación, se comentan las posibles ventajas e inconvenientes de cada arquitectura que se esperan antes de realizar los experimentos.

Capa oculta con la mitad de neuronas que parámetros de entrada

En este apartado se comentan cuales son las principales ventajas e inconvenientes, a priori, de un modelo de clasificación multiclase con un número de neuronas en la capa oculta igual a la mitad del número de parámetros que recibe el modelo.

Ventajas ($n/2$)

- Un número reducido de neuronas en la capa oculta limita la capacidad del modelo para memorizar patrones dominantes, lo cual puede ayudar a mitigar el sobreajuste hacia las clases mayoritarias del conjunto de entrenamiento.
- Al reducirse la complejidad del modelo, se disminuye el tiempo necesario para calcular los gradientes durante el entrenamiento y, en consecuencia, también el tiempo de respuesta durante la inferencia.
- En presencia de clases desbalanceadas, un modelo menos expresivo puede favorecer una mayor estabilidad en el aprendizaje, evitando que la red se especialice únicamente en las clases más frecuentes.

Inconvenientes($n/2$)

- Un número bajo de neuronas puede limitar la capacidad del modelo para aprender representaciones suficientes de las clases minoritarias, especialmente si estas presentan patrones complejos o poco diferenciados.
- La capacidad reducida del modelo puede dificultar la separación entre clases en problemas multiclase, afectando negativamente al rendimiento en métricas sensibles al desbalance como el macro recall o el macro F1-score.
- Si el número de parámetros de entrada es muy alto, una capa oculta con solo la mitad de neuronas podría actuar como un cuello de botella, perdiéndose información útil para distinguir correctamente entre clases poco representadas.

Tras comentar cuales son las principales ventajas e inconvenientes de utilizar una capa oculta con un número de neuronas igual a la mitad de los parámetros del modelo, aparentemente, las propiedades del conjunto de datos que se utiliza para entrenar al modelo no son las idóneas para esta arquitectura y no se esperan resultados especialmente destacables. Debido al alto desbalanceo entre clases, se espera que esta arquitectura proporcione resultados peores que su equivalente del modelo de clasificación binaria.

Capa oculta con el mismo número de neuronas que parámetros de entrada

En este apartado se comentan cuales son las principales ventajas e inconvenientes, a priori, de un modelo de clasificación multiclase con un número de neuronas en la capa oculta igual al número de parámetros que recibe el modelo.

Ventajas (n)

- Un número de neuronas igual al de parámetros de entrada proporciona al modelo una capacidad representativa suficiente para capturar patrones relevantes tanto en clases frecuentes como en las minoritarias.
- Esta configuración ofrece un equilibrio razonable entre complejidad y eficiencia computacional, manteniendo un coste de entrenamiento y de inferencia moderado sin comprometer la expresividad del modelo.
- Permite al modelo adaptar sus pesos de manera más flexible, facilitando el aprendizaje de límites de decisión más precisos entre clases desbalanceadas sin riesgo inmediato de sobreajuste.

Inconvenientes(n)

- Aun sin ser excesivo, este tamaño puede generar cierta sobreadaptación a las clases mayoritarias si no se aplican mecanismos de regularización adecuados.
- En conjuntos de datos muy desbalanceados, esta configuración puede no ser suficiente para compensar la escasa representación de clases minoritarias, especialmente si estas requieren una mayor capacidad para ser correctamente diferenciadas.
- Si las características de entrada tienen una alta correlación entre sí, mantener el mismo número de neuronas puede introducir redundancia, afectando la eficiencia del aprendizaje y aumentando el riesgo de converger a soluciones deficientes.

Tras comentar cuales son las principales ventajas e inconvenientes de utilizar una capa oculta con el mismo número de neuronas que parámetros recibe el modelo, aparentemente, esta arquitectura ofrecera resultados significativamente positivos gracias a su equilibrio. Aunque, si las clases que tienen un número inferior de muestras, tienen relaciones complejas entre sus datos, esta arquitectura puede ser insuficiente para obtener un modelo capaz de distinguir entre algunas de las clases más minoritarias.

