MÁSTER EN FORMACIÓN PERMANENTE EN INTELIGENCIA ARTIFICIAL

Impacto y degradación de modelos explicables de machine learning aplicados a las áreas de salud, seguridad y defensa nacional de España ante ataques de envenenamiento de datos

Presentado por:

Jenny Aguilar Guiñez

Dirigido por:

Vicente Castillo Faulí

CURSO ACADÉMICO 2024-2025

Resumen

En el contexto actual del avance acelerado de las tecnologías basadas en inteligencia artificial, ante la creciente presión por alcanzar una posición competitiva a nivel internacional, la adopción de modelos de aprendizaje automático ha crecido exponencialmente en los sectores públicos y privados, en áreas tan críticas como el sector salud, seguridad y defensa nacional. Estos sectores exigen cada vez más el uso de modelos de inteligencia artificial que sean explicables, como los árboles de decisión, debido a la necesidad de garantizar la trazabilidad, interpretabilidad y rendición de cuentas en los procesos que han sido automatizados. Sin embargo, esta misma característica, que los hace valiosos, también los convierte en objetivos vulnerables para ataques como el envenenamiento de datos, capaces de alterar su funcionamiento desde la etapa de entrenamiento sin generar indicadores detectables inicialmente.

El objetivo de este trabajo es demostrar cómo, pese a sus ventajas, estos modelos pueden deteriorarse de manera significativa ante ataques dirigidos o indiscriminados, comprometiendo no solo su precisión, sino también su lógica interna. Para ello, se llevará a cabo un experimento en dos fases: una primera con entrenamiento con datos sin envenenar para establecer métricas de referencia, y una segunda con datos envenenados que permitirá observar el impacto adverso en su rendimiento y estructura.

A través de este análisis, se pretende alertar sobre la necesidad de implementar estrategias de defensa para modelos explicables en contextos críticos, subrayando que la transparencia que proporcionan no garantiza por sí sola la seguridad. Esta línea de investigación se alinea con los Objetivos de Desarrollo Sostenible, al contribuir al desarrollo de sistemas confiables, éticos y tecnológicamente responsables en servicios esenciales para la sociedad.

Palabras clave: Aprendizaje automático, Explicabilidad, Ciberseguridad, Árboles de decisión, Envenenamiento de datos, Evaluación de modelos

Abstract

In the current context of the rapid advancement of technologies based on artificial intelligence, and under growing pressure to achieve a competitive international position, the adoption of machine learning models has increased exponentially in both public and private sectors, particularly in critical areas such as health, security, and national defense. These sectors increasingly demand the use of explainable artificial intelligence models, such as decision trees, due to the need to ensure traceability, interpretability, and accountability in automated processes. However, this very characteristic, which makes them valuable, also renders them vulnerable targets for attacks such as data poisoning, capable of altering their functioning from the training stage without generating initially detectable indicators.

The aim of this work is to demonstrate how, despite their advantages, these models can significantly deteriorate under targeted or indiscriminate attacks, compromising not only their accuracy but also their internal logic. To this end, the study will carry out a two-phase experiment: the first with training on clean data to establish baseline metrics, and the second with poisoned data to observe the adverse impact on performance and structure.

Through this analysis, the intention is to highlight the urgent need for defense strategies in explainable models within critical contexts, emphasizing that transparency alone does not guarantee security. This line of research aligns with the Sustainable Development Goals by contributing to the development of trustworthy, ethical, and technologically responsible systems in essential services for society.

Keywords: Machine Learning, Explainability, Cybersecurity, Decision Trees, Data Poisoning, Model Evaluation

Índice

Índice de figuras

Índice de cuadros

1. Introducción

1.1. Justificación

La creciente adopción de sistemas de inteligencia artificial (IA) en sectores críticos como la salud, la seguridad y la defensa nacional, ha sido motivada por múltiples factores. Las políticas públicas como la aprobación de la *Estrategia de Inteligencia Artificial 2024* (Ministerio para la Transformación Digital y de la Función Pública, 2024) han supuesto un impulso institucional. Sin embargo, también han puesto de manifiesto la necesidad de abordar los desafíos éticos fundamentales de la IA, que la Unión Europea (UE) resume en cuatro: la justicia, respeto por la autonomía humana, la previsión del daño y la explicabilidad (Ortiz de Zárate Alcarazo, 2022). Por tanto, en este contexto, el presente Trabajo Fin de Máster (TFM), se justifica desde una triple perspectiva: institucional, científica y ética.

Desde el punto de vista legal e institucional, la *Agenda España Digital 2026* (Ministerio de Asuntos Económicos y Transformación Digital, 2022) establece una hoja de ruta para la transformación digital del país. Según esta hoja de ruta, la ciberseguridad y la IA son pilares fundamentales para el desarrollo económico y social. Esta agenda enfatiza la necesidad de garantizar la seguridad y la confianza en los sistemas digitales, especialmente en sectores estratégicos. Además, el Reglamento de Inteligencia Artificial de la Unión Europea (Unión Europea, 2024) establece normas armonizadas en materia de IA y subraya la importancia de la trazabilidad y la explicabilidad de los sistemas. Esto resulta especialmente crítico para modelos aplicados a ámbitos sensibles como la salud, la seguridad o la defensa.

Desde una perspectiva científica, el envenenamiento de datos se reconoce como una amenaza silenciosa y progresiva en los entornos de aprendizaje automático (Biggio & Roli, 2018). Este tipo de ataque compromete la integridad del modelo puesto que introduce ejemplos maliciosos en el conjunto de entrenamiento, sin generar alertas inmediatas en su funcionamiento. Estudios recientes en el ámbito de la ciberseguridad y la IA, como los realizados por organizaciones internacionales (OWASP, 2024) e investigadores (Calzavara et al., 2025) estudian el modo en que este tipo de amenazas puede modificar

significativamente el comportamiento de los sistemas. Estas amenazas plantean un desafío particular en modelos explicables como los árboles de decisión, cuya facilidad de interpretación y transparencia puede inducir a una falsa sensación de seguridad. Es posible, por tanto, que esta aparente sensación de seguridad dé lugar a la errónea percepción de que las medidas de protección pueden ser postergadas o completamente mitigadas si se introducen métodos para mejorar la robustez como los Random Forest, no obstante, estos enfoques son equivocados como exponen estudios previos (Drews et al., 2020; Chang & Im, 2020; Chen et al., 2019).

Este TFM contribuirá a visibilizar los riesgos asociados al envenenamiento de datos en modelos explicables, con un enfoque particular en los árboles de decisión, por ser uno de los modelos más extendidos gracias a su simplicidad, transparencia y capacidad de explicación (Mienye & Jere, 2024).

Desde una perspectiva ética, se pretende impulsar una labor continua y comprometida tanto en el diseño de estrategias de defensa para los modelos de IA, como en el desarrollo de líneas de investigación orientadas a una IA coherente con una visión explicativa, ética y profundamente responsable. Esta labor no solo busca resguardar la integridad técnica de los modelos, sino también reconocer su verdadero propósito, más allá de las métricas, entendiendo que la motivación subyacente de este trabajo es proteger a las personas cuyas realidades pueden verse profundamente afectadas por decisiones automatizadas en sectores críticos que impactan directamente en sus vidas y su dignidad.

1.1.1. Objetivos de Desarrollo Sostenible (Agenda 2030).

Este trabajo aspira a contribuir a los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030, propuestos por la Organización de las Naciones Unidas. En particular, se relaciona directamente con los siguientes objetivos:

- **ODS 3 – Salud y Bienestar:** Al centrarse en la confiabilidad de modelos de IA en el ámbito sanitario, este trabajo promueve un uso más seguro y efectivo de las tecnologías en la salud pública.
- **ODS 9 – Industria, Innovación e Infraestructura:** Aporta al desarrollo de infraestructuras digitales resilientes mediante el análisis y fortalecimiento de modelos

de aprendizaje automático ante posibles ataques de envenenamiento de datos.

- **ODS 16 – Paz, Justicia e Instituciones Sólidas:** Al estudiar algoritmos explicables y la degradación de modelos utilizados en seguridad y defensa, se contribuye a una toma de decisiones más transparente y confiable en instituciones públicas.

Estos objetivos proporcionan un marco de referencia adicional que guía el propósito social y ético de esta investigación.

1.1.2. Motivación personal

El presente trabajo se enmarca en una inquietud que ha acompañado mi desarrollo profesional a lo largo de más de 20 años de experiencia laboral en sectores estratégicos como la seguridad ciudadana, la defensa de la nación y la gestión de instituciones de salud. Mi paso por instituciones como la Defensa Civil, Escuela de Investigaciones Policiales y el Ejército de Chile, así como mi experiencia dirigiendo organizaciones vinculadas a servicios sanitarios y tecnológicos, han reforzado mi convicción sobre la importancia de construir sistemas confiables, auditables y éticamente responsables. En los últimos años, esta preocupación se ha profundizado en mi rol como ingeniera en IA, desde donde he podido observar los desafíos reales que enfrenta la adopción de modelos de aprendizaje automático en contextos sensibles.

La elección de esta temática responde, por tanto, a una motivación ética, profesional y técnica. Abordar el estudio de cómo los ataques de envenenamiento de datos afectan a modelos explicables, como los árboles de decisión, no solo representa una oportunidad académica, sino una contribución práctica a concienciar sobre el problema, lo cual es un paso fundamental para plantearse cualquier posible vía de solución en aras de diseñar sistemas seguros y transparentes. Este trabajo constituye también el punto de partida para una línea de investigación en el ámbito de la ciberseguridad aplicada a la IA, con un enfoque compartido entre las crecientes necesidades en esta área desde el sector de la empresa privada, y también desde el sector público. Frente al interés común que representa este último, se busca aportar en áreas como la defensa de infraestructuras críticas, la mejora de servicios esenciales y la contribución a la credibilidad y fortalecimiento institucional dado su alto impacto social.

1.1.3. Metodología para el desarrollo del TFM

El desarrollo de este trabajo se estructura siguiendo el modelo Cross Industry Standard Process for Data Mining (CRISP-DM), ampliamente utilizado en proyectos de ciencia de datos y aprendizaje automático. Esta metodología contempla seis fases interdependientes: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. En el contexto del presente TFM, se adaptarán estas etapas de la siguiente manera:

- La fase de comprensión del negocio se centra en la identificación de las vulnerabilidades en modelos que utilizan algoritmos explicables como los árboles de decisión, aplicados en sectores críticos como en salud, seguridad y defensa.
- La comprensión y preparación de los datos se llevará a cabo mediante la selección y análisis de datasets apropiados para el entrenamiento y la generación de datos envenenados.
- El modelado incluirá la implementación de modelos con árboles de decisión en ambos contextos (entrenados con datos sin y con envenenamiento).
- La evaluación considerará métricas clásicas para problemas de regresión y clasificación, así como para el análisis de la estructura explicativa de los árboles.
- Finalmente, los hallazgos serán analizados en función del comportamiento observado de los modelos bajo condiciones adversariales, con el objetivo de evaluar cambios en la estructura y en el rendimiento ante manipulaciones en los datos de entrenamiento. No se contempla una fase de despliegue en producción, ya que el alcance del trabajo se limita a la experimentación controlada.

Este enfoque estructurado garantiza la coherencia metodológica del trabajo y permite una trazabilidad clara de cada una de las decisiones tomadas a lo largo del desarrollo del TFM. Además, este marco de gestión se complementa con la perspectiva epistemológica y experimental desarrollada en la sección 3, en coherencia con el paradigma que busca objetividad, medición, comprobación y predicción, es decir, el paradigma positivista adoptado, tal como se explica en esa sección. De forma transversal, las fases de

CRISP-DM se han articulado de manera que cada una contribuya al cumplimiento de los objetivos específicos planteados en el TFM, asegurando la alineación entre metodología y resultados esperados.

1.2. Problema y finalidad

Tal como se expuso en el apartado anterior, en sectores críticos como la salud, la seguridad y la defensa nacional, la adopción de IA ha sido especialmente promovida por políticas públicas nacionales y normativas internacionales recientes.

No obstante, esta búsqueda de una hiperautomatización de procesos y la adopción de modelos de IA en áreas críticas, ha generado también un creciente interés por parte de los reguladores y la sociedad civil (Preece et al., 2018) en torno a la explicabilidad y gobernanza en el uso de estos sistemas (Grimmelikhuijsen & Meijer, 2022).

En estos contextos, donde las decisiones automatizadas pueden tener consecuencias críticas en la vida de las personas o en la confianza hacia nuestras estructuras organizacionales, la explicabilidad de los modelos no son una opción, sino un requisito ético cada vez más exigible, ya que se pretende garantizar la trazabilidad, auditar decisiones y facilitar la rendición de cuentas por parte de las instituciones que las implementan (Ortiz de Zárate Alcarazo, 2022). Este requisito de transparencia ha favorecido la difusión de los modelos explicables, como los árboles de decisión, para la resolución de diversas problemáticas en áreas críticas tanto en seguridad (Gómez et al., 2023; Cuesta Calvo et al., 2018) y defensa (Lewis et al., 2016), como en salud (Cordero et al., 2024; Ministerio de Sanidad, s.f.), puesto que ofrecen una representación clara y comprensible de las decisiones tomadas por el modelo, lo que los convierte en piezas clave para la legitimación del uso de la IA en dominios de alto impacto social (Bishop, 2009; Russell & Norvig, 2004).

Pero la real problemática surge de esta misma característica de transparencia y facilidad de interpretación, ya que se expone como una de las principales causas de vulnerabilidad de estos algoritmos (Barredo Arrieta et al., 2020). Al ser modelos cuyas reglas de decisión pueden anticiparse con relativa facilidad, los algoritmos explicables como los árboles de decisión, se exponen a manipulaciones dirigidas desde la fase de entrenamiento. Así sucede con el envenenamiento de datos, que puede alterar la estructura interna del algoritmo sin generar alertas evidentes.

Una vez expuesto el contexto, el problema adquiere mayor relevancia: si bien los ataques de envenenamiento de datos han sido ampliamente estudiados en el ámbito de redes neuronales y otros modelos complejos, los árboles de decisión han recibido una atención comparativamente menor, a pesar de su uso extensivo en el análisis de datos tabulares debido a su eficacia (Calzavara et al., 2025). Esta brecha en la literatura pone de relieve la necesidad de examinar el impacto de estas amenazas en dicho algoritmo, especialmente en entornos sensibles, donde la trazabilidad y la rendición de cuentas resultan fundamentales (Ramirez et al., 2022).

Al considerar que los árboles de decisión, así como otros modelos explicables, seguirán siendo utilizados como una herramienta útil en sectores críticos, este trabajo busca analizar empíricamente el comportamiento de árboles de decisión entrenados con datos manipulados, comparándolos con modelos entrenados con datos no alterados. La finalidad es evidenciar el deterioro potencial de su rendimiento, proporcionar métricas que permitan demostrar su degradación y generar conciencia sobre los riesgos asociados, con el fin de promover futuras estrategias de defensa adaptadas a entornos de alta sensibilidad.

1.3. Objetivos del TFM

Objetivo principal

Analizar y documentar el impacto y la degradación de los ataques de envenenamiento de datos sobre modelos de ML explicables de regresión y clasificación, utilizando árboles de decisión para el caso de estudio, con el fin de evidenciar su vulnerabilidad ante ataques adversariales de envenenamiento, obtener criterios de evaluación técnica y evaluar las implicaciones de su implementación en sectores críticos como salud, seguridad y defensa.

Objetivos específicos

a) **Vulnerabilidad asociada a la explicabilidad:** Evidenciar la vulnerabilidad ante ataques adversariales de envenenamiento, asociada a la explicabilidad de los árboles de decisión.

b) **Modos de envenenamiento de datos:** Identificar tipos de ataque de envenenamiento de datos aplicables a árboles de decisión.

c) **Deterioro por envenenamiento de datos:** Obtener métricas para medir el potencial deterioro de un modelo explicable basado en árbol de decisión, sometido a envenenamiento de datos.

d) **Riesgos en contextos críticos:** Evaluar el *impacto*, en forma de riesgos operativos y de gobernanza, del uso de este modelo en contextos críticos como salud, seguridad y defensa nacional.

e) **Recomendaciones para prevenir y mitigar los riesgos:** Recopilar y recomendar medidas de prevención y mitigación del riesgo de envenenamiento de datos en modelos explicables, particularmente sobre árboles de decisión.