Capa oculta con el doble de neuronas que parámetros de entrada

En este apartado se comentan cuales son las principales ventajas e inconvenientes, a priori, de un modelo de clasificación multiclase con el doble de neuronas en la capa oculta que parámetros recibe el modelo.

Ventajas (n*2)

- Una capa oculta más amplia incrementa la capacidad del modelo para aprender representaciones complejas, lo que puede ser especialmente útil para distinguir correctamente clases minoritarias con patrones sutiles o poco definidos.
- Esta configuración ofrece una mayor flexibilidad al modelo para aproximar funciones no lineales, favoreciendo una mejor separación entre clases en escenarios multiclase con alta variabilidad entre categorías.
- Al disponer de más neuronas, el modelo tiene mayor margen para explorar combinaciones de características relevantes durante el entrenamiento, lo cual puede traducirse en una mejora de métricas como macro F1 o recall por clase.

Inconvenientes(n*2)

- El aumento en el número de parámetros incrementa el riesgo de sobreajuste, especialmente en presencia de clases mayoritarias dominantes o cuando el conjunto de entrenamiento es reducido
- Un modelo con mayor complejidad requiere más recursos computacionales y tiempos de entrenamiento más prolongados, lo que puede dificultar su implementación en entornos con restricciones de capacidad o temporales.
- Si no se utilizan mecanismos de control adecuados (como regularización o balanceo de clases), es posible que el modelo aprenda a ajustar principalmente las clases frecuentes, sin mejorar sustancialmente el rendimiento en las minoritarias.

Tras comentar cuales son las principales ventajas e inconvenientes de utilizar una capa oculta con un número de neuronas igual al doble de los parámetros del modelo, aparentemente, esta arquitectura será la que mejores resultados obtenga en la detección de clases minoritarias al tener la capacidad de aprender representaciones más complejas. Sin embargo, esta arquitectura es más propensa a sobreajustes debido la presencia de las clases mayoritarias.

6.5.5. Matriz de confusión para clasificación multiclase

Como se explica en la sección anterior [6.4.1Matriz de confusión para clasificación binaria](#), la matriz de confusión

Para un modelo con múltiples clases, como es el caso del modelo de que se desarrolla en este documento, se debe tener en cuenta que en un modelo neuronal con 9 salidas, la matriz tiene las siguientes interpretaciones:

- **Diagonal principal:** Cada celda de la diagonal principal de la matriz representa cuántas veces la clase real fue correctamente predicha como clase. Este es el caso de las instancias que fueron correctamente clasificadas, y es lo más cercano a un "verdadero positivo" para esa clase específica. Sin embargo, en clasificación multiclase, se suele hablar de aciertos o instancias correctamente clasificadas para cada clase.
- **Fuera de la diagonal:** Las celdas fuera de la diagonal representan falsas clasificaciones. Es decir, representan cuántas veces una instancia de la clase real fue predicha incorrectamente como otra clase.

	Predicción Clase 1	Predicción Clase 2	Predicción Clase 3	Predicción Clase 4	Predicción Clase 5	Predicción Clase 6	Predicción Clase 7	Predicción Clase 8	Predicción Clase 9
Real Clase 1	VP ₁	FP ₁₂	FP ₁₃	FP ₁₄	FP ₁₅	FP ₁₆	FP ₁₇	FP ₁₈	FP ₁₉
Real Clase 2	FP ₂₁	VP ₂	FP ₂₃	FP ₂₄	FP ₂₅	FP ₂₆	FP ₂₇	FP ₂₈	FP ₂₉
Real Clase 3	FP ₃₁	FP ₃₂	VP ₃	FP ₃₄	FP ₃₅	FP ₃₆	FP ₃₇	FP ₃₈	FP ₃₉
Real Clase 4	FP ₄₁	FP ₄₂	FP ₄₃	VP ₄	FP ₄₅	FP ₄₆	FP ₄₇	FP ₄₈	FP ₄₉
Real Clase 5	FP ₅₁	FP ₅₂	FP ₅₃	FP ₅₄	VP ₅	FP ₅₆	FP ₅₇	FP ₅₈	FP ₅₉
Real Clase 6	FP ₆₁	FP ₆₂	FP ₆₃	FP ₆₄	FP ₆₅	VP ₆	FP ₆₇	FP ₆₈	FP ₆₉
Real Clase 7	FP ₇₁	FP ₇₂	FP ₇₃	FP ₇₄	FP ₇₅	FP ₇₆	VP ₇	FP ₇₈	FP ₇₉
Real Clase 8	FP ₈₁	FP ₈₂	FP ₈₃	FP ₈₄	FP ₈₅	FP ₈₆	FP ₈₇	VP ₈	FP ₈₉
Real Clase 9	FP ₉₁	FP ₉₂	FP ₉₃	FP ₉₄	FP ₉₅	FP ₉₆	FP ₉₇	FP ₉₈	VP ₉