1.4. Planificación

Fase 1: Estado del arte y preparación de datos

- Revisión de la literatura.
- Búsqueda y selección de la bibliografía y referencias.
- Búsqueda y selección de los dataset.
- Preparación de los dataset, preprocesamiento y limpieza de datos.

Fase 2: Diseño e implementación de los experimentos

- Diseño de los experimentos y selección de métricas.
- Desarrollo del mecanismo de generación de datos envenenados.
- Definición de criterios de evaluación
- Entrenamiento de modelos con datos no alterados y envenenados.

Fase 3: Análisis de resultados y conclusiones

- Evaluación de métricas de rendimiento por experimento.
- Análisis comparativo de resultados entre experimentos.
- Conclusiones.
- Ajustes finales.

2. Marco teórico

2.1. Aprendizaje supervisado: fundamentos, clasificación y regresión

El aprendizaje supervisado es una de las ramas fundamentales del aprendizaje automático (Machine Learning o ML), y se caracteriza por la existencia de un conjunto de entrenamiento conformado por pares de datos del tipo entrada-salida, donde el objetivo es aprender una función que pueda predecir la salida correspondiente a nuevas entradas no vistas durante el entrenamiento del modelo (Russell & Norvig, 2004). Este paradigma se utiliza principalmente para realizar dos tareas: clasificación, cuando las salidas pertenecen a un conjunto discreto de clases, y regresión, cuando las salidas son valores continuos.

En la clasificación, el modelo intenta asignar cada observación de entrada a una de varias categorías predefinidas. Por ejemplo, en el ámbito médico, esto puede traducirse en determinar si un paciente padece o no una enfermedad específica, o clasificar el nivel de riesgo de hospitalización según factores de morbilidad. Por su parte, en la regresión, el objetivo es predecir una variable cuantitativa, como lo sería en el ámbito de la seguridad y defensa nacional, predecir la tasa de criminalidad en una región determinada, la probabilidad de éxito de una operación de inteligencia, o el tiempo de respuesta ante una emergencia.

Ambos enfoques comparten técnicas y modelos base, entre ellos los árboles de decisión, redes neuronales, máquinas de vectores de soporte, y modelos basados en inferencia bayesiana.

2.1.1. Métricas de evaluación de rendimiento para modelos supervisados

En el aprendizaje automático supervisado, las métricas de evaluación permiten cuantificar el rendimiento de un modelo respecto a su capacidad de generalizar sobre datos no vistos durante el entrenamiento. Estas métricas externas son independientes del algoritmo utilizado, y se aplican tras el entrenamiento para comparar resultados en tareas de clasificación o regresión.

A continuación, se describen solo algunas métricas comúnmente usadas para medir el desempeño de los modelos, enfocándose en aquellas que serán utilizadas en la fase

experimental de este TFM, diferenciadas según el tipo de experimento (regresión o clasificación). La implementación de estas métricas, sobre árboles de decisión, se detalla la sección 2.3.2.

2.1.1.a Evaluación de rendimiento de modelos de clasificación

Los modelos de clasificación predicen una etiqueta discreta. Para evaluar su rendimiento, una herramienta ampliamente utilizada es la Matriz de confusión.

	Predicción Positiva	**Predicción Negativa**
Real Positivo	Verdaderos Positivos (TP)	Falsos Negativos (FN)
Real Negativo	Falsos Positivos (FP)	Verdaderos Negativos (TN)

Cuadro 1: Matriz de confusión

De esta matriz se desprenden las métricas más comunes como:

- **Accuracy (Exactitud):** proporción de instancias correctamente clasificadas entre el total de predicciones. Es sensible a clases desbalanceadas. Formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- **Precision (Precisión):** proporción de verdaderos positivos entre todos los elementos clasificados como positivos.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

- **Recall (Sensibilidad o Tasa de verdaderos positivos):** proporción de verdaderos positivos entre todos los elementos que realmente pertenecen a la clase positiva.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

- **F1-Score:** media armónica entre precisión y recall, útil cuando se requiere equilibrio entre ambas.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4}$$

- **Average Precision - AP (Precisión Promedio):** métrica que resume el área bajo la curva de precisión-recall (PR), calculada como el promedio de las precisiones obtenidas para cada nivel de recall. Es especialmente útil en contextos con clases desbalanceadas.

$$AP = \sum_{n} (R_n - R_{n-1}) \cdot P_n \tag{5}$$

  donde P_n es la precisión en el umbral n, y R_n es el recall correspondiente.

- **Receiver Operating Characteristic - ROC (Curva de Característica operativa del receptor):** es la representación gráfica de la capacidad del modelo para distinguir entre clases, representando la tasa de verdaderos positivos frente a la tasa de falsos positivos.

2.1.1.b Evaluación de rendimiento de modelos de regresión

Los modelos de regresión predicen un valor continuo. Sus métricas más comunes incluyen:

- **MAE (Mean Absolute Error):** error absoluto medio entre las predicciones y los valores reales. Fórmula:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{6}$$

  Interpretación de resultados: valores más bajos indican mejor ajuste del modelo, pero no penaliza errores grandes.

- **MSE (Mean Squared Error):** error cuadrático medio entre las predicciones y los valores reales. Fórmula :

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{7}$$

  Interpretación de resultados: valores más bajos indican mejor ajuste del modelo. Penaliza más los errores grandes debido al cuadrado de las diferencias.

- **R^2 Score (Coeficiente de determinación):** mide la proporción de varianza explicada por el modelo, independientemente de la escala de la variable objetivo. Fórmula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{8}$$

Interpretación de resultados: valores cercanos a 1 indican un buen ajuste del modelo, mientras que valores negativos sugieren un mal desempeño.

2.2. Modelos explicables de IA: principios y taxonomía

Como ya se ha justificado inicialmente en este trabajo, la explicabilidad de los modelos de IA es un requisito fundamental en sectores críticos, donde la transparencia y la rendición de cuentas son esenciales. Atender a este requisito se ha vuelto relevante, especialmente en dichos sectores, puesto que las decisiones automatizadas pueden tener un impacto significativo en la vida de las personas. Ante estos desafíos de transparencia surge el campo de la inteligencia artificial explicable (eXplainable AI - XAI) (Barredo Arrieta et al., 2020), cuyo objetivo es garantizar que los sistemas de IA puedan justificar sus decisiones de forma inteligible para los humanos.

La XAI se enfrenta a la tensión clásica entre precisión predictiva y comprensión del modelo, y se la entiende como el conjunto de métodos que "aportan evidencia o razones comprensibles que justifican cada salida del sistema" (Phillips et al., 2021)

2.2.1. Principios fundamentales

El Instituto Nacional de Estándares y Tecnología (NIST, por sus siglas en inglés) de EE.UU. ha propuesto cuatro principios clave que toda IA explicable debe cumplir para ser considerada confiable y transparente (Phillips et al., 2021):

- *Explicabilidad (Explainability)* La IA debe ofrecer explicaciones claras sobre cómo genera sus resultados.
- *Justificabilidad (Meaningful)* Las explicaciones deben ser comprensibles y relevantes para el usuario final.
- *Exactitud (Accuracy)* Las explicaciones deben reflejar fielmente el comportamiento real del sistema.
- *Límites del conocimiento (Knowledge Limits)* La IA debe identificar cuándo no tiene suficiente confianza para emitir una respuesta.

El hecho de que en la literatura se ha tendido a confundir explicabilidad con interpretabilidad, haciendo uso indistinto de esta y otras terminologías similares, se reconoce como el primer obstáculo para la comprensión de la XAI (Barredo Arrieta et al., 2020).

Sin embargo, investigaciones recientes, como las recogidas por Ortiz de Zárate (Ortiz de Zárate Alcarazo, 2022), enfatizan la necesidad de diferenciarlas. Este trabajo recoge las definiciones del NIST, diferenciando que mientras la **interpretabilidad** se refiere a la transparencia intrínseca del modelo (comprender directamente su estructura y funcionamiento), la **explicabilidad** abarca también aquellas técnicas externas que permiten hacer comprensibles los modelos complejos mediante mecanismos de post análisis (Phillips et al., 2021).

Por esta razón, se habla no sólo de algoritmos explicables, sino de **modelos explicables** (Barredo Arrieta et al., 2020), incluyendo tanto los algoritmos como las herramientas externas que permiten dotarlos de significado. Esta distinción es relevante, ya que permite abordar la explicabilidad con una perspectiva más amplia, considerando la naturaleza del modelo y las técnicas utilizadas para interpretarlo.

2.2.2. Clasificación de modelos explicables

Desde un punto de vista conceptual, la literatura distingue dos perspectivas diferentes (Barredo Arrieta et al., 2020):

2.2.2.a Modelos de ML transparentes

Estos modelos son comprensibles por sí mismos, es decir, son interpretables por diseño sin necesidad de aplicar técnicas adicionales, por ejemplo:

- Árboles de decisión (Decision Trees)
- Aprendizaje basados en reglas (Ruled-based Learning)
- Regresión lineal o logística
- Naïve Bayes (en ciertos contextos simples)

Los árboles de decisión pertenecen a esta categoría. Ofrecen una estructura jerárquica de decisiones, donde cada nodo representa una condición de decisiones que puede se-

guirse de forma lógica, visual y textual. Cada nodo representa una condición claramente interpretable, y cada rama un camino de decisión verificable por humanos.

2.2.2.b Modelos de ML con técnicas de explicabilidad post-hocs

Estos modelos no son interpretables por sí mismos, se consideran de "caja negra", esto quiere decir que su funcionamiento interno no es comprensible por sí mismo. Sin embargo, pueden ser analizados mediante herramientas externas que generan explicaciones aproximadas de su comportamiento. Ejemplos de algoritmos complejos incluyen:

- Redes neuronales profundas o Deep Neural Networks (DNN)
- Redes Neuronales Convolucionales o Convolutional Neural Networks (CNN)
- Redes Neuronales Recurrentes o Recurrent Neural Networks (RNN)
- Máquina de Soporte Vectorial o Support Vector Machines (SVM)
- Métodos ensembled (como Random Forest o XGBoost)

Herramientas de explicabilidad comunes:

- Explicaciones locales e independientes del Modelo (Local Interpretable Model-Agnostic Explanations, LIME)
- Explicaciones Aditivas de Shapley (Shapley Additive Explanations, SHAP)
- Explicaciones Contrafactuales (Counterfactual Explanations)
- Mapas de Saliencia (Saliency Maps, aplicados principalmente a imágenes)

2.2.2.c Otros modelos emergentes

Cabe mencionar que, actualmente y en forma incipiente, comienzan a desarrollarse técnicas que podrían constituir una nueva clasificación como modelos híbridos o modelos que integran explicabilidad mediante el uso de arquitecturas tipo Transformers, conocidas por sus mecanismos de atención, que permiten identificar qué partes de la entrada han influido más en la predicción del modelo pero como están en un estado de desarrollo inicial, no se incluyen como una clasificación aceptada aún. Ejemplos de estas técnicas emergentes incluyen:

- Modelos de atención interpretativa con Procesamiento de Lenguaje Natural (como Transformers con visualización de atención) (Fantozzi & Naldi, 2024)
- Modelos destilados + reglas simbólicas (Tan et al., 2018)
- Sistemas de Reglas Inducidas a partir de Redes Neuronales (Distill & Match) (Sun et al., 2025)
- Frameworks como AUTOLYCOUS, que usan XAI para extraer modelos interpretables desde un modelo original bajo "acceso de caja negra". (Oksuz et al., 2024)

A pesar de que existen investigaciones que promueven los beneficios de los modelos transparentes por sobre modelos de caja negra que deban ser explicados (Rudin, 2019), los usos de ambos enfoques son ya tangibles en dominios donde la trazabilidad de la decisión es un requisito legal o ético. En salud, los árboles clínicos y modelos de riesgo basados en reglas facilitan la auditoría médica y la comunicación con el paciente. En finanzas se emplean explicaciones de tipo SHAP o LIME para justificar concesiones de crédito y detectar sesgos sistémicos. En seguridad se han utilizado árboles de decisión para la predicción de delitos (Cuesta Calvo et al., 2018).

No obstante lo anterior, recientes trabajos demuestran que muchas explicaciones también pueden ser vector de diversos ataques o hacer más efectivos los ataques (Barredo Arrieta et al., 2020; Ramirez et al., 2022), como la extracción de modelos o el envenenamiento de datos, subrayando la necesidad de salvaguardas adicionales (Oksuz et al., 2024; Alruwaili & Moulahi, 2025). En el sector público europeo, la explicabilidad figura ya como uno de los cuatro principios éticos imprescindibles para una IA "digna de confianza" como se expuso en la justificación de este TFM.

En suma, las investigaciones actuales convergen en que la explicabilidad no es un accesorio, sino un requerimiento funcional que condiciona la adopción segura de IA en ámbitos regulados.

Finalmente, se puede evidenciar que esta taxonomía permite posicionar los árboles de decisión como una de las pocas técnicas que son explicables por diseño, lo cual justifica su uso como modelo base en este trabajo, ya que no solo permiten una evaluación directa del impacto del envenenamiento de datos sobre las predicciones, sino también sobre la lógica

interna de las decisiones, lo cual no es posible en modelos opacos sin aplicar técnicas externas.

2.3. Árboles de decisión

2.3.1. Introducción a los árboles de decisión

Los árboles de decisión son algoritmos de aprendizaje supervisado ampliamente utilizados, tanto en tareas de clasificación como de regresión. Su estructura jerárquica permite representar decisiones mediante divisiones sucesivas del espacio de características y sus posibles consecuencias se observan de manera gráfica, facilitando la interpretación y comprensión de los resultados por parte de los humanos (Duda et al., 2001).

Un árbol comienza con un nodo raíz que representa todo el conjunto de datos. Luego, en cada nodo interno, se selecciona una característica y un umbral de decisión que maximiza la medida de pureza que evalúa la homogeneidad (como la entropía, la ganancia de información o el índice de Gini) para dividir los datos en subconjuntos. Este proceso se repite recursivamente hasta que se cumplen ciertos criterios de parada, como alcanzar un número mínimo de ejemplos en un nodo o una profundidad máxima del árbol. Al final del proceso recursivo, cuando este alcanza el nodo hoja, se asigna una clase (en clasificación) o un valor promedio (en regresión). La poda es una técnica utilizada para evitar el sobreajuste, eliminando ramas que aportan poca información y mejorando la generalización del modelo (Duda et al., 2001).

La naturaleza binaria de la toma de decisiones en los árboles y su representación explícita mediante reglas del tipo “si-entonces” (Russell & Norvig, 2004) convierten a estos modelos en herramientas de alta explicabilidad.

> «A key property of tree-based models, which makes them popular in fields such as medical diagnosis, for example, is that they are readily interpretable by humans because they correspond to a sequence of binary decision applied to the individual input variables.» (Bishop, 2009, p. 664).
>
> “Una propiedad clave de los modelos basados en árboles, que los hace populares en campos como el diagnóstico médico, por ejemplo, es que son fácilmente interpretables por los humanos porque corresponden a una secuencia

de decisiones binarias aplicadas a las variables de entrada individuales." (traducción propia).

No obstante, esta misma estructura explicable puede ser también una fuente de fragilidad. Estudios han demostrado que pequeñas alteraciones en los datos de entrenamiento, como las introducidas en ataques de envenenamiento de datos, pueden afectar las divisiones, alterar la estructura del árbol, degradar el accuracy del clasificador o forzar ciertas predicciones (Cinà et al., 2024; Drews et al., 2020). Esta susceptibilidad es aún más crítica en contextos donde se espera que el modelo mantenga la consistencia y justificación lógica en su comportamiento.

Por ello, aunque los arboles de decisión ofrecen ventajas claras en términos de interpretabilidad y transparencia, su uso en entornos críticos requiere una evaluación cuidadosa de su robustez frente a ataques adversariales. Este trabajo utilizará árboles como modelo base precisamente por su valor explicativo, lo que permite no solo medir la degradación de rendimiento, sino también evidenciar directamente cómo los ataques comprometen su lógica interna. Adicionalmente, su uso en áreas como la salud, la seguridad y la defensa nacional, los convierte en un caso de estudio relevante para explorar las implicaciones de los ataques de envenenamiento de datos en sistemas críticos.

2.3.2. Métricas de evaluación de los árboles de decisión

En el contexto de los árboles de decisión, se distinguen dos tipos de métricas: las métricas internas, utilizadas durante la construcción del árbol para seleccionar las divisiones óptimas en cada nodo, como el error cuadrático medio, la entropía o el índice de Gini que han sido detalladas en el Anexo A, y las métricas externas, empleadas para evaluar el rendimiento general del modelo una vez entrenado y que detallaron en la sección 2.1.1

Las métricas internas permiten guiar el crecimiento del árbol optimizando la separación de clase, de esta forma el árbol aprende seleccionando la partición que mejor separa los datos, guiado por criterios de pureza (Duda et al., 2001, p. 398) o por la reducción del error en cada partición o nodo. Por su parte, las métricas externas, como Accuracy, F1-score o el error cuadrático medio (MSE), entre otras, son fundamentales para cuantificar la capacidad de generalización del modelo, compararlo con otros enfoques y validar su desempeño en datos no vistos.

Ambos tipos de métricas son esenciales: las internas definen la estructura del modelo y las externas permiten evaluar su comportamiento global, tanto en tareas de clasificación como en aquellas de regresión. Esta distinción será especialmente relevante cuando se desarrolle la metodología de este trabajo, donde se emplearán métricas externas para analizar los efectos del envenenamiento de datos sobre el rendimiento de los modelos.

2.3.3. Variantes en árboles de decisión

Hasta ahora se ha descrito el funcionamiento general de los árboles de decisión binarios clásicos. Existen, sin embargo, múltiples variantes desarrolladas con el objetivo de mejorar su eficiencia, precisión o adaptabilidad a contextos específicos.

Una revisión ampliada de estas variantes, incluyendo los algoritmos ID3, C4.5, CART y sus desarrollos posteriores como los Optimal Classification Trees (OCT), se presenta en el Anexo B. Allí se analizan sus fundamentos, diferencias clave y ámbitos de aplicación.

En síntesis, la evolución desde los modelos pioneros hasta enfoques contemporáneos como los OCT refleja una línea de investigación orientada a maximizar la precisión sin comprometer la interpretabilidad.

2.3.4. Ventajas y desventajas de los árboles de decisión

La documentación consultada menciona multiples ejemplos de ventajas y desventajas del uso de los árboles de decisión. A continuación se compilan algunas de las encontradas en el marco de este trabajo:

Ventajas:

- **Interpretabilidad**: Son fáciles de interpretar por los humanos (Bishop, 2009)
- **Enfoque similar al humano**: La estructura de los árboles de decisión se asemeja al proceso de toma de decisiones humano, lo que facilita su comprensión y aceptación en aplicaciones prácticas.(James et al., 2013)
- **Visualización gráfica**: La representación gráfica de los árboles de decisión permite una visualización clara de las decisiones facilitando la interpretación y comunicación de los resultados a audiencias no técnicas.(James et al., 2013)

- **Multitarea**: Los árboles de decisión pueden utilizarse tanto para clasificación como para regresión. (Murphy, 2022)
- **Robustez ante outliers**: Son relativamente menos sensibles a los valores atípicos (Murphy, 2022)
- **Son rápidos de entrenar y escalan bien**: Los árboles de decisión son eficientes en términos de tiempo de entrenamiento y pueden manejar grandes conjuntos de datos sin problemas significativos de escalabilidad (Murphy, 2022).
- **Manejo de datos faltantes**: Pueden manejar valores faltantes de manera efectiva, ya sea ignorándolos o asignando probabilidades basadas en el resto de los datos (Murphy, 2022).

Desventajas:

- **Inestabilidad**: Los árboles de decisión tienen una alta varianza, lo que puede ser su mayor desventaja, ya que pequeñas variaciones en los datos de entrenamiento pueden resultar en árboles muy diferentes, lo que afecta la consistencia del modelo debido a su naturaleza jerárquica (Hastie et al., 2009; Bishop, 2009)
- **Subóptimos**: Aunque los árboles de decisión son fáciles de interpretar, pueden no ser tan óptimos como otros modelos más complejos, dado que las divisiones están alineadas a los ejes y no pueden capturar relaciones más complejas entre las características (Bishop, 2009). Por otra parte, el óptimo global de un árbol de decisión no es necesariamente el óptimo local, lo que puede llevar a soluciones subóptimas en la práctica (Duda et al., 2001).
- **Complejidad computacional**: La construcción de árboles de decisión puede ser computacionalmente costosa, especialmente en conjuntos de datos grandes o con muchas características (Duda et al., 2001).
- **Poca precisión**: Aunque son fáciles de interpretar, los árboles de decisión pueden no ser tan precisos como otros modelos más complejos (Murphy, 2022).

2.4. Envenenamiento de datos: definición y tipologías

Al revisar algunas de las investigaciones más actuales, se encuentra que una influyente fundación estadounidense en el campo de la ciberseguridad, Open Worldwide Application Security Project (OWASP), en una de sus publicaciones más recientes, presenta una clasificación de los diez riesgos de seguridad más críticos actuales relacionados con el uso de modelos de IA Generativa y LLM (OWASP, 2024).

A continuación, se enumeran estos riesgos, con su nombre original y una traducción adaptada (traducción propia):

- Inyección de instrucciones (Prompt Injection)
- Divulgación de información sensible (Sensitive information Disclosure).
- Vulnerabilidades de la cadena de suministro (Supply Chain Vulnerability)
- Envenenamiento de modelos y datos (Data and Model Poisoning)
- Manejo inadecuado de salidas (Improper output Handling)
- Autonomía excesiva (Excessive Agency)
- Filtración de instrucciones del sistema (System Prompt Leakage)
- Debilidades en vectores y embeddings (Vector and Embedding Weaknesses)
- Desinformación (Misinformation)
- Consumo descontrolado de recursos (Unbounded Consumption)

Entre todos los mencionados en esta lista de riesgos, los más relevantes a profundizar en el marco del presente trabajo son los mencionados en el cuarto item: *Envenenamiento de modelos y datos*, que buscan comprometer la integridad del modelo desde su fase de entrenamiento.

Con la finalidad de dar un marco contextual a este tipo de ataque, es importante destacar que en ámbito de la ciberseguridad y la ciberinteligencia, el envenenamiento de datos es uno de los muchos tipos de ataques que pueden enmarcarse dentro de lo que se conoce

como *ataques adversariales* (adversarial attacks). Estos ataques buscan explotar vulnerabilidades en los modelos de aprendizaje automático para manipular sus predicciones o comportamientos, y pueden clasificarse en varias categorías según su objetivo y metodología.

El NIST de Estados Unidos, nos proporciona un exhaustivo marco teórico sobre la taxonomía de los diversos tipos de ataques adversariales (Vassilev, 2025), haciendo una primera distinción entre ataques a modelos de IA predictivos y ataques a modelos de IA generativos. Como los árboles de decisión son modelos predictivos, este trabajo se centrará en los ataques adversariales que afectan a estos modelos.

De este modo, en el contexto de los modelos predictivos, se define los ataques adversariales como aquellos que buscan comprometer la integridad, confidencialidad o disponibilidad de un modelo de IA, afectando su capacidad para realizar predicciones precisas y confiables.

Otros criterios que se tienen en cuenta para clasificar los ataques adversariales son los objetivos, capacidades y conocimientos del atacante, lo que permite agrupar las variantes de los ataques para una mejor comprensión de sus características y potenciales impactos.

Finalmente, se identifica que los modelos de ML predictivos diferencian un *estado de entrenamiento* donde el modelo aprende y un segundo estado, el *estado de despliegue* también llamado de *inferencia* (Papernot et al., 2018), donde el modelo realiza predicciones sobre datos no vistos en la fase de entrenamiento.

Los *ataques por envenenamiento*, en el contexto de sistemas de IA, se refiere a ataques adversariales diseñados para interferir con el proceso de entrenamiento de un modelo de ML, introduciendo datos maliciosos en el conjunto de entrenamiento o manipulando directamente los parámetros del modelo (Vassilev, 2025).

Dependiendo del objetivo del atacante, se pueden encontrar ataques aleatorios o dirigidos, y los últimos son por lo general más efectivos y difíciles ya que tienen un objetivo específico (Lyu et al., 2020).

También es posible diferenciar dos categorías más, mencionadas en la literatura, según hacia dónde son dirigidos los ataques, a los datos o al modelo (Wan et al., 2023; Vassilev, 2025):

- **Envenenamiento de datos** (Data poisoning): el objetivo de este ataque son los

datos de entrenamiento, y corresponde a la manipulación de ejemplos dentro del conjunto de entrenamiento, para alterar el modelo y producir salidas influenciadas por el atacante. Esta es la categoría en la que se enfoca el presente trabajo.

- **Envenenamiento del modelo** (Model poisoning): el objetivo de este ataque es el modelo e implica manipular directamente los parámetros del modelo durante su entrenamiento. Se suele asociar a modelos de aprendizaje distribuido (como ocurre en federated learning), generalmente por parte de un cliente comprometido. Aunque igualmente riesgoso, este tipo de ataque se considera fuera del alcance del presente estudio.

2.4.1. Definición de envenenamiento de datos

El ataque de *envenenamiento de datos*, también conocido como *data poisoning*, es una técnica de ataque adversarial que busca comprometer la integridad de un modelo de aprendizaje automático. Su naturaleza rompe la suposición implícita de que los datos de entrenamiento son representativos de los datos de prueba reales, logrando su objetivo final que es determinar una forma efectiva de contaminar los datos de entrenamiento para forzar predicciones erróneas del modelo en el momento de la prueba. (Calzavara et al., 2025).

La metodología de este ataque implica la introducción de ejemplos, diseñados cuidadosamente, en el conjunto de datos de entrenamiento, con el fin de alterar el comportamiento del modelo final.

Estos ataques son particularmente problemáticos en modelos de alta sensibilidad a los datos de entrenamiento, como es el caso de los árboles de decisión, ya que incluso pequeños cambios pueden alterar la estructura del árbol y modificar por completo las reglas de decisión aprendidas (Bishop, 2009).

2.4.2. Clasificación del envenenamiento de datos

Existen múltiples variantes de este ataque, que pueden clasificarse según el objetivo del atacante y la forma en que se construyen los datos contaminados:

Según el objetivo del atacante, se identifican tres categorías principales (Cinà et al.,

2024):

- **Indiscriminado (Indiscriminate):** El atacante manipula una fracción del conjunto de entrenamiento para maximizar el error de clasificación del modelo.
- **Dirigido (Targeted) :** El atacante manipula de nuevo un subconjunto de los datos de entrenamiento, pero esta vez con el objetivo de provocar la clasificación errónea de un conjunto específico de muestras (limpias).
- **Puerta trasera (Backdoor):** El atacante manipula los datos de entrenamiento para que el modelo aprenda a clasificar correctamente las muestras limpias, pero al mismo tiempo introduce una serie de muestras envenenadas con un patrón específico (detonador de puerta trasera) que activan un comportamiento específico del modelo que se presenta durante la fase de inferencia.

Según la forma en que se construyen los datos contaminados, se pueden distinguir dos tipos de ataques (Lyu et al., 2020):

- **Ataques de etiqueta limpia (Clean label attacks):** El atacante no puede modificar las etiquetas de los datos, pero puede introducir ejemplos que alteren el comportamiento del modelo. Estos ataques son más difíciles de detectar, ya que los datos envenenados pueden parecer legítimos a simple vista y el rendimiento del modelo no se ve afectado .
- **Ataques de etiqueta sucia (Dirty label attacks):** El atacante puede modificar las etiquetas de los datos, introduciendo ejemplos que desea etiquetar erróneamente con una etiqueta objetivo. Estos ataques son más fáciles de detectar, ya que los datos envenenados pueden ser identificados por su inconsistencia con el resto del conjunto de entrenamiento.

Complementariamente, el trabajo de Ramirez (Ramirez et al., 2022) nos proporciona un listado de variantes de ataques de envenenamiento de datos que se ha incluido en el Anexo. No obstante, es relevante incluir en esta sección dos de los ataques incluidos en el listado ya que son representativos de las amenazas que enfrentan los árboles de decisión y son utilizados en la fase experimental de este TFM:

- **Ataque de manipulación de etiquetas (Label-Flipping Attacks):** Consiste en la alteración maliciosa de las etiquetas en los datos de entrenamiento, lo que puede realizarse de forma aleatoria o específica para reducir la precisión general o causar una clasificación errónea de una clase específica, respectivamente.

- **Ataques de Envenenamiento basados en características (Feature-Based Poisoning Attacks):** Crean muestras de entrenamiento envenenadas que son indistinguibles de las muestras originales para la inspección visual humana, preservando así la privacidad y mostrando una alta resistencia a los métodos de defensa existentes.

Esta sección permite comprender la variedad de vectores de ataque, objetivos y metodologías que pueden emplearse en el envenenamiento de datos, lo que es crucial para el desarrollo de defensas efectivas y la evaluación de la robustez de los modelos de machine learning, especialmente aquellos basados en árboles de decisión.

2.4.3. Vulnerabilidad de los árboles de decisión frente al envenenamiento de datos

Los árboles de decisión, si bien destacan por su interpretabilidad y simplicidad, también presentan una elevada sensibilidad a pequeñas modificaciones en el conjunto de entrenamiento, lo que es una desventaja de su uso como se señaló en la sección 2.3.4. Esto los convierte en objetivos especialmente vulnerables a ataques de envenenamiento de datos, dado que su estructura se basa en divisiones secuenciales guiadas por métricas de pureza, incluso unos pocos ejemplos maliciosos pueden alterar drásticamente las decisiones tomadas en los nodos superiores del árbol, propagando su efecto hacia todo el modelo.

Esta problemática ha sido abordada por diversos autores. Por ejemplo, *Antidote*, es un sistema para verificar la robustez de árboles de decisión ante este tipo de ataques. En su estudio demuestran que es posible construir ejemplos adversarios que cambian significativamente las divisiones del árbol con un número sorprendentemente bajo de instancias envenenadas(Drews et al., 2020).

Por su parte, investigaciones más actuales dan paso al estudio de ataques específicamente diseñado para árboles de decisión, como es el caso de Timber (Calzavara et al., 2025). Esta técnica utiliza un enfoque de conocimiento completo del modelo para selec-

cionar ejemplos de entrenamiento cuya modificación maximice el cambio estructural y el deterioro del rendimiento global del clasificador. Su investigación confirma que incluso modelos entrenados sobre grandes conjuntos de datos pueden ser profundamente afectados mediante perturbaciones dirigidas a puntos críticos del espacio de entrada.

Estos estudios coinciden en señalar que los árboles de decisión, son susceptibles a degradaciones significativas ante la inserción de ejemplos maliciosos. Esta vulnerabilidad no solo afecta el rendimiento, sino también la coherencia lógica y la trazabilidad de las decisiones, elementos fundamentales en aplicaciones donde la explicabilidad es un requisito.

2.4.4. Métodos de ensamble basados en árboles de decisión: su aparente robustez y vulnerabilidad ante ataques

Los métodos de ensamble constituyen una estrategia de aprendizaje supervisado para mejorar el rendimiento, la estabilidad y la generalización de los modelos. Su fundamento se basa en la combinación de múltiples clasificadores individuales, por ejemplo árboles de decisión, para formar un modelo compuesto que supere las limitaciones de sus componentes.

Desde la perspectiva de los árboles de decisión, estos métodos permiten mitigar problemas como reducir la alta varianza (Murphy, 2022), al combinar las predicciones de varios árboles entrenados sobre diferentes subconjuntos del conjunto de datos original.

Estas técnicas no solo mejoran el desempeño, sino que también proporcionan una mayor robustez frente a variaciones en los datos.

Entre los métodos de ensamble más utilizados para árboles de decisión encuentran (James et al., 2013):

- **Bagging (Bootstrap Aggregation)**: Es un procedimiento de propósito general diseñado para reducir la varianza de un método de aprendizaje estadístico. Es particularmente útil y frecuente en el contexto de los árboles de decisión. Esta técnica consiste en que en vez de entrenar el modelo una sola vez con el conjunto de datos completo, se generan múltiples subconjuntos de entrenamiento mediante muestreo con reemplazo (bootstrap sampling). Esto se logra tomando repetidas muestras del conjunto de datos de entrenamiento original, con reemplazo. Cada uno de estos

subconjuntos se utiliza para entrenar un modelo base (por ejemplo, un árbol de decisión), y luego las predicciones de todos los modelos se combinan, generalmente mediante votación mayoritaria (en clasificación) o promediado (en regresión).

- **Random Forests**: Es una mejora sobre los árboles bagged, que busca reducir la correlación entre los árboles individuales. Funciona muy similar al bagging, pero la diferencia clave es que cada vez que se considera una división en un árbol, se selecciona un subconjunto aleatorio de características en lugar de considerar todas las características disponibles. Esto introduce una mayor diversidad entre los árboles y reduce la varianza del modelo final. Los Random Forests son conocidos por su robustez y capacidad para manejar grandes conjuntos de datos con muchas características, además de ser menos propensos al sobreajuste en comparación con un único árbol de decisión.