Tabla 6.2: Matriz de confusión para clasificación con 9 clases.

6.5.6. Métricas útiles en clasificación multiclase

En el desarrollo de modelos neuronales para clasificación multiclase, es fundamental contar con métricas que permitan evaluar el desempeño general y específico del modelo. A continuación, se presentan las métricas más relevantes, agrupadas según su naturaleza: generales, macro-promediadas y ponderadas.

Métricas Generales

- **Exactitud (*Accuracy*):**

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\{y_i = \hat{y}_i\}} \quad (6.9)$$

La exactitud cuantifica la proporción de predicciones correctas sobre el total de instancias evaluadas. En contextos de clasificación multiclase, proporciona una medida global del rendimiento del modelo. Sin embargo, su uso puede resultar limitado en escenarios donde las clases están desbalanceadas, ya que tiende a favorecer aquellas clases con mayor representación. En estos casos, una alta exactitud puede enmascarar un bajo rendimiento en clases minoritarias, lo que podría comprometer la robustez del modelo. Se considera una métrica de referencia general, pero insuficiente para evaluar en profundidad el comportamiento del modelo en todos los grupos de clase.

Métricas Macro-promediadas

Las métricas macro-promediadas se calculan evaluando individualmente el rendimiento del modelo para cada clase y luego promediando estos resultados. Esta metodología asigna igual peso a todas las clases, independientemente de su frecuencia, lo que permite detectar deficiencias de desempeño en clases poco representadas. Son particularmente útiles en problemas de clasificación multiclase con desequilibrios significativos.

- **Precisión Macro (*Macro Precision*):**

$$\text{Precision}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^C \text{Precision}_c \quad (6.10)$$

La precisión macro evalúa la proporción de verdaderos positivos entre todas las predicciones positivas, calculada individualmente para cada clase y promediada de forma equitativa. Esta métrica permite identificar si el modelo tiende a emitir predicciones incorrectas en ciertas clases, lo cual puede ser crítico en aplicaciones donde los falsos positivos resultan costosos. En entornos multiclase, su utilidad radica en evaluar la calidad de las predicciones para todas las clases de forma uniforme.

- **Sensibilidad Macro (*Macro Recall*):**

$$\text{Recall}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^C \text{Recall}_c \quad (6.11)$$

La sensibilidad macro (o exhaustividad) mide la proporción de verdaderos positivos sobre el total de instancias reales de cada clase. Al promediarla sin ponderar por la frecuencia de clase, esta métrica refleja la capacidad del modelo de detectar correctamente todas las clases, incluidas aquellas menos frecuentes. Resulta esencial en situaciones donde los falsos negativos tienen consecuencias graves o donde se busca una cobertura completa del espacio de clases.

- **F1 Macro (*Macro F1 Score*):**

$$F1_{\text{macro}} = \frac{1}{C} \sum_{c=1}^C F1_c \quad (6.12)$$

El F1 macro se obtiene calculando el F1 score individual por clase (media armónica entre precisión y sensibilidad), y promediándolo uniformemente. Esta métrica proporciona una visión equilibrada del desempeño del modelo en todas las clases, y es especialmente útil cuando se busca un rendimiento homogéneo en un entorno multiclase, independientemente de la distribución de las muestras.

Métricas Ponderadas (*Weighted*)

Las métricas ponderadas (weighted) también se calculan por clase, pero se ajustan según la frecuencia de aparición de cada clase en el conjunto de datos. Este enfoque permite obtener una representación más realista del rendimiento global del modelo, otorgando mayor peso a las clases con mayor número de instancias. Son adecuadas para obtener una evaluación que respete la distribución natural del conjunto de datos, sin perder información relevante de las clases menos representadas.