- **Boosting**: Es otra técnica para mejorar las predicciones de los árboles de decisión. En este caso los árboles se construyen de manera secuencial, donde cada nuevo árbol se entrena para corregir los errores del árbol anterior. En lugar de entrenar todos los árboles de forma independiente, el boosting ajusta el modelo en función de los errores cometidos por los árboles anteriores. Los modelos más conocidos son *AdaBoost* y *Gradient Boosting*, que ajustan iterativamente los pesos de las instancias mal clasificadas para mejorar la precisión del modelo final. La literatura también menciona variantes como Forward Stagewise Additive Modeling (FSAM) y Gradient Boosting Machines (GBM), que se centran en minimizar una función de pérdida específica durante el entrenamiento. Logit Boosting es una variante que se utiliza para problemas de clasificación binaria(Murphy, 2022). Uno que merece mención especial es *XGBoost* (Extreme Gradient Boosting), que ha ganado popularidad por su eficiencia y rendimiento. XGBoost implementa técnicas avanzadas de regularización y optimización, lo que lo hace especialmente efectivo para conjuntos de datos grandes y complejos.

Complementariamente, el **Stacking**, que es un enfoque de ensamble que combina múltiples modelos base (que pueden ser árboles de decisión) y utiliza un modelo meta para hacer la predicción final. En este caso, los modelos base se entrenan por separado

y sus predicciones se utilizan como características de entrada para el modelo meta, que aprende a combinar las salidas de los modelos base para mejorar la precisión general (Hastie et al., 2009). Si bien este modelo no es señalado como un método de ensamble basado en árboles según la referencia señalada, es importante mencionarlo ya que puede incluir árboles de decisión como modelos base y sí lo consideran como tal otros autores (Murphy, 2022).

Finalmente, y en contexto con este TFM, se destaca que esta aparente robustez de los modelos ensamble no implica inmunidad al enfrentar ataques adversariales. Como señalan (Chang & Im, 2020), incluso modelos de tipo *Random Forest* pueden ser vulnerables a estrategias de envenenamiento de datos. El impacto del envenenamiento puede intensificarse cuando las instancias maliciosas son diseñadas con características que tienden a distribuirse de forma uniforme, afectando de manera consistente los patrones de decisión del modelo. Aunque el estudio no menciona directamente su propagación en múltiples árboles, la manipulación deliberada de atributos con baja desviación estándar sugiere un efecto acumulativo en el conjunto del Random Forest.

2.5. Aplicación de la IA en sectores críticos: Defensa, Seguridad y Salud

En esta sección se explora la aplicación de la IA en proyectos de España dentro de los sectores críticos de la defensa, de la seguridad y de la salud. Estos tres sectores, tienen en común que generan un impacto directo sobre la vida de las personas con implicaciones significativas y, en algunos casos, irreversibles.

2.5.1. IA en Defensa

No es nuevo que los avances tecnológicos, quizás más importantes para la sociedad, se producen en el ámbito militar y en un entorno bélico. En el trabajo de (Roldán Tudela, 2017) un conjunto de autores, recogen un análisis holístico de la IA aplicada, justamente al ámbito militar.

Se destaca el valor de aplicar IA, junto a otras técnicas relacionadas, como una exigencia de transformación de las capacidades militares en orden de mantener la ventaja

militar. Esta necesidad se refuerza con el creciente volumen de datos que para ser tratados requieren de técnicas de IA. El tratamiento de los datos es un factor clave en el ámbito militar, ya que la información es un recurso estratégico y su correcta interpretación puede marcar la diferencia en la toma de decisiones tácticas y estratégicas. Por otra parte, la necesidad de inteligencia en el ámbito militar sobre los medios del adversario, la anticipación de sus movimientos y la identificación de patrones de comportamiento es un factor valioso del que ningún sector militar puede prescindir, y la IA se presenta como una herramienta clave para potenciarla.

La incorporación del combatiente como un elemento clave en las funciones del combate, resalta la necesidad de mantenerlos preparados, por lo que se incentiva la incorporación de herramientas de IA para mejorar la instrucción avanzada, considerando además que la IA puede ayudar a reducir la carga cognitiva del combatiente, permitiéndole centrarse en tareas más críticas y estratégicas, pudiendo anticiparse a las decisiones del adversario.

La delegación de la toma de decisiones a sistemas automatizados, sin embargo, no es tomado a la ligera y en el mencionado trabajo, se destina un capitulo completo a la ética en el uso de la IA en el ámbito militar. En este sentido, se destaca la importancia de la trazabilidad y la explicabilidad de las decisiones automatizadas, especialmente en contextos donde las decisiones pueden tener consecuencias humanas significativas (Lewis et al., 2016). La ética militar exige que las decisiones tomadas por sistemas automatizados sean auditables y comprensibles, lo que implica que los modelos utilizados deben ser transparentes y sus decisiones justificables.

Por otra parte, el trabajo de (Alcántara Suárez, 2023) que analiza la aplicación de machine learning en sistemas de defensa, complementa esta mirada con ejemplos concretos de proyectos de defensa que aplican ML. Parece apropiado, entonces, incorporar al menos dos de estos ejemplos ya que ilustran la aplicabilidad de estas técnicas en proyectos reales del ámbito militar en España, y adicionalmente resaltar lo valioso del aporte debido a la escasa información que está disponible en español sobre estos temas. Ambos proyectos se incluyen en el anexo F.

2.5.2. IA en Seguridad

En el ámbito de la seguridad, la IA se ha convertido en una herramientas útil para la detección de amenazas, la prevención del delito y la mejora de la seguridad pública, en muchos países.

En España, se puede encontrar la publicación realizada en la XVIII Conferencia de la Asociación Española de Inteligencia Artificial, donde se presenta un trabajo que analiza el uso de técnicas de ML con algoritmos como los árboles de decisión M5P, M5Rules, Regresión lineal, k-NN, Random Forest y varios más en la predicción del delito (Cuesta Calvo et al., 2018). Se destaca el uso de variantes del algoritmo árboles de decisión ya que refuerza el enfoque de este trabajo.

El trabajo de (Ocaña, 2024) presenta una visión de investigador independiente sobre la contribución de la IA en la seguridad, destacando las implicaciones jurídicas del uso de la IA para la prevención y disminución de los riesgos de seguridad. Se menciona el desarrollo de aplicaciones de IA para la predicción del delito, los cuales tienen vías multimodales de ejecución y por tanto, el análisis también requiere una perspectiva multimodal, que permita integrar datos de diferentes fuentes, como imágenes, texto y audio.

Por otra parte, (Monforte, 2023) nos propone una clasificación de tres categorías, para agrupar las herramientas que usan IA y que podrían ser aplicadas para investigar delitos; las Herramientas de predicción y evaluación de riesgos, las Herramientas de investigación de delitos y las Herramientas de tramitación.

Presentando técnicas utilizadas como reconocimiento facial y de voz para verificar identidad, reconocimiento de emociones para detectar veracidad en las declaraciones, procesamiento de lenguaje natural (PLN) para análisis documental, reconocimiento de imágenes para identificar objetos y personas en contextos de investigación, etc.

El trabajo de (Cuenca & Medina, 2023) proporciona ejemplos de proyectos implementados en España y que se han realizado para seguridad de la población. De ellos, dos proyectos se incluyen en el anexo F.

2.5.3. IA en Salud

La relación entre la ciencias de la computación y el ámbito de la salud tiene una larga trayectoria que dio lugar al surgimiento de términos específicos como la informática médica, la bioinformática, procesamiento electrónico de datos médicos, etc. Hoy en día, la IA, está apoyando a los médicos en tareas que son esenciales y limitadas, dejando la responsabilidad de manejar a los pacientes a los médicos humanos. No obstante, se valora su apoyo en la optimización de los procesos de prevención, diagnóstico y tratamiento de enfermedades (Molina, 2024)

El Reglamento Europeo (Unión Europea, 2025) que regula el espacio europeo de datos de salud, busca establecer un marco normativo que garantice la protección de los datos personales de salud y promueva su uso para fines de investigación y mejora de la atención médica. Esta regulación complementa el Reglamento General de Protección de Datos (RGPD), proporcionando un marco específico para el tratamiento de datos de salud en la Unión Europea y en él, se regula también el uso de la IA en el ámbito sanitario, estableciendo requisitos específicos para garantizar la seguridad, la privacidad y la ética en el tratamiento de estos datos.

Como ejemplos de proyectos de IA en el ámbito de la salud, desarrollados en España, se incluyen dos en el anexo F.

2.6. Técnicas de defensa ante ataques y escenarios de riesgo

Uno de los principales desafíos en materia de seguridad, ya sea desde un enfoque técnico o estratégico, es la ausencia de soluciones universales capaces de garantizar una protección total de los sistemas. En consecuencia, se requiere adoptar medidas de seguridad en capas, que aborden en profundidad distintos vectores de ataque y vulnerabilidades. Esta lógica aplica también a los escenarios adversariales del aprendizaje automático, donde deben considerarse tanto las características del modelo como el contexto de aplicación y los posibles objetivos del atacante.

Este trabajo no pretende abordar de forma exhaustiva todas las técnicas defensivas existentes, sino ofrecer una visión general de las estrategias recogidas en la literatura, especialmente aquellas relacionadas con los ataques de envenenamiento de datos en mo-

delos predictivos explicables, como los árboles de decisión.

2.6.1. Estrategias y modelos de amenazas

Siguiendo la estructura planteada por (Vassilev, 2025), una primera recomendación estratégica es modelar al adversario y proyectar diferentes escenarios de ataque. Esta anticipación resulta clave para implementar medidas defensivas acordes al tipo de amenaza. En este sentido, el trabajo de (Muñoz-González et al., 2017) propone un marco sistemático para entender los ataques de envenenamiento de datos en modelos de aprendizaje profundo, que puede adaptarse a otros algoritmos.

Dicho marco considera tres dimensiones principales:

- **Objetivo del ataque:** ¿Busca degradar el rendimiento global del modelo o alterar específicamente una clase?
- **Nivel de conocimiento del atacante:** ¿Posee acceso completo al modelo, o su conocimiento es parcial o nulo?
- **Capacidad para manipular los datos:** ¿Puede alterar el conjunto de entrenamiento, o solo intervenir en la fase de inferencia?

De forma complementaria, (Biggio & Roli, 2018) introduce tres "reglas de oro" aplicables a la narrativa de seguridad en el aprendizaje automático:

- **Conoce a tu adversario (Know your adversary):** modelar las amenazas contra el sistema que se diseña.
- **Sé proactivo (Be proactive):** simular ataques y diseñar contramedidas antes del despliegue.
- **Protégete (Protect yourself):** implementar medidas de defensa tanto reactivas como proactivas.

Esta visión estratégica es crucial, dado que muchas técnicas defensivas actuales presentan limitaciones importantes: algunas reducen la precisión del modelo, otras no son aplicables a todos los algoritmos, y algunas son vulnerables a ataques más sofisticados, como los llamados *clean-label*.

2.6.2. Escenario de riesgo en modelos explicables

Un aspecto particular en el caso de los modelos explicables, como los árboles de decisión, es que su lógica interna puede ser fácilmente inferida mediante herramientas de XAI, facilitando así la ingeniería inversa de su comportamiento lo que representa un riesgo. Esta facilidad para inferir su lógica interna es una vulnerabilidad que es explotada por el sistema **AUTOLYCUSE** (Oksuz et al., 2024), que emplea explicaciones generadas por el propio modelo para reconstruir su estructura incluso con acceso limitado, lo cual facilita la generación de ataques dirigidos.

Por tanto, los modelos explicables enfrentan amenazas particulares que, aunque aumentan la transparencia, también pueden incrementar la superficie de ataque.

2.6.3. Técnicas de defensa documentadas

Frente a los riesgos mencionados, diversos estudios han propuesto mecanismos de defensa para mitigar el impacto de los ataques de envenenamiento en modelos de aprendizaje automático. Estas estrategias pueden clasificarse en tres grandes enfoques (Ramirez et al., 2022; Drews et al., 2020):

- **Filtrado previo al entrenamiento:** aplicación de técnicas estadísticas o de aprendizaje no supervisado para detectar y eliminar instancias sospechosas o inconsistentes.
- **Modificación del proceso de aprendizaje:** adaptación de algoritmos para limitar la influencia de datos individuales (por ejemplo, mediante regularización adversarial o modificación de criterios de partición en árboles).
- **Auditoría posterior al entrenamiento:** análisis del comportamiento del modelo para detectar reglas de decisión inesperadas o ejemplos con influencia desproporcionada.

En cuanto a herramientas específicas, se destaca **Antidote** (Drews et al., 2020), diseñada para evaluar y mejorar la robustez de modelos frente al envenenamiento, en particular en árboles de decisión.

De forma inversa, la investigación de (Calzavara et al., 2025) describe el ataque **Timber**, un método de envenenamiento de tipo caja blanca que aplica *label-flipping* para deteriorar el rendimiento de clasificadores basados en árboles. Aunque no existen aún defensas específicamente robustas para árboles de decisión, se han explorado mecanismos como saneamientos basados en *k-Nearest Neighbors* (kNN) y estrategias como Baggins, que si bien no eliminan el ataque, permiten reducir su efecto.

Este último trabajo es especialmente relevante para el presente TFM, ya que demuestra tanto las vulnerabilidades específicas de los árboles de decisión como la necesidad urgente de diseñar defensas más adecuadas, más allá de las soluciones agnósticas al modelo actualmente disponibles.

La revisión realizada presenta un espectro relativamente acotado de técnicas de defensa ante ataques adversariales en general, y de envenenamiento de datos en particular. Estas estrategias abarcan diferentes enfoques como la detección y filtrado de instancias sospechosas usando diversas técnicas de sanitización de datos, la modificación del proceso de aprendizaje con técnicas como regularización, métodos de agregación y ensembled, datos aumentados, privacidad diferenciada y entrenamiento adversarial, todas ellas para incrementar la robustez del entrenamiento del modelo (Carnerero-Cano, 2023). Sin embargo, también se identifican limitaciones importantes, como su escasa efectividad frente a ataques sofisticados, su reducida capacidad de generalización a distintos tipos de modelos y, especialmente, su enfoque reactivo más que preventivo.

Este enfoque toma especial importancia además dada las investigaciones que demuestran el concepto de que los ejemplos de entrenamiento adversarial pueden transferirse entre algoritmos, lo que implica que un ataque diseñado para un modelo específico podría ser efectivo contra otros modelos, incluso si estos no comparten la misma arquitectura o metodología de entrenamiento (Muñoz-González et al., 2017). Esto sugiere que las defensas deben ser diseñadas con una perspectiva más amplia, considerando no solo el modelo específico en cuestión, sino también el ecosistema más amplio de modelos y técnicas de aprendizaje automático.

En este contexto, se observa que la mayoría de las propuestas revisadas en este trabajo se centran en detectar y mitigar los efectos del envenenamiento una vez que el ataque ha ocurrido, dejando relativamente inexplorado el desarrollo de medidas preventivas que

dificulten el uso de estos datos en el entrenamiento. Esta observación abre, por una parte, una oportunidad para que el presente TFM, sin ser ese su foco principal, pueda aportar algunas reflexiones iniciales sobre posibles recomendaciones orientadas a la prevención, las cuales serán abordadas en las conclusiones. Y, por otra parte, refuerza la necesidad de continuar investigando en este campo emergente, dinámico y de alto impacto, que aún presenta importantes desafíos abiertos.

2.7. Cierre del marco teórico

La literatura reciente sobre ataques a modelos y riesgos derivados del uso de XAI se ha centrado en escenarios con redes neuronales y explicaciones ricas en gradientes, o en marcos de extracción de modelo que explotan LIME/SHAP para aproximar fronteras de decisión (p. ej., AUTOLYCUS). Aunque estos trabajos avanzan la comprensión de la superficie de ataque, persiste una brecha aplicada: hay poca evidencia en modelos explicables desplegados en sectores críticos con datos públicos españoles y análisis fino del cambio estructural del clasificador (raíces, reglas, profundidad) bajo envenenamiento leve. Este TFM contribuye en esa intersección: (i) operacionaliza ataques simples pero reproducibles sobre árboles de decisión, (ii) evalúa el deterioro métrico y estructural, y (iii) contextualiza su impacto en salud, seguridad y defensa, sentando bases empíricas para trabajos futuros con ensembles y ataques más sofisticados.

3. Metodología

Esta investigación adopta un enfoque metodológico fundamentado en los supuestos filosóficos que subyacen a toda actividad investigativa, específicamente en términos de ontología (la naturaleza de la realidad) y epistemología (cómo puede conocerse dicha realidad). Siguiendo la clasificación propuesta por Oates (2006), se reconocen tres paradigmas principales en el ámbito de los sistemas de información y la computación: el positivista, el interpretativo y el crítico.

El paradigma positivista, adoptado en este trabajo, asume que la realidad es objetiva, única y medible. Desde esta perspectiva, el investigador actúa como un observador independiente, cuyo propósito es descubrir relaciones causales y leyes generales a través de métodos cuantitativos rigurosos. Este paradigma es consistente con la estrategia seleccionada para la investigación: el experimento, una técnica típicamente positivista que permite manipular variables independientes en entornos controlados para evaluar su efecto sobre variables dependientes (Oates, 2006). En este sentido, el paradigma interpretativo, orientado a la subjetividad y la comprensión, se presenta como una vía más idónea para explorar significados y experiencias sociales, mientras que el paradigma crítico, enfocado en la emancipación y la transformación, se centra en cuestionar las estructuras de poder y en impulsar cambios sociales.

En este trabajo, el marco de referencia de CRISP-DM, descrito en la sección 1.1.3, se emplea como guía operativa para la gestión y desarrollo del TFM, mientras que el enfoque positivista de Oates orienta el diseño experimental. De esta forma, las fases de comprensión, preparación, modelado y evaluación propias de CRISP-DM se alinean con el objetivo experimental de establecer relaciones causales entre el tipo de envenenamiento de datos y el desempeño de los modelos de aprendizaje automático. Así, CRISP-DM proporciona la estructura secuencial y trazable del proceso, y el enfoque experimental positivista garantiza el rigor científico en la contrastación de hipótesis.

En particular, se diseñaron dos escenarios experimentales. El primer escenario se centró en un dataset de criminalidad por comunidad autónoma en España, incluyendo un baseline y cuatro experimentos de envenenamiento de datos. En todos los casos, se aplicaron modelos de regresión basados en árboles de decisión para evaluar el impacto de

las alteraciones. El segundo escenario trabajó con un dataset del ámbito de la salud relacionado con la COVID-19, con un baseline y tres experimentos de envenenamiento de datos, aplicando modelos de clasificación también basados en árboles de decisión. Esta configuración experimental permite establecer relaciones causales entre el tipo de envenenamiento de datos y el desempeño de los modelos de aprendizaje automático, en línea con el enfoque positivista adoptado.

En síntesis, la metodología combina dos niveles complementarios: el nivel filosófico-epistemológico, sustentado en el paradigma positivista y en el uso del experimento como estrategia investigativa; y el nivel práctico-operativo, sustentado en CRISP-DM como marco de gestión y desarrollo. De forma transversal, ambos niveles convergen en el cumplimiento de los objetivos específicos del TFM, garantizando la coherencia entre el diseño metodológico, las tareas realizadas y los resultados esperados.

3.1. Objetivos y tareas

Esta sección define las tareas que guían el diseño experimental del presente trabajo, según la metodología descrita previamente. Para cada objetivo, descritos en la sección 1.3, se detallan las tareas principales, los indicadores de éxito y los entregables esperados.

Objetivo: a) Vulnerabilidad asociada a la explicabilidad

- **Tareas:** Analizar los puntos críticos de la estructura interna del modelo, mediante la identificación de los nodos más relevantes y las características que influyen en las decisiones del modelo, así como problemas para que pueda tener el modelo para generalizar, o cómo le afectan los sesgos en los datos. Esto permite determinar las vulnerabilidades de los modelos entrenados con datos no alterados.

- **Indicadores:** Variación en las características del árbol entre réplicas (predicciones, reglas de decisión, estructura en ramas y nodos) dentro de un umbral definido; estabilidad de métricas de rendimiento (Accuracy/F1 en clasificación, MAE/MSE en regresión) durante la validación.

- **Entregables:** Figuras de los árboles de referencia; cuadros con indicadores de estabilidad de la estructura y métricas.

Objetivo: b) Modos de envenenamiento de datos

- **Tareas:** Implementar las técnicas de envenenamiento de datos *feature-based* y *label-flipping*, sobre los datos usados para entrenar los modelos de árboles de decisión.
- **Indicadores:** Reproducibilidad de los ataques (scripts parametrizados); verificación de que la tasa de envenenamiento aplicada coincide con la nominal (1 %, 3 %, 5 %).
- **Entregables:** Scripts, bloques o notebooks de generación de ataques; cuadros comparativos con las tasas efectivas de envenenamiento aplicadas a cada dataset.

Objetivo: c) Deterioro por envenenamiento de datos

- **Tareas:** Entrenar modelos de árboles de decisión sobre datasets envenenados con diferentes parámetros de ataque; registrar cambios en la estructura del modelo y en su rendimiento general; comparar métricas entre modelos con y sin datos envenenados.
- **Indicadores:** Cambios en las métricas entre modelos (ej. Accuracy/Recall en clasificación; MAE/MSE/R^2 en regresión); cambios estructurales significativos en el árbol (modificación en el número de nodos o en las variables seleccionadas en la raíz y niveles superiores; alteración de la ruta de decisión principal).
- **Entregables:** Cuadros comparativos por tasa de ataque; figuras de árboles antes y después del envenenamiento.

Objetivo: d) Riesgos en contextos críticos

- **Tareas:** Implementar *y contextualizar* experimentos de clasificación y regresión que permitan evaluar el comportamiento de los modelos en sectores críticos como salud y seguridad/defensa, considerando de manera independiente las particularidades de cada dominio.
- **Indicadores:** Evidencia de impacto del envenenamiento dentro de cada contexto (p.ej., deterioro de métricas en el escenario de salud o en el de criminalidad); consistencia y replicabilidad de los resultados obtenidos en cada dominio.

- **Entregables:** Informe de resultados por dominio; cuadros y gráficas de desempeño diferenciadas para cada dataset. Subsección en conclusiones: **Evaluación de los riesgos en contextos críticos**.

Objetivo: e) Recomendaciones para prevenir y mitigar riesgos

- **Tareas:** Concluir medidas orientadas a la prevención y mitigación del impacto del envenenamiento de datos en modelos de árboles de decisión.
- **Indicadores:** Generación de recomendaciones basadas en la evidencia empírica de los experimentos; coherencia con la literatura revisada.
- **Entregables:** Subsección de conclusiones: **Recomendaciones de seguridad** con propuestas de medidas preventivas.

3.2. Diseño experimental

Como se estableció en la metodología, este trabajo adopta un **diseño experimental de tipo cuantitativo y comparativo**, con un enfoque de laboratorio. Este diseño se articula en las fases de modelado y evaluación de CRISP-DM, bajo el enfoque positivista descrito por Oates, que sustenta la elección del experimento como estrategia investigativa.

El experimento se basa en la manipulación de variables independientes (introducción de envenenamiento en los conjuntos de datos de entrenamiento) y la observación de los efectos sobre variables dependientes, medidos a través de métricas objetivas y cuantitativas. Los indicadores de éxito incluyen tanto la variación en métricas de rendimiento de uso extendido en modelos supervisados (Accuracy, Recall, F1-score, MAE, MSE, ROC, AP) como los cambios observables en la estructura interna de los árboles de decisión (profundidad, número de nodos, rutas de decisión, estabilidad de reglas).

De esta manera, se diseñan dos experimentos principales: uno de regresión (predicción de la cantidad de delitos) y otro de clasificación (predicción de hospitalizaciones). Ambos experimentos responden a los objetivos de identificar vulnerabilidades, evaluar el deterioro inducido por el envenenamiento y comparar los riesgos en diferentes contextos críticos.

Su estructura incluye los siguientes ejes:

- **Escenario de envenenamiento:** Se simulan los objetivos del atacante como:
  - Ataques dirigidos a clases específicas.
- **Proporciones de datos envenenados:** Se implementan distintos niveles de manipulación: 1 %, 3 % y 5 % de registros envenenados sobre el total del dataset.
- **Magnitud de la perturbación del ataque:** Se aplica una intensidad de alteración sobre los datos que se define en función de la lógica del ataque y el contexto del dataset.
- **Tipos de ataque aplicados:**
  - *Según la forma de construcción de los datos:*
    - **Ataques de etiqueta limpia (Clean label attacks):** Se introducen ejemplos adversarios sin modificar las etiquetas.
    - **Ataques de etiqueta sucia (Dirty label attacks):** Se introducen ejemplos adversarios que alteran las etiquetas originales para inducir errores de clasificación.
  - *Según el tipo de manipulación:*
    - **Label-flipping:** Intercambio de etiquetas en un porcentaje definido del dataset.
    - **Feature-based:** Modificación adversaria de atributos clave que afectan la estructura del árbol.
- **Medios para generar los datos envenenados:** Se utilizan implementaciones manuales propias, asistidas por la herramienta *ChatGPT*, para la generación de ejemplos adversarios.

Los ataques implementados han sido seleccionados por ser técnicamente aplicables y efectivamente observables en modelos de árboles de decisión, permitiendo analizar su impacto tanto en rendimiento como en estructura. En contraste, se han excluido métodos que

requieren gradientes diferenciables, retro propagación o estructuras densamente conectadas, características propias de modelos como redes neuronales profundas o Transformers, pero ausentes en los árboles de decisión.

Por otra parte el rango de proporciones de datos envenenados escogidos de (1 %, 3 % y 5 %) son considerados como realista, difícil de detectar y bajo pero con potencial de generar un alto impacto en el rendimiento con base a tres referencias de la literatura (Calzavara et al., 2025; Cinà et al., 2024; Chang & Im, 2020). Estas proporciones incorporadas a este TFM, nos otorga un set de opciones para evaluar el impacto del envenenamiento desde niveles realistas, que podrían pasar desapercibidos y serían menos exigentes de cumplir para el atacante. Además, estas proporciones permiten iteraciones rápidas y pruebas reproducibles, facilitando la experimentación y comparación de resultados.

El flujo metodológico esquemático de estos ataques se puede ver en la Figura 1.

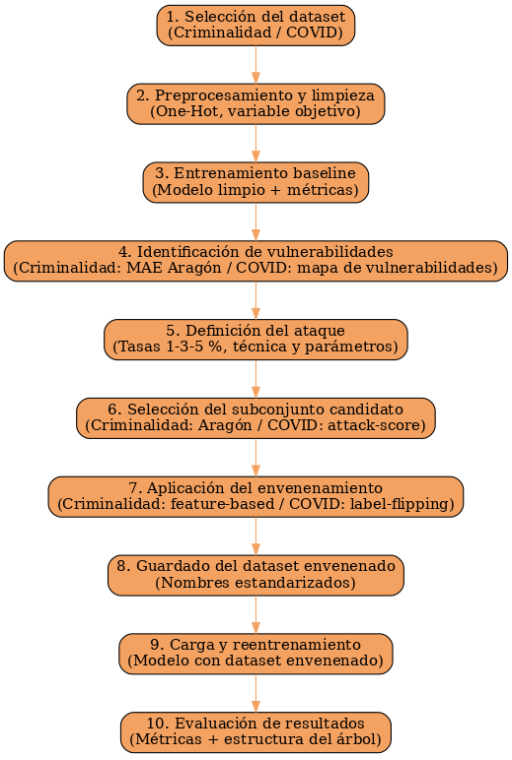


Figura 1: Flujo metodológico del procedimiento de envenenamiento de datos aplicado en los experimentos

3.3. Datasets

La experimentación se realiza sobre datasets estructurados, que proporcionan datos que pueden ser utilizados para tomar decisiones en sectores críticos (salud y seguridad) y que permiten dar contexto a los experimentos para la implementación de los modelos de árboles de decisión.

Criterios de selección:

- Disponibilidad pública o privada de los datasets.
- Licencia abierta, de uso académico o permitido su uso para este TFM.
- Relevancia para sectores críticos (salud y seguridad).
- Tamaño que permite iteraciones rápidas y pruebas reproducibles.
- Compatibilidad con tareas de clasificación o regresión.

Criterios de exclusión:

- Datasets que requerían preprocesamiento intensivo no relevante para el análisis.
- Datasets sensibles, sin anonimización o que contenían información personal identificable.

Fuente inicial de los datasets:

- Portal Estadístico de Criminalidad de España (https://estadisticasdecriminalidad.ses.mir.es/).
- Portal de Datos Abiertos del Gobierno de España (https://datos.gob.es/).
- Portal Oficial de Datos Europeos (https://data.europa.eu/en).
- Plataforma de contratación del Estado (https://contrataciondelestado.es/).
- Plataforma gestión de datos de contratación pública (https://contratos.gobierto.es/)
- Centro Nacional de Epidemiología del Instituto de Salud Carlos III (https://cnecovid.isciii.es/).

Dataset utilizados:

A continuación se presentan los dataset seleccionados para los experimentos, separados por sector al que representan, junto con una breve descripción de su origen y características principales:

- **Seguridad:** El dataset original contiene datos estadísticos de criminalidad obtenidos del Portal Estadístico de Criminalidad del Ministerio del Interior de España. Los datos son provenientes de la Policía Nacional, Guardia Civil, policías autonómicas y policías locales que proporcionan datos al Sistema Estadístico de Criminalidad. No se incluyen datos de los Mossos d'Esquadra de robos con fuerza o con violencia/intimidación en establecimientos, ni de robos con violencia/intimidación en vía pública hasta el año 2019. A partir del año 2020 se encuentran incluidos. No se incluyen datos de los Mossos d'Esquadra ni Ertzaintza de estafas informáticas hasta 2014, no obstante, a partir de 2015 se encuentran incluidos. La adaptación de este dataset es denominada "dataset de criminalidad", para referenciarlo en este trabajo.

- **Salud:** El dataset original contiene la declaración de los casos de COVID-19 a la Red Nacional de Vigilancia Epidemiológica (RENAVE) a través de la plataforma informática vía Web SiViES (Sistema de Vigilancia de España) que gestiona el Centro Nacional de Epidemiología (CNE). Esta información procede de la encuesta epidemiológica de caso que cada Comunidad Autónoma cumplimentaba ante la identificación de un caso de COVID-19. La adaptación de este dataset es denominada "dataset de covid19", para referenciarlo en este trabajo.

Los dataset seleccionados fueron adaptados durante el presente trabajo con el fin de simular los conjuntos de datos para los entrenamientos aplicados a tareas de regresión y clasificación en contextos de los sectores de salud y seguridad. Adicionalmente, se crearon copias con los datos envenenados en un entorno controlado.

Links de descarga datasets originales:
- Seguridad: Dataset de criminalidad por comunidad
- Salud: Dataset situación y evolución de la pandemia de COVID-19 en España

3.4. Recursos e Instrumentos

Se utilizan los siguientes recursos:

- Lenguaje: Python 3.12.3
- Librerías: Scikit-learn, NumPy, Pandas, Matplotlib, Seaborn, entre otras.
- Modelos utilizados: árbol de decisión (DecisionTreeClassifier y DecisionTreeRegressor de la librería Scikit-learn).
- Aplicaciones: ChatGPT modelos GPT-4o y GPT-5.
- Entorno: Jupyter Notebook y Visual Studio Code.
- Métricas: Para clasificación Accuracy, Precision, Recall, F1-score, Matriz de confusión, ROC; y para regresión RMSE, MAE, R^2 score.
- Equipo con tarjeta gráfica: Se utiliza una GPU NVIDIA GeForce RTX 4070 para acelerar el entrenamiento de los modelos.

3.5. Procedimiento

Los pasos que se siguieron en el diseño experimental se repitieron por los sectores críticos (seguridad y salud) y se detallan a continuación:

- Preprocesamiento de los datos: limpieza y transformación de variables según fue necesario.
- Análisis exploratorio de los datos: visualización de distribuciones, correlaciones y patrones relevantes.
- Selección de variable objetivo y predictoras.
- Entrenamiento de un modelo base sin envenenamiento.
- Obtención de las métricas de rendimiento del modelo base.
- Análisis de vulnerabilidades del modelo.