- **Precisión Ponderada (*Weighted Precision*):**

$$\text{Precision}_{\text{weighted}} = \sum_{c=1}^C \frac{N_c}{N} \cdot \text{Precision}_c \quad (6.13)$$

La precisión ponderada evalúa la calidad de las predicciones positivas, ajustada por la proporción de instancias de cada clase. Esta métrica resulta útil para capturar el rendimiento general del modelo en conjuntos de datos con distribución desigual, sin que las clases poco frecuentes dominen el resultado final.

- **Sensibilidad Ponderada (*Weighted Recall*):**

$$\text{Recall}_{\text{weighted}} = \sum_{c=1}^C \frac{N_c}{N} \cdot \text{Recall}_c \quad (6.14)$$

La sensibilidad ponderada proporciona una medida del porcentaje de instancias reales correctamente identificadas, teniendo en cuenta la frecuencia de cada clase. En modelos de clasificación multiclase, permite analizar cómo se comporta el modelo con respecto a la cobertura general de los datos, favoreciendo una evaluación proporcional al tamaño de las clases.

- **F1 Ponderado (*Weighted F1 Score*):**

$$F1_{\text{weighted}} = \sum_{c=1}^C \frac{N_c}{N} \cdot F1_c \quad (6.15)$$

El F1 ponderado combina precisión y sensibilidad ajustadas por la proporción de cada clase. Es una de las métricas más importantes cuando se desea evaluar un rendimiento global equilibrado que refleje el impacto relativo de cada clase en el conjunto de datos, resultando especialmente útil para comparar modelos entrenados sobre datasets desbalanceados.

Área Bajo la Curva ROC (AUC)

El área bajo la curva ROC (AUC) mide la capacidad del modelo para distinguir entre clases. En clasificación multiclase, se emplean dos métodos principales para su cálculo:

- **AUC One-vs-One (OVO):**

$$\text{AUC}_{\text{ovo}} = \frac{1}{C(C-1)/2} \sum_{i < j} \text{AUC}_{i,j} \quad (6.16)$$

El enfoque One-vs-One calcula el AUC para cada par de clases, evaluando la capacidad del modelo para distinguir una clase frente a otra. El resultado final se obtiene promediando los valores obtenidos para todos los pares posibles. Esta estrategia proporciona una evaluación detallada de la separabilidad entre clases específicas. Es útil cuando se quiere analizar el comportamiento del modelo en decisiones binarias dentro del espacio multiclase.

- **AUC One-vs-Rest (OVR):**

$$\text{AUC}_{\text{ovr}} = \frac{1}{C} \sum_{c=1}^C \text{AUC}_c \quad (6.17)$$

El enfoque One-vs-Rest evalúa, para cada clase, la capacidad del modelo de distinguir dicha clase frente a todas las demás combinadas. Esta métrica permite examinar el rendimiento del modelo desde la perspectiva de cada clase individual y es adecuada para diagnósticos detallados en entornos multiclase.

6.5.7. Mejores hiperparámetros para cada arquitectura

En esta sección se pueden observar los resultados de los experimentos realizados con los posibles valores de los hiperparámetros comentados en la sección [6.5.2 Diseño del modelo e hiperparámetros seleccionados](#).

Para medir la eficiencia y eficacia del modelo, se han obtenido las métricas derivadas de la matriz de confusión comentadas en la sección [?? ??](#). El objetivo de este modelo es clasificar intrusiones en redes, ya detectadas en diferentes tipos específicos de ataque enumerados en la sección [5.2](#) para facilitar una respuesta adecuada. Aunque una clasificación incorrecta puede afectar la eficiencia de la mitigación, las consecuencias no son tan catastróficas como en el caso de falso negativo del modelo de clasificación binaria. Sin embargo, la precisión en la identificación del tipo de ataque sigue siendo clave para optimizar los recursos y la estrategia de defensa. Teniendo esto en cuenta, la métrica más importante es Weifhted F1 socre, ya que como se comenta en el apartado anterior, refleja el impacto relativo de cada clase en el conjunto de datos, resultando especialmente útil para comparar modelos entrenados con datasets desbalanceados.