- Diseño del ataque por envenenamiento: selección del tipo de ataque y parámetros del mismo.
- Implementación de un primer ataque de envenenamiento (p. ej. label-flipping del 1 % de los datos con magnitud 3).
- Guardado de los datos envenenados.
- Entrenamiento del modelo sobre el dataset envenenado.
- Comparación de métricas de rendimiento y estructura.
- Visualización de los árboles generados.
- Repetición del experimento con diferentes proporciones de datos envenenados.
- Análisis de los resultados obtenidos, comparando los modelos entrenados con y sin envenenamiento.
- Documentación de los hallazgos y conclusiones.

3.6. Análisis de datos

3.6.1. Dataset de criminalidad

El dataset *criminalidad*, como se menciona en la sección 3.3, corresponde a una adaptación del dataset original para efectos de este trabajo. Este contiene un conjunto de datos agregados de criminalidad en España, donde cada registro recoge la cantidad de delitos registrados y agrupados por comunidad autónoma, categoría general de delito y por año. No incluye información individualizada, sino totales agregados para cada combinación de variables.

3.6.1.a Procesado y preprocesamiento de datos

Dada la naturaleza del dataset, datos agregados, y a que contenía cálculos de subtotales y totales, fue necesario realizar una limpieza de valores irrelevantes, detectar valores faltantes y mantener la consistencia de los datos.

Como parte del procesado y preprocesamiento de los datos se realizaron las siguientes actividades:

- Eliminación de registros con subtotales y totales generales.
- Eliminación de subcategorías de delitos, manteniendo solo las 12 categorías generales para mantener la simplicidad en el análisis.
- Reorganización de los datos en formato largo, para facilitar el análisis temporal y el uso de herramientas estadísticas.
- Verificación de valores nulos (sin imputar valores porque no se encontraron nulos).
- Codificación one-hot de las variables categóricas (*Comunidad*, *Categoría* y *Año*) para su uso en el entrenamiento.
- División del dataset en 80 % entrenamiento y 20 % validación, con semilla fija para reproducibilidad (`random_state=42`).

Cabe señalar que al aplicar one-hot encoding (One-Hot Encoded ó OHE), sobre las variables categóricas, se crean columnas dummy (término usado en estadística) para cada valor único de las variables categóricas. Durante el desarrollo de este TFM se les nombrará como columnas "dummies" (del término variables dummy), "características codificadas" o "features codificadas".

Finalmente, no se aplicó normalización ni escalado, ya que los árboles de decisión no requieren estas transformaciones para su funcionamiento, priorizando así la simplicidad y reproducibilidad del experimento. Bajo este mismo enfoque de priorizar la reproducibilidad y aislar la procedencia de los cambios durante la evaluación del deterioro del modelo entrenado con datos sin y con envenenamiento, se abstuvo de usar técnicas de validación cruzada.

3.6.1.b Análisis exploratorio de los datos

A continuación se describe el Análisis exploratorio de datos o EDA (por sus siglas en inglés, Exploratory Data Analysis), para datos agregados, realizado sobre el dataset de criminalidad:

- Revisión de la estructura del dataset, identificando cantidad de registros, columnas y tipos de datos de las variables.
- Búsqueda de valores nulos.
- Visualización de la distribución de delitos agrupados por año, comunidad autónoma y categoría del delito, con uso de gráficas de barras para identificar tendencias y patrones.
- El análisis de correlaciones entre variables numéricas fue descartado ya que el dataset solo contiene una variable numérica (Cantidad de delitos).

3.6.1.c Configuración del modelo y parámetros de entrenamiento

Para el experimento de regresión, que busca la predicción de la cantidad de delitos, se empleó un modelo de árbol de regresión (`DecisionTreeRegressor`) de la librería *scikit-learn*.

El modelo fue configurado con los siguientes parámetros:

Parámetro	**Valor**	**Descripción**
criterion	squared_error	Mide la calidad de una división basándose en la minimización del error cuadrático medio (MSE).
splitter	best	Selecciona automáticamente la mejor división posible para cada nodo.
max_depth	4	Limita la profundidad del árbol a 4 niveles para evitar sobreajuste y facilitar la interpretación.
random_state	42	Fija una semilla para garantizar la reproducibilidad.
Otros	Por defecto	`max_features`, `max_leaf_nodes`, `min_samples_split`, `min_samples_leaf` y `ccp_alpha` mantienen valores por defecto.

Cuadro 2: Parámetros de configuración del DecisionTreeRegressor

Se utilizó como variable objetivo, la variable *Cantidad* (Cantidad de delitos), y como variables predictoras: *Comunidad*, *Categoría* y *Año*. Se mantuvieron constantes los pará-

metros y se evitó introducir técnicas o configuraciones aleatorias (random) para asegurar que las diferencias observadas en el rendimiento del modelo se deban específicamente al envenenamiento de datos.

3.6.1.d Análisis de vulnerabilidad y diseño del ataque

Para llevar a cabo el análisis de vulnerabilidades del modelo, se implementaron dos enfoques: el primero evaluar la importancia de las características para comprender qué dimensiones dominan el modelo (qué sostiene el modelo). El segundo enfoque evalúa los valores de error por característica para identificar aquellas categorías donde el modelo tenía mayor dificultad para predecir (puntos débiles).

A partir de este análisis, se procede a diseñar un ataque de envenenamiento de datos dirigido a las clases específicas *Comunidad*, utilizando la técnica *feature-based*.

Como parte del diseño del ataque, se selecciona como objetivo de ataque la característica codificada *Comunidad_Aragón*, que presenta el mayor error (MAE).

El ataque se diseña estableciendo una proporción, respecto a la cantidad total de registros, como cantidad de datos a envenenar. En este experimento, se seleccionaron tres proporciones: 1 %, 3 % y 5 %.

El diseño del ataque, permitió definir los requisitos para construir un subconjunto de registros "candidato" a ser envenenado. Se estableció que los registros que lo componen debían pertenecer a la comunidad autónoma objetivo del ataque, *Comunidad_Aragón*, y debían estar distribuidos dentro de los años 2021, 2022 y 2023 del dataset, por ser los años más actualizados. Además, se decide mantenerlos concentrados en solo 3 años para evitar diluir el efecto del ataque.

Para efectos de la experimentación, cuando el subconjunto "candidato" tuvo suficientes registros para cubrir la proporción de datos a envenenar, se le mencionó como *con reemplazo o cr*, es decir, los datos envenenados reemplazan los datos originales. De lo contrario, se duplican registros seleccionados para completar la proporción deseada aumentando la cantidad de registros totales del dataset.

Consecuente con lo anterior, este trabajo cubre solo un ejemplo de envenenamiento del 1 % del total de registros del dataset con reemplazo. Los experimentos restantes, se realizaron con tasas de envenenamiento del 1 %, 3 % y 5 % sin reemplazo de los regis-

tros, permitiendo que un mismo registro pudiera ser seleccionado más de una vez para ser envenenado o se agregaran registros duplicados, aumentando así el tamaño total del dataset.

Finalmente, el valor de la característica *Cantidad* de los registros correspondientes fueron multiplicados por un factor de perturbación de 3. Esta magnitud fue seleccionada para mantener los valores envenenados dentro del rango empírico, sin superar el umbral $\mu + 2\sigma$, evitando así generar outliers fácilmente detectables, como se dejó registrado en el cuadro 5.

Nota: El concepto **con reemplazo** utilizado en el contexto de este trabajo está considerado desde una perspectiva descriptiva del experimento y no debe confundirse con el concepto de reemplazo del método `sample` en python y su configuración `replace=True`, que es el concepto opuesto al utilizado en este trabajo.

3.6.2. Dataset de covid19

Tal como mencionamos en la sección 3.3, el *dataset de covid19*, corresponde a una adaptación del dataset original para efectos de este trabajo. En él se encuentra un conjunto de datos agregados de vigilancia epidemiológica, en el que cada registro corresponde a un conteo de casos, de hospitalizaciones, de ingresos en UCI y de defunciones agrupadas por provincia, sexo, grupo de edad, año, mes y semana epidemiológica, de modo que no contiene observaciones individuales, sino totales consolidados.

3.6.2.a Procesado y preprocesamiento de datos

Dada la naturaleza del dataset, datos agregados, y a que contenía datos irrelevantes para el experimento, fue necesario realizar una limpieza para eliminar registros irrelevantes, tratar los valores nulos, adaptar la estructura del dataset y mantener la consistencia de los datos.

Como parte del procesado y preprocesamiento de los datos, se realizaron las siguientes actividades:

- Eliminación de registros, sin eventos confirmados, para eliminar datos irrelevantes para la predicción y obtener clases más balanceadas.

- Eliminación de la variable fecha, previa segmentación en variables de mes, año y semana.
- Eliminación de la columna *num_hosp*.
- Adición de la columna *hospitalizado* de tipo binaria (0, 1).
- Verificación de valores faltantes y aplicación de imputación (valor "Desconocido").
- Codificación one-hot de las variables categóricas (*provincia*, *sexo*, *grupo_edad*) para su uso en el entrenamiento.
- División del dataset en 80 % entrenamiento y 20 % validación, con semilla fija para reproducibilidad (`random_state=42`).

Al igual que en el experimento de regresión, en este experimento de clasificación tampoco se aplicó normalización, escalado, técnicas de validación cruzada, ni configuraciones aleatorias para priorizar simplicidad y reproducibilidad del experimento.

3.6.2.b Análisis exploratorio de los datos

A continuación se describe el EDA realizado sobre el dataset de covid19.

Se realizó un análisis estadístico descriptivo para entender la distribución de los datos y la relación de las variables predictivas con la variable objetivo. Se generaron visualizaciones para identificar patrones y tendencias:

- Revisión de la estructura del dataset, identificando cantidad de registros, columnas y tipos de datos de las variables.
- Búsqueda de valores nulos con gráfica para visualizar su proporción por características.
- Visualización de la distribución de los datos de la clase objetivo, mediante gráfico de barras y de pastel.
- Visualización de la correlación entre variables, utilizando un mapa de calor.

3.6.2.c Configuración del modelo y parámetros de entrenamiento

Para el experimento de clasificación, que busca la predicción de hospitalizaciones frente a casos confirmados de COVID-19, se empleó el modelo de árbol de decisión para clasificación o modelo de clasificación (`DecisionTreeClassifier`) de la librería *scikit-learn*.

El modelo fue configurado con los siguientes parámetros:

Parámetro	**Valor**	**Descripción**
criterion	Gini	Mide la calidad de una división basándose en la impureza de Gini. Solo aplicable en tareas de clasificación en la API de scikit-learn.
splitter	best	Selecciona automáticamente la mejor división posible para cada nodo.
max_depth	4	Limita la profundidad del árbol a 4 niveles para evitar sobreajuste y facilitar la interpretación del modelo.
random_state	42	Fija una semilla para garantizar la reproducibilidad.
min_samples_split	2	Establece un mínimo de 2 muestras para realizar una división.
min_samples_leaf	1	Establece un mínimo de 1 muestra por hoja.
Otros parámetros	Por defecto	`max_features`, `max_leaf_nodes` y `ccp_alpha` mantienen valores por defecto, sin restricciones adicionales ni poda post-entrenamiento.

Cuadro 3: Parámetros de configuración del DecisionTreeClassifier

Se utilizó como variable objetivo, la clase *hospitalizado*, con valor binario (0, 1), mientras que como variables predictoras se usaron las variables *provincia*, *sexo*, *grupo_edad*, *num_casos*, *num_uci*, *num_def*, *anio*, *mes* y *semana*.

3.6.2.d Análisis de vulnerabilidad y diseño del ataque

Se realiza un análisis en búsqueda de puntos críticos en la estructura interna del modelo que permitiera identificar las vulnerabilidades posibles de explotar en el ataque de

envenenamiento de datos.

Para lo anterior, se identifican las características que más influyen en las decisiones del modelo mediante la obtención de las importancias de las características. Adicionalmente, se construyó un mapa de vulnerabilidad por nodo usando el grado de impureza por número de muestras. Los nodos con mayor puntaje *attack score* concentran el poder separador del árbol y, por tanto, constituyen los vectores de ataque más eficientes.

A partir de este análisis, se procede a diseñar un ataque de envenenamiento de datos dirigido a la clase objetivo, utilizando la técnica de *label-flipping*.

El ataque se diseña estableciendo una proporción, respecto a la cantidad total de registros, como cantidad de datos a envenenar. En este experimento se seleccionaron tres proporciones: 1 %, 3 % y 5 %.

El diseño del ataque definió un subconjunto "candidato" de envenenamiento a partir de los nodos más influyentes del árbol de decisión, identificados en el análisis de vulnerabilidad mediante la métrica *attack score* (producto de la reducción de impureza y el número de muestras en el nodo). Dentro de estos nodos, se priorizaron los registros de entrenamiento que se encontraban más próximos a los umbrales de división en las variables numéricas, o bien, pertenecían directamente a categorías relevantes en el caso de variables categóricas codificadas con *One-Hot*.

Al calcular el tamaño del subconjunto "candidato", se garantizó que éste fuese suficiente para cubrir los diferentes porcentajes de envenenamiento definidos en los experimentos (1 %, 3 % y 5 % del conjunto de entrenamiento). La selección de los registros se realizó con reemplazo de las etiquetas originales, de modo que cada instancia podía ser volteada solo una vez, asegurando la trazabilidad del experimento.

Finalmente, la magnitud del ataque correspondió al propio porcentaje de envenenamiento aplicado, ya que se utilizó un esquema de *label-flipping*, donde la etiqueta *hospitalizado* fue invertida ($0 \rightarrow 1$ o $1 \rightarrow 0$) en los registros seleccionados. Este criterio fue suficiente para inducir contradicciones en zonas estratégicas de la frontera de decisión del árbol, sin necesidad de alterar los valores originales de las otras características, como se registró en los cuadros y gráficas comparativas de este TFM.

A diferencia del experimento de regresión, en el cual se manipuló una característica numérica, en este caso el ataque se produjo directamente por la alteración de las etiquetas

de la clase objetivo, lo que permite evaluar la sensibilidad del clasificador ante cambios en la consistencia de la información de entrenamiento.

Nota: El concepto **con reemplazo** utilizado en el contexto de este trabajo está considerado desde una perspectiva descriptiva del experimento y no debe confundirse con el concepto de reemplazo en python y su configuración `replace=True`, que es el concepto opuesto al utilizado en este trabajo.

4. Resultados

En esta sección se presentan los resultados obtenidos de los experimentos realizados con los datasets adaptados de criminalidad y covid19. Se incluyen métricas de rendimiento, análisis de vulnerabilidades y comparaciones entre modelos entrenados con datos sin y con envenenamiento. Cabe destacar que los resultados aquí presentados aplican al contexto específico de los experimentos desarrollados para este TFM y no son necesariamente generalizables a otros contextos o datasets. La investigación y las validaciones necesarias para generalizar estos resultados quedan fuera del alcance de este trabajo.

4.1. Dataset de criminalidad, experimento de regresión

4.1.1. Resultados del EDA

Del análisis exploratorio de los datos se obtienen los siguientes resultados:

La estructura del dataset inicial contiene 3.360 filas y 4 columnas. Estas columnas están compuestas por la variable *Cantidad* (variable objetivo), que es de tipo numérica, y las tres restantes son categóricas (*Comunidad*, *Categoría* y *Año*). Esta estructura se mantiene para la obtención de los resultados del EDA.

Posteriormente, tal como se señala en la sección 3.6.1.a, se aplica la codificación one-hot a las variables categóricas, lo que modifica la estructura del dataset para el entrenamiento del modelo. De la estructura modificada se obtienen 44 columnas, donde una contiene datos de tipo numéricos *Cantidad* y las 43 restantes contiene datos categóricos codificados.

Durante la exploración, no se encontraron valores nulos, ni faltantes en el dataset.

Respecto a la distribución de los datos, en la figura 2 se muestra la distribución de la cantidad total de delitos, de todas las comunidades autónomas de España y de todas las categorías de delitos, agrupados por años. El periodo inicia en el 2010 y se extiende hasta el 2023.

Figura 2: Dataset de criminalidad: Cantidad de delitos totales por años.

En la figura 3, se muestra la distribución de la cantidad total de delitos, de todos los años y de todas las categorías de delitos, agrupados por comunidad autónoma. Están incluidas como comunidades autónomas, las ciudades de *Ceuta* y *Melilla*, y un valor registrado como *En el extranjero*.

La comunidad con mayor cantidad de delitos es *Cataluña* con 6.158.792 delitos, seguida de *Madrid* con 5.320.672 delitos, *Andalucía* con 5.009.778 delitos y *Comunidad Valenciana* con 3.482.551 delitos. Las comunidades con menor cantidad de delitos son *La Rioja* con 130.362 delitos, *Ceuta* con 68.777 delitos y *Melilla* con 66.341 delitos.

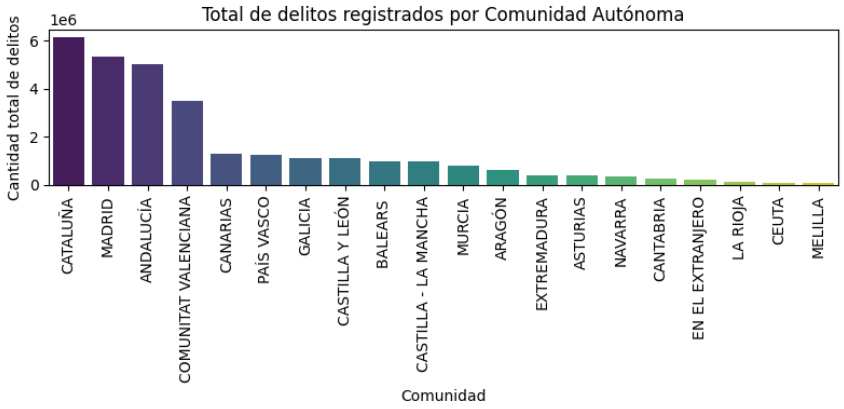


Figura 3: Dataset de criminalidad: Cantidad de delitos totales por comunidad autónoma en España.

En la figura 4 se muestra la distribución de la cantidad total de delitos, de todos los años y de todas las comunidades autónomas, agrupados por categoría del delito.

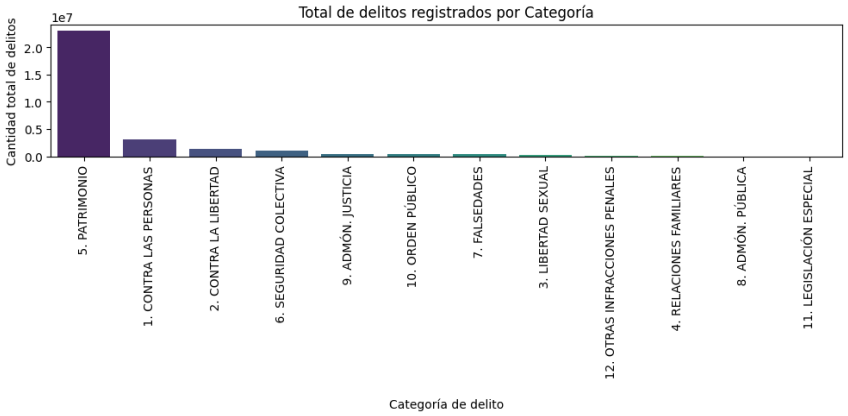


Figura 4: Dataset de criminalidad: Cantidad de delitos totales por categoría de delito.

La categoría de delito más frecuente es el efectuado contra el *Patrimonio* con 22.996.509 delitos, seguido por los pertenecientes a la categoría *Contra las personas* con 3.153.321 delitos y *Contra la libertad* con 1.346.046 delitos. Mientras los delitos menos frecuentes fueron los pertenecientes a las categorías de delitos de la *Admón. pública* con 6.580 delitos y los pertenecientes a *Legislación especial* con 3.623 delitos.

4.1.2. Resultados del análisis de vulnerabilidades

La figura 5 muestra las cinco características codificadas más importantes para la predicción de la criminalidad utilizadas en el árbol de regresión. Los valores de importancia son los siguientes: *Categoría_Patrimonio* con 0.39, seguida de *Comunidad_Cataluña* con 0.34, *Comunidad_Madrid* con 0,18, *Comunidad_Comunitat Valenciana* con 0,06 y *Año_2020* con 0,003.

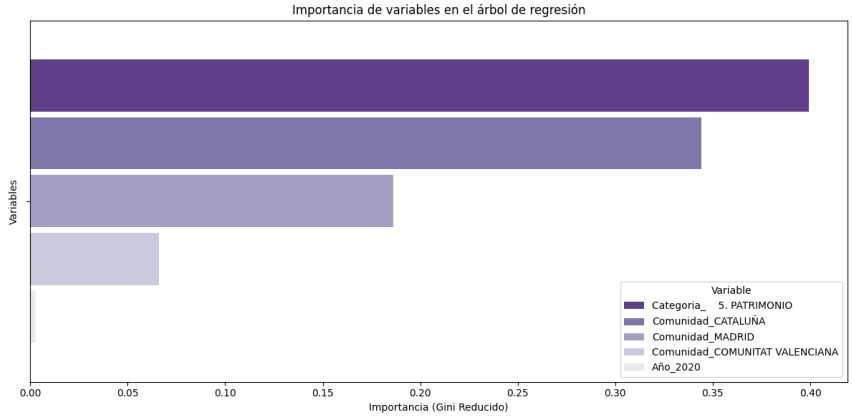


Figura 5: Dataset de criminalidad: Importancia de las características para el modelo.

Se obtienen los MAE de todas las comunidades, de los cuales los 5 primeros se muestran en la cuadro 4.

Posición	Comunidad Autónoma	MAE
1	ARAGÓN	12.038,58
2	MADRID	6.853,16
3	COMUNITAT VALENCIANA	6.780,99
4	EXTREMADURA	5.272,67
5	PAÍS VASCO	4.087,86

Cuadro 4: Dataset de criminalidad: Comunidades autónomas con mayor MAE en la predicción de criminalidad

La comunidad autónoma que presenta mayor MAE es Aragón, por lo cual se efectúa un análisis estadístico adicional que se muestra en el cuadro 5.

Este cuadro muestra los valores estadísticos observados en los datos de Aragón. Considera la media, desviación estándar, el valor máximo observado en Aragón y el valor máximo general en el dataset. Adicionalmente, se incorpora al cuadro el valor resultante de multiplicar la media por un factor de 3. Este factor representa la magnitud de la perturbación implementada sobre los valores del subconjunto "candidato" seleccionado para el ataque de envenenamiento de datos, dando como resultado 10.802,71.

Descripción	Valor
Media ARAGÓN	3.600,90
Desviación estándar ARAGÓN	8.842,31
Media × 3 (ataque)	10.802,71
Máximo observado en ARAGÓN	39.806,00
Máximo general en el dataset	431.028,00

Cuadro 5: Dataset de criminalidad: Datos estadísticos de ARAGÓN

4.1.3. Resultados de la evaluación de los modelos

En la siguiente figura, se observa gráficamente la estructura del árbol de decisión entrenado con el dataset sin envenenamiento de datos.

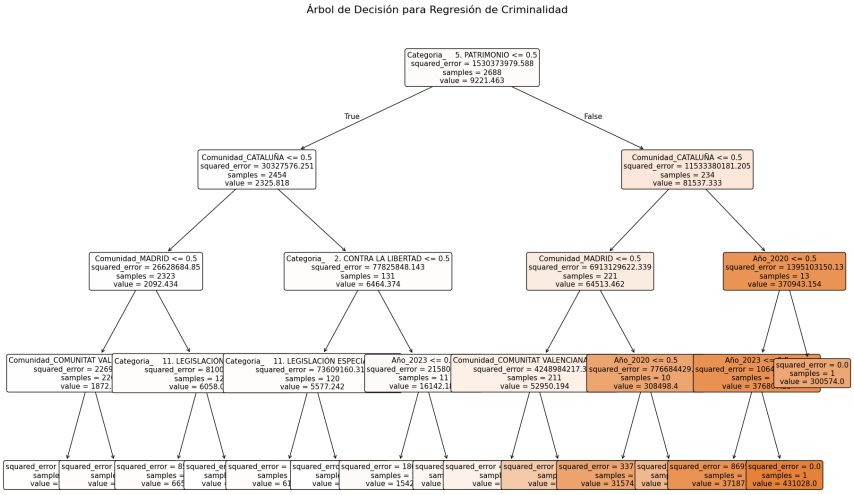


Figura 6: Dataset criminalidad: Estructura del árbol entrenado con datos sin envenenamiento.

El modelo entrenado con el dataset sin envenenamiento de datos, inicia en el nodo raíz con la característica codificada *Categoría_Patrimonio*. Usa un umbral de 0,5 para segmentar las decisiones por nodo. Cuando la característica codificada *Categoría_Patrimonio* es menor o igual a 0,5, el modelo toma la rama izquierda (False) de lo contrario, toma la rama derecha (True) y consulta por la siguiente característica codificada que es *Comunidad_Cataluña*, manteniendo el mismo umbral en todos los nodos. Este proceso sigue sucesivamente hasta llegar a las hojas del árbol.

En la figura 7 se observa la estructura del árbol de decisión entrenado con la mayor proporción de datos envenenados del experimento, correspondiente a una proporción del 5 % de datos envenenados sobre el total de registros del dataset. El ataque utilizado es del tipo *feature-based* como se explicó en la sección 3.6.1.d.

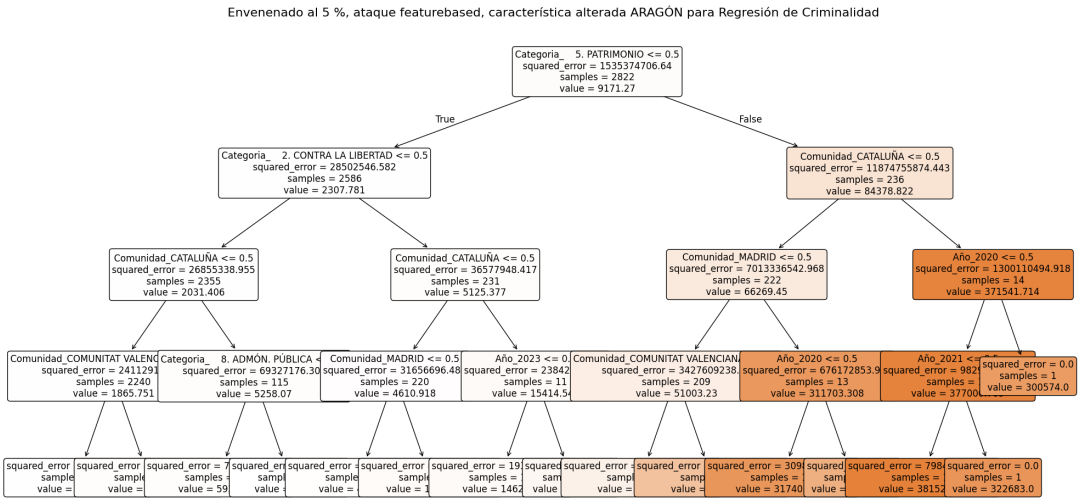


Figura 7: Dataset criminalidad: Estructura del árbol entrenado con datos envenenados (5 % de envenenamiento).

Comparativamente, entre el árbol de decisión entrenado sin y con el 5 % de datos envenenados, se observan los siguientes puntos de interés:

El nodo raíz mantiene la característica, *Categoría_Patrimonio*, como la primera variable de decisión. Sin embargo, en el modelo envenenado, aumenta el error cuadrático medio en 5.000.727,052 unidades. La cantidad de registros en el nodo aumentan de 2.688 a 2.822 y, la predicción de la cantidad de delitos en ese nodo disminuye de 9.221,463 a 9.171,27.

A partir del segundo nivel del árbol, se observan cambios en la elección de características para hacer el split. El nodo de la izquierda del nodo raíz, sustituye la característica codificada *Comunidad_Cataluña* por la característica codificada *Categoría_Contra la libertad*. En el nodo de la derecha, se mantiene la característica codificada *Comunidad_Cataluña*, pero con cambios en las otras medidas.

A tercer nivel, siguiendo solo la rama izquierda del nodo raíz, se observa que las características codificadas *Comunidad_Madrid* y *Categoría_Contra la libertad* se sustituyen por *Comunidad_Cataluña*, en ambos casos.

El detalle de los cambios en las estructuras de los árboles de todos los experimentos de regresión se incluyen en el cuadro 13 del anexo A.

Respecto de los resultados de la medición del rendimiento de los todos los modelos,

entrenados con y sin datos envenenados, han sido detallados en el cuadro 6, que puede consultarse a continuación.

Tasa de Envenenamiento	MAE	MSE	R^2
0 % (sin envenenar)	4315,89	194364378,05	0,8415
1 % (con reemplazo)	4343,74	194273114,85	0,8415
1 %	4787,53	254497063,18	0,6256
3 %	4459,14	187372813,69	0,8555
5 %	6195,83	497810941,56	0,5047

Cuadro 6: Dataset de criminalidad: Comparativa de métricas de rendimiento con distintas tasas de envenenamiento

4.2. Dataset de covid19, experimento de clasificación

4.2.1. Resultados del EDA

Del análisis exploratorio de los datos se obtienen los siguientes resultados:

La estructura del dataset inicial contiene 300.372 filas y 10 columnas. Las características están compuestas por tres columnas (*provincia*, *sexo*, *grupo_edad*) con datos de tipo categóricos y las restantes (*num_casos*, *num_uci*, *num_def*, *anio*, *mes*, *semana*, *hospitalizado* la variable objetivo) contienen datos de tipo numéricos. Esta estructura se mantiene para la obtención de los resultados del EDA.

Posterior a la aplicación de la codificación one-hot a las variables categóricas, el dataset para el entrenamiento del modelo es modificado y las columnas aumentan a 65, donde las 6 columnas iniciales con datos de tipo numéricos se mantienen y las 54 restantes contienen datos categóricos codificados.

Durante la exploración, se identificaron 6.386 valores nulos en total, todos pertenecientes a la variable *provincia*. La proporción de valores nulos por característica se muestra en la figura 8.

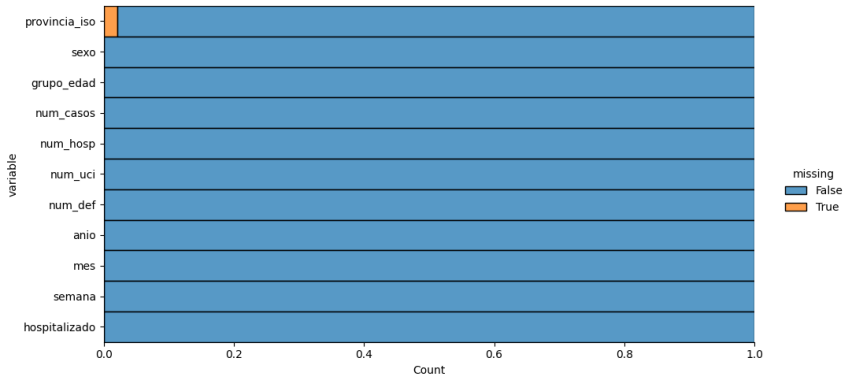


Figura 8: Dataset de covid19: Proporción de valores nulos por característica.

Los valores nulos fueron imputados a valor *Desconocido* para efectos del entrenamiento, cuya justificación se detalla en la sección 5.2.1.

Respecto a la distribución de la clase objetivo, en la figura 9 se muestra la distribución de la clase *hospitalizado*, con valor binario (0, 1), donde 1 indica que hay al menos un caso de paciente hospitalizado y 0 que no hubo ningún caso hospitalizado. Se observó que 156.419 registros están etiquetados como *hospitalizado*, lo que representa un 52,1 %

de registros con valor 1 y, los registros de los casos en que no hubo hospitalizados son 143.953, que representa un 47,9 % de registros con valor 0.

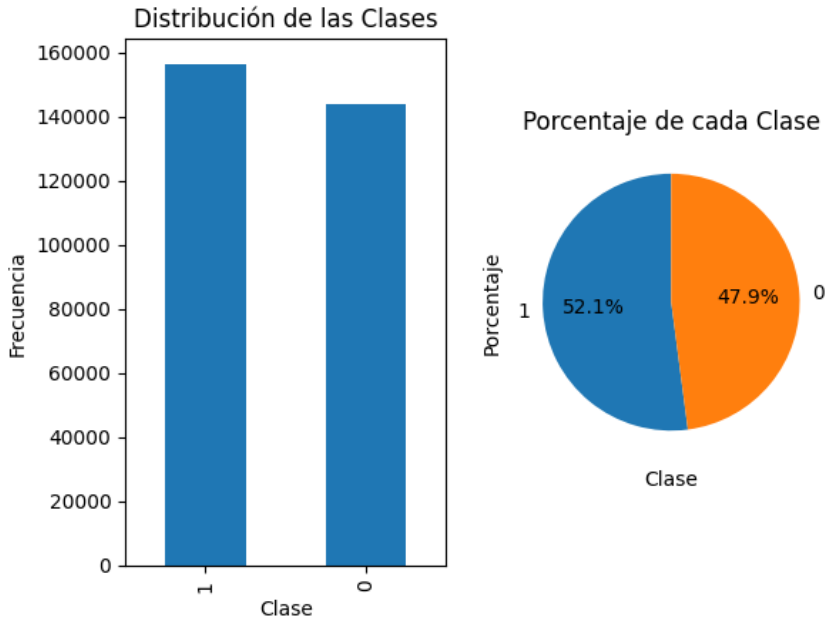


Figura 9: Dataset de covid19: Distribución de las clases hospitalizado (1) y no hospitalizado (0) frente a evento de covid19 confirmado.