Otras métricas relevantes que se muestran en las siguientes figuras son:

- **Macro F1 Score:** Indica un buen equilibrio entre precisión y recall promedio por clase, sin importar su frecuencia, lo que refleja buen rendimiento en todas las clases, incluidas las minoritarias.

- **Macro Recall:** Un valor alto significa que el modelo detecta correctamente una alta proporción de instancias reales en cada clase, lo que es crucial para no ignorar clases poco representadas.
- **Macro Precision:** mide qué proporción de predicciones por clase son correctas. Un valor alto implica que el modelo no clasifica erróneamente otras clases como una clase minoritaria.

En la figura 6.11, se encuentran representadas las cinco mejores configuraciones de hiperparámetros para el modelo de clasificación multiclase con una arquitectura de 25 neuronas ($n/2$) en la capa oculta. Coinciden con lo esperado tras el análisis de la sección 6.5.4 para esta arquitectura. La mejor configuración para la arquitectura con la mitad de neuronas en la capa oculta que de parámetros de entrada tiene el modelo, ha sido:

- **Batch size:** 64
- **Epochs:** 100
- **Learning rate:** 0,001

batch_size	epochs	learning_rate	avg_f1_macro	avg_f1_weighted	avg_precision_macro	avg_recall_macro
64	100	0.001	0.3761312308	0.5546364155	0.357460992	0.5508763109
512	100	0.001	0.3747157213	0.5521536932	0.371903664	0.5408947466
128	80	0.001	0.3762934178	0.5513289533	0.3678181316	0.5511853724
256	80	0.001	0.3794112205	0.550052853	0.3645469232	0.5536027517
128	50	0.001	0.3685039647	0.5470481528	0.3687535085	0.5353746999

Figura 6.11: Mejores cinco configuraciones de hiperparámetros del modelo de clasificación multiclase con una capa oculta de 25 neuronas.

La figura 6.12, muestra la representación de las cinco mejores configuraciones de hiperparámetros para el modelo de clasificación multiclase con una arquitectura de 49 neuronas (n) en la capa oculta. Los resultados medios obtenidos son algo peores de lo comentado en el análisis de la sección 6.5.4 para esta arquitectura. La mejor configuración para la arquitectura con la el mismo número de neuronas en la capa oculta que de parámetros de entrada tiene el modelo, ha sido:

- **Batch size:** 256
- **Epochs:** 50
- **Learning rate:** 0,001

batch_size	epochs	learning_rate	avg_f1_macro	avg_f1_weighted	avg_precision_macro	avg_recall_macro
256	50	0.001	0.3893548463	0.5698160035	0.3691349102	0.5548088333
128	100	0.001	0.3938800702	0.5655814612	0.3766473369	0.5620950682
128	80	0.001	0.3883275447	0.5654025284	0.3750873415	0.5486260218
64	50	0.001	0.3825767191	0.5618128611	0.3667403786	0.5507329851
256	80	0.001	0.3902195824	0.558701689	0.3781765329	0.5650141555

Figura 6.12: Mejores cinco configuraciones de hiperparámetros del modelo de clasificación multiclase con una capa oculta de 49 neuronas.

En la figura 6.13, se encuentran representadas las cinco mejores configuraciones de hiperparámetros para el modelo de clasificación multiclase con una arquitectura de 98 neuronas ($n*2$) en la capa oculta. Los resultados obtenidos en esta arquitectura son los que más coinciden con lo explicado en el análisis de la sección 6.5.4 para esta arquitectura. La mejor configuración para la arquitectura con el doble de neuronas en la capa oculta que de parámetros de entrada tiene el modelo, ha sido:

- **Batch size:** 256
- **Epochs:** 100
- **Learning rate:** 0,001

batch_size	epochs	learning_rate	avg_f1_macro	avg_f1_weighted	avg_precision_macro	avg_recall_macro
256	100	0.001	0.4131798458	0.5831225815	0.3948980948	0.5770690195
256	80	0.001	0.4067727363	0.5775525405	0.3853218281	0.5649617083
512	100	0.01	0.3888079348	0.5741997027	0.3728356158	0.5730832422
64	80	0.001	0.3987197961	0.5710195986	0.3775434452	0.5660456156
256	50	0.001	0.397307994	0.5698308505	0.387028818	0.550390812

Figura 6.13: Mejores cinco configuraciones de hiperparámetros del modelo de clasificación multiclase con una capa oculta de 98 neuronas.

Estos resultados no son definitivos, puesto que no se está evaluando la posibilidad de que el modelo haya sufrido sobreajuste incluso utilizando validación cruzada. Para obtener resultados reales del desempeño del modelo, se realizan pruebas, también conocidas como test, en las que se obtienen las métricas del modelo al recibir datos que no ha visto nunca. Este proceso y sus resultados se comentan en el capítulo 7.

6.5.8. Comparación de las arquitecturas

En esta sección se comapan los resultados obtenidos entre las tres arquitecturas propuestas. Una vez comparadas, se desarrolla una posible explicación para comprender los resultados de los experimentos.

Para comprender mejor los resultados globales de los experimentos, se han recogido en la figura 6.14, las 10 configuraciones de hiperparámetros con mayor weighted f1 score independientemente de su arquitectura.

batch_size	epochs	hidden_size	learning_rate	avg_f1_weighted
256	100	98	0.001	0.5831225815
256	80	98	0.001	0.5775525405
512	100	98	0.01	0.5741997027
64	80	98	0.001	0.5710195986
256	50	98	0.001	0.5698308505
256	50	49	0.001	0.5698160035
128	80	98	0.001	0.5669125716
128	100	98	0.001	0.5657792633
128	100	49	0.001	0.5655814612
128	80	49	0.001	0.5654025284

Figura 6.14: Mejores diez configuraciones de hiperparámetros del modelo de clasificación multiclase.

Para poder comparar los resultados de los experimentos del modelo de clasificación multiclase entre varias arquitecturas, es necesario recoger más combinaciones de hiperparámetros independientemente de las arquitecturas que en la sección 6.4.4. Esto denota una gran diferencia entre los resultados de los experimentos del modelo de clasificación multiclase y los resultados del modelo de clasificación binaria. Como se puede observar en la figura 6.8, los resultados del modelo de clasificación binaria, demuestran que la arquitectura no supone una gran diferencia a la hora de que el modelo converja hacia una solución. En cambio, en los resultados de los experimentos del modelo de clasificación multiclase, la arquitectura con un tamaño de la capa oculta igual al doble del número de entradas del modelo ha sido la dominante.

Los resultados de la figura 6.14 reflejan la complejidad del problema. La hegemonía que presenta la arquitectura con un mayor número de neuronas implica que la relación entre los datos es muy compleja.

La dificultad del problema que el modelo debe resolver reside en tres características del mismo:

1. **Clases muy desbalanceadas:** Al igual que en el mundo real, los datos utilizados para entrenar el modelo representan varios tipos de intrusiones informáticas a sistemas. Estos tipos de ataques no tienen los mismos objetivos, tal y como se comenta en el capítulo ?? ???. Como la finalidad de cada tipo de ataque es diferente, el número de ataques de cada clase que se realizan depende de lo que los usuarios que los lanzan deseen conseguir. Esta característica del problema se ve reflejada en la proporción de muestras de cada clase

que contiene el dataset utilizado. Para comprender la dimensión del problema, de las muestras utilizadas para el entrenamiento del modelo, 31 052 pertenecen a ataques de tipo Exploits, mientras que solo 109 de las muestras son de tipo Worms. Esto implica que el número de muestras que son Exploits es 284 veces mayor que el número de muestras de ataques Worms.

2. **Complejidad para relacionar los datos:** Como se comenta en varias de las secciones anteriores de este capítulo, las relaciones entre los datos que se utilizan para el entrenamiento del modelo no son lineales. La complejidad de los datos implica un mayor número de ajustes por retropropagación en cada época de entrenamiento. Esta característica además de aumentar el coste computacional del entrenamiento y el uso del modelo, requiere de poder ajustar los pesos de la red con mayor precisión que en modelos con una complejidad menor, como es el caso del modelo de clasificación binaria propuesto en este proyecto.
3. **Número de clases elevado:** Cuanto mayor es el número de clases entre las que debe diferenciar un modelo de clasificación multiclase, mayor número de muestras de cada clase son necesarias para entrenar el modelo y mayor número de situaciones de clasificación erróneas existe.

El modelo que se ha diseñado debe identificar 9 clases distintas de intrusiones pero debido a la complejidad de obtener datos representativos de cada intrusión, el número de muestras utilizado para entrenar al modelo es de aproximadamente 72 000. De ese número de muestras, 60 000 pertenecen solo a 3 tipos de clases.

Si tomamos de punto de partida la matriz de confusión del modelo de clasificación binaria, existen 4 posibles respuestas del modelo, de las cuales 2 de ellas son correctas. Esto implica que, en el hipotético caso, de que el modelo simplemente diese respuestas aleatorias, este daría la respuesta correcta 2 de cada 4 veces, o lo que es lo mismo el 50 % de las ocasiones. En el caso del modelo de clasificación multiclase, la matriz de confusión tiene 81 posibilidades diferentes. De estas posibilidades, solo las que pertenecen a la diagonal principal son respuestas correctas, lo que equivale a 9 respuestas correctas. Aplicando la misma lógica que se ha aplicado para el modelo de clasificación binaria, si el modelo de clasificación multiclase solo diese respuestas de manera aleatoria, acertaría en 9 de cada 81 ocasiones, lo que se traduce en una tasa de acierto del 11,1 %. Estas tasas de acierto corresponden con la métrica de precisión de cada modelo. Sin embargo, las métricas utilizadas para evaluar qué combinación de hiperparámetros es mejor para cada modelo, se ven especialmente afectadas por los falsos negativos y por la clasificación errónea de clases minoritarias, correspondientemente.

Todas estas características comentadas provocan que la matriz de confusión del modelo de clasificación multiclase sea la que se muestra en la figura 6.15. La figura muestra que de las nueve clases que el modelo es capaz de identificar, solo 3 de ellas se clasifican con exactitud en la mayoría de las ocasiones. Destaca el número de predicciones erróneas de las clases más minoritarias que han sido clasificadas como clase 3 (Exploits). Esto se debe al desbalanceo presente en los datos comentado a lo largo de este capítulo.

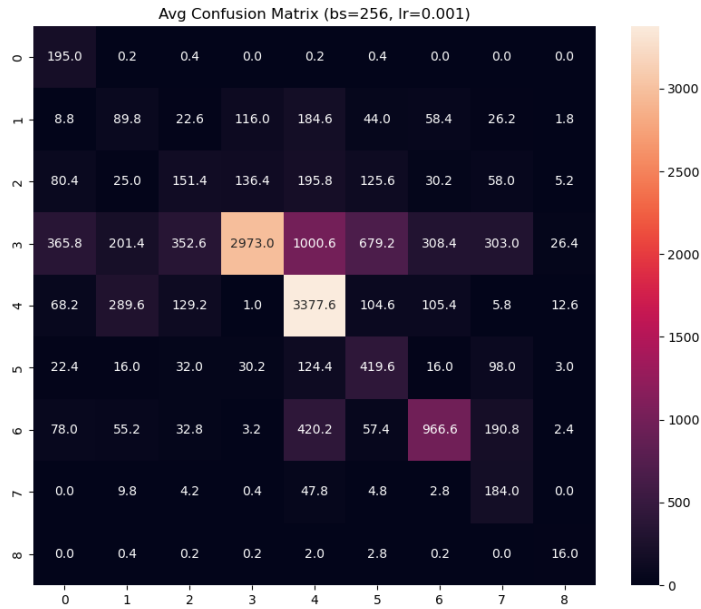


Figura 6.15: Matriz de confusión de la mejor configuración de hiperparámetros encontrada para el modelo de clasificación multiclase.

Las clases de la figura 6.15 se encuentra codificadas de la siguiente manera:

- Clase 0: Analysis,
- Clase 1: Backdoor.
- Clase 2: DoS.
- Clase 3: Exploits.
- Clase 4: Fuzzers.
- Clase 5: Generic.
- Clase 6: Reconnaissance.
- Clase 7: Shellcode.
- Clase 8: Worms.

Capítulo 7

Test

Capítulo 8

Despliegue

Capítulo 9

Tecnologías usadas

Capítulo 10

Seguimiento del proyecto

Capítulo 11

Conclusiones

Apéndice A

Manuales

A.1. Manual de despliegue e instalación

A.2. Manual de mantenimiento

A.3. Manual de usuario

Apéndice B

Resumen de enlaces adicionales

Los enlaces útiles de interés en este Trabajo Fin de Grado son:

- Repositorio del código: <https://gitlab.inf.uva.es/>.

Bibliografía

- [1] Pablo Haya. La metodología crisp-dm en ciencia de datos, noviembre 2021. Consultado el 18 de mayo de 2025.
- [2] Ken Schwaber and Jeff Sutherland. The scrum guide: The definitive guide to scrum – the rules of the game. <https://www.scrum.org/resources/scrum-guide>, 2020.
- [3] IONOS España. Tcp protocol: así funciona el protocolo de transmisión, 2020.
- [4] Vicente González Ruiz. Tcp (transmission control protocol), December 2014.
- [5] Vicente González Ruiz. El ip (internet protocol), December 2014.
- [6] AprendeIA. ¿qué es deep learning?, 2021.
- [7] Daniel Nielson Unite.AI. ¿qué son las redes neuronales convolucionales (cnn)?, 2020.
- [8] MSMK University. Red neuronal de retropropagación (backpropagation neural network), 2023.
- [9] James Rundle. The ai effect: Amazon sees nearly 1 billion cyber threats a day. *The Wall Street Journal*, December 2024.
- [10] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc., 2000.
- [11] Kent Beck, Mike Beedle, Arie van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, Jon Kern, Brian Marick, Robert C. Martin, Steve Mellor, Ken Schwaber, Jeff Sutherland, and Dave Thomas. Manifesto for agile software development. <https://agilemanifesto.org/>, 2001.
- [12] Project Management Institute. *A Guide to the Project Management Body of Knowledge (PMBOK Guide)*. Project Management Institute, 2021.
- [13] Diego Carmona Fernández and Silvia Román Suero. *Gestión de Riesgos: Fundamentos sobre la gestión de riesgos en los proyectos*. Universidad de Cádiz, 2021.

- [14] Miguel Ángel Oliveros Villegas and Haydeé Cecilia Rincón de Parra. Gestión de costos en los proyectos: un abordaje teórico desde las mejores prácticas del project management institute. *Visión Gerencial*, 10(1):85–94, 2011.
- [15] A. S. Tanenbaum. *Computer Networks*. Prentice Hall, 5th edition, 2009.
- [16] J. Scott. A comprehensive study on denial of service attacks. *International Journal of Computer Science*, 12(4):45–60, 2015.
- [17] International Information System Security Certification Consortium. Phishing attacks and how to avoid them, 2018. <https://www.isc2.org>.
- [18] Cosmikal. ¿qué es y cómo funciona un firewall? guía básica 2025, december 2024.
- [19] Geekflare. Ids vs ips: A comprehensive guide to network security solutions, december 2024.
- [20] Palo Alto Networks. ¿qué es la microsegmentación?
- [21] National Institute of Standards and Technology (NIST). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations, 2021.
- [22] John Brooke. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194), 1996.
- [23] Majed Luay, Siamak Layeghy, Seyedehfaezeh Hosseininoorbin, Mohanad Sarhan, Nour Moustafa, and Marius Portmann. Temporal analysis of netflow datasets for network intrusion detection systems, 2025.
- [24] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [25] Simon Haykin. *Neural Networks and Learning Machines*. Pearson, 2009.
- [26] Michael Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.
- [27] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.
- [28] EITCA Academy. ¿cuál es el papel de la función de pérdida en el aprendizaje automático? Accedido: 2025-05-22.
- [29] Ultralytics. Función de pérdida. Accedido: 2025-05-22.
- [30] DataCamp. Explicación de las funciones de pérdida en el machine learning. Accedido: 2025-05-22.
- [31] BigDataFran. 7. introducción problemas de clasificación — trabajando con pytorch. Accedido: 2025-05-22.
- [32] Léon Bottou. Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT 2010*, pages 177–186, 2010.
- [33] D.P. Kingma and J.B. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [34] François Chollet. *Deep learning with Python*. Manning Publications Co., 2018.