También se realizó un análisis de correlación entre las variables del dataset, cuyos resultados se presentan en la figura 10.

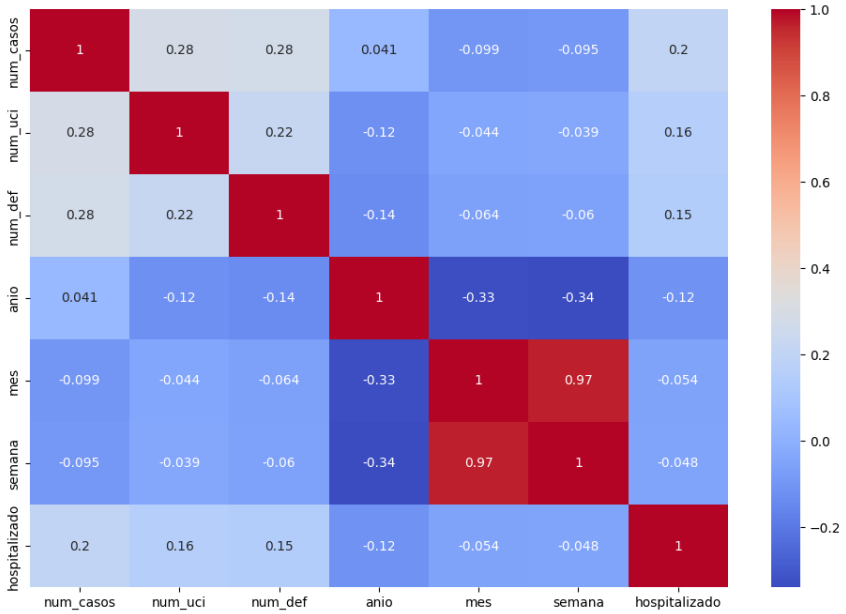


Figura 10: Dataset de covid19: Matriz de correlación entre variables.

4.2.2. Resultado del análisis de vulnerabilidades

La figura 11 presenta las 5 características expandidas (con características codificadas) más relevantes en el árbol de clasificación. Se observa que los valores de importancia de las características son: *num_num_casos* con 0,56, seguida por *num_num_def* con 0,14 y la característica codificada *cat_grupo_edad_80+* con 0,14. Finalmente, *num_num_uci* con 0,12 y *num_anio* con 0,013.

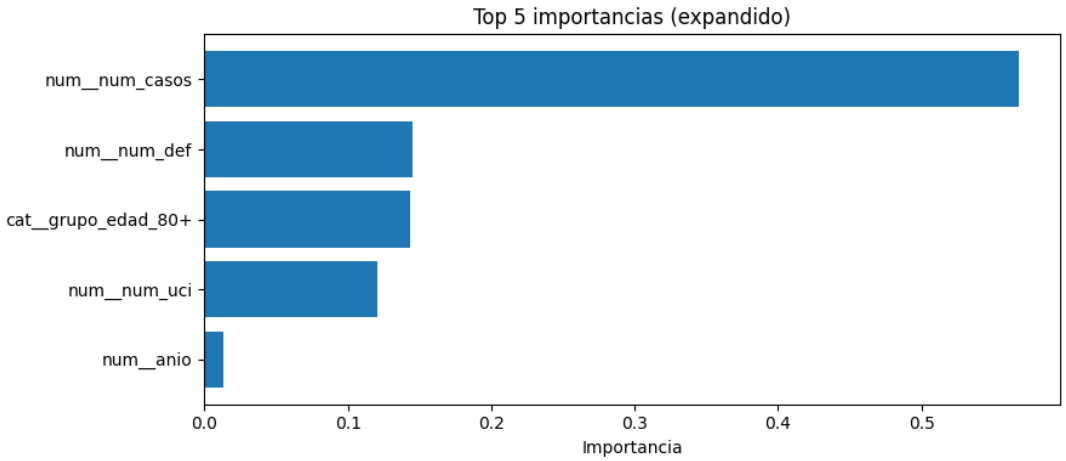


Figura 11: Dataset de covid19: Importancia de las características para el modelo.

En la figura 12 se presenta el mapa de vulnerabilidades del árbol de clasificación. En el eje horizontal se representan los nodos del modelo, mientras que el eje vertical indica el valor de *threshold* (umbral) en escala logarítmica. El color de cada punto corresponde al valor normalizado del *attack_score* (puntuación del ataque), calculado como el producto entre la variación de impureza y el número de muestras en el nodo, siendo los valores más cercanos al amarillo los de mayor puntuación relativa. Se observa que la característica *num_casos* concentra los nodos con mayor puntuación, incluyendo el nodo raíz (node_id $\approx$ 0), que aparece con el círculo más grande y de color más claro. En contraste, otros nodos como el 14 y el 28 presentan umbrales altos, pero con círculos más pequeños y oscuros, lo que refleja puntuaciones de vulnerabilidad más bajas.

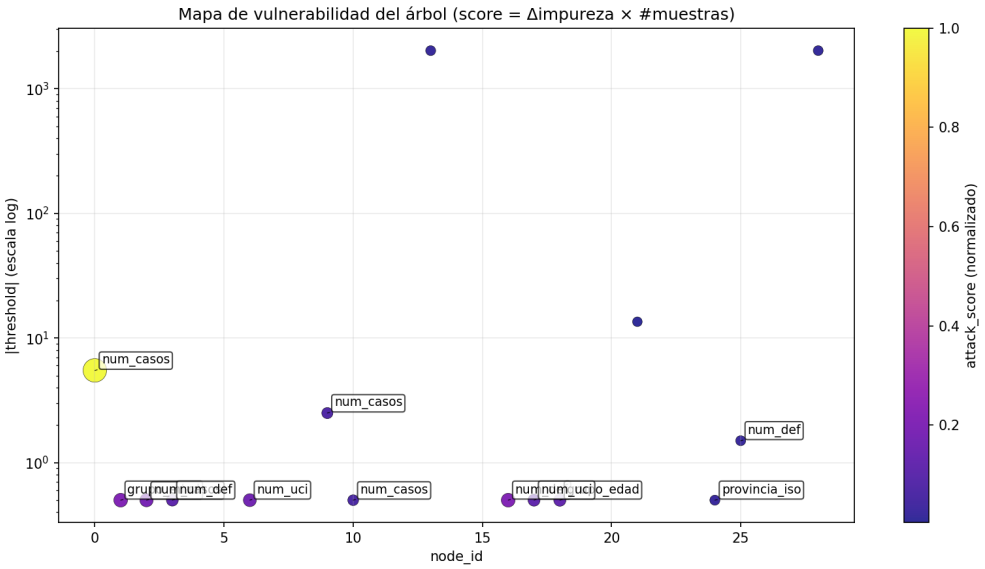


Figura 12: Dataset de covid19: Mapa de vulnerabilidades.

4.2.3. Resultados de la evaluación de los modelos

En la siguiente figura, se observa gráficamente la estructura del árbol de decisión entrenado con el dataset sin envenenamiento de datos.

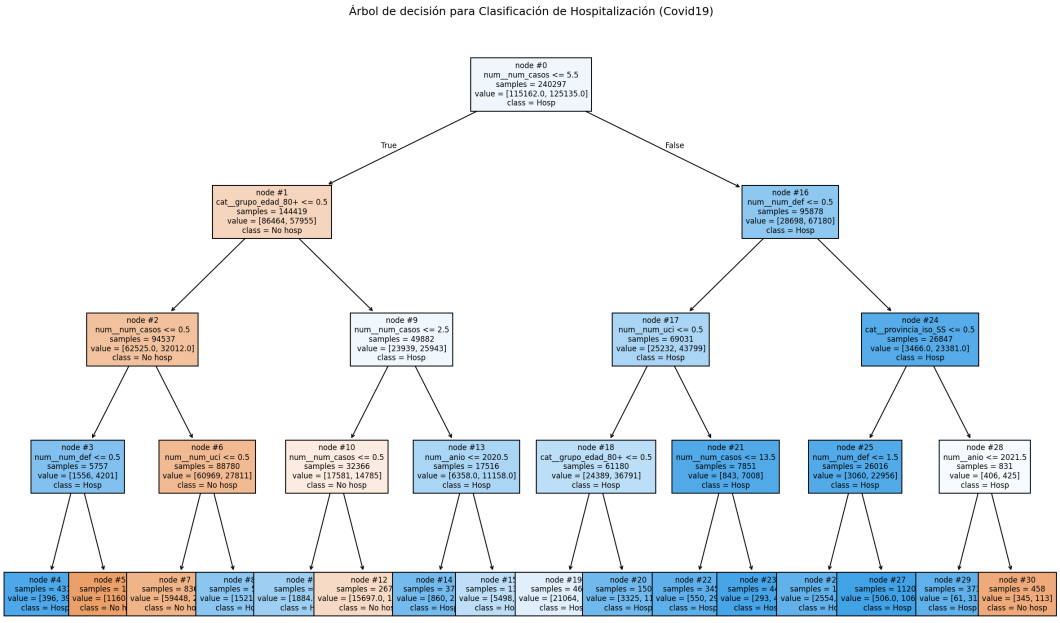


Figura 13: Dataset de covid19: Estructura del árbol de decisión entrenado con datos sin envenenamiento.

El modelo entrenado con el dataset sin envenenamiento de datos, inicia en el nodo

raíz con la característica *num_num_casos*. Usa un umbral de 5,5 para segmentar en ese nodo. Cuando la característica *num_num_casos* es menor o igual a 5,5, el modelo toma la rama izquierda (True) y consulta por la siguiente característica que es la característica codificada *cat_grupo_edad_80+*, usando un umbral diferente correspondiente a 0,5. Volviendo al nodo raíz, si la característica *num_num_casos* es mayor a 5,5, entonces toma la rama derecha (False) y consulta por la siguiente característica que es *num_num_def*, modificando también el umbral a 0,5. Este proceso sigue sucesivamente hasta llegar a las hojas del árbol.

En la figura 14 se observa la estructura del árbol de decisión entrenado con la mayor proporción de datos envenenados del experimento, correspondiente a una proporción del 5 % de datos envenenados sobre el total de registros del dataset. El ataque utilizado es del tipo *label-flipping* como se explicó en la sección 3.6.2.d.

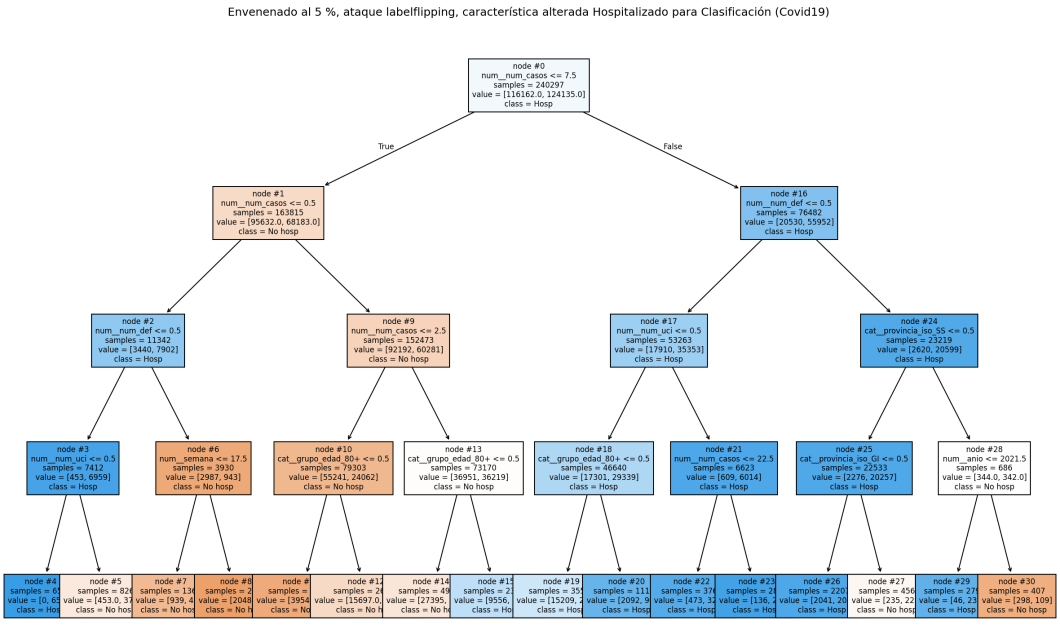


Figura 14: Dataset de covid19: Estructura del árbol de decisión entrenado con datos envenenados (5 % de envenenamiento).

Comparativamente, entre los árboles de decisión entrenados sin y con datos envenenados, se observan los siguientes puntos de interés:

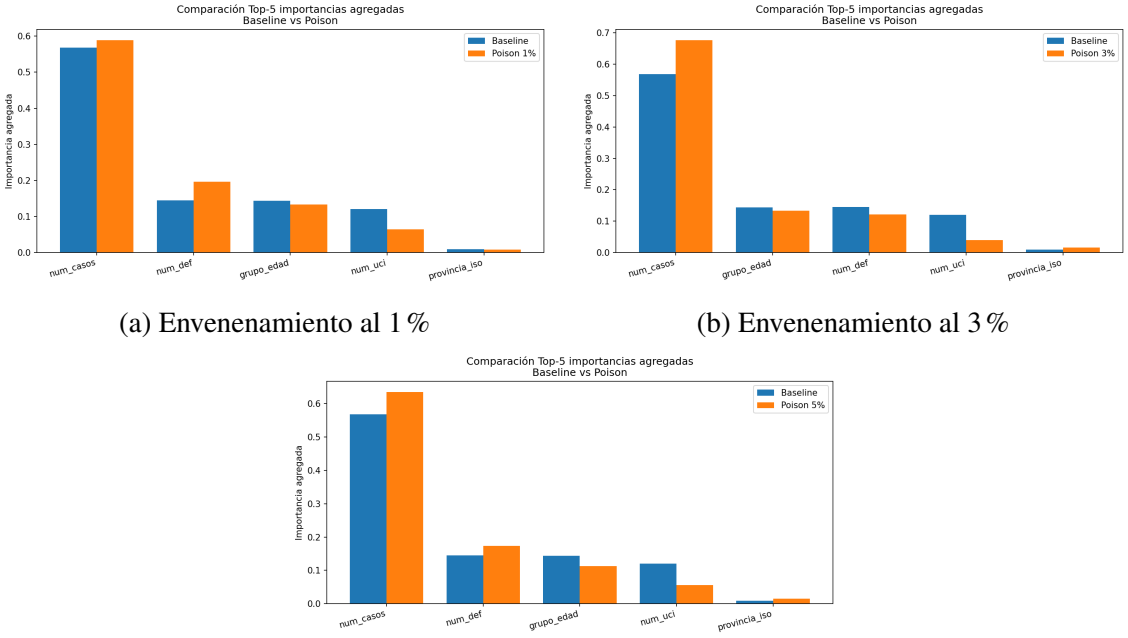
El nodo raíz mantiene la característica *num_num_casos*, como la primera variable de decisión. Sin embargo, en el modelo envenenado, aumenta el umbral de 5,5 a 7,5 y el

índice Gini cambia de 0,499139 a 0,499450. La cantidad de ejemplos no cambia.

A partir del segundo nivel del árbol, se observan cambios en la elección de características para hacer el split. En el nodo de la izquierda del nodo raíz, sustituye la característica codificada *cat_grupo_edad_80+* por la característica *num_num_casos*. En el nodo de la derecha, se mantiene la característica *num_num_def*, pero varía la cantidad de registros.

El detalle de los cambios en la estructura de los árboles de todos los experimentos de clasificación se incluyen en el cuadro 14 del anexo A.

La figura 15 presenta una comparación de las importancias de las características entre las encontradas en los modelos entrenados con datos envenenados al 1 %, 3 % y 5 %, y con datos sin envenenar o *baseline*.



(a) Envenenamiento al 1 %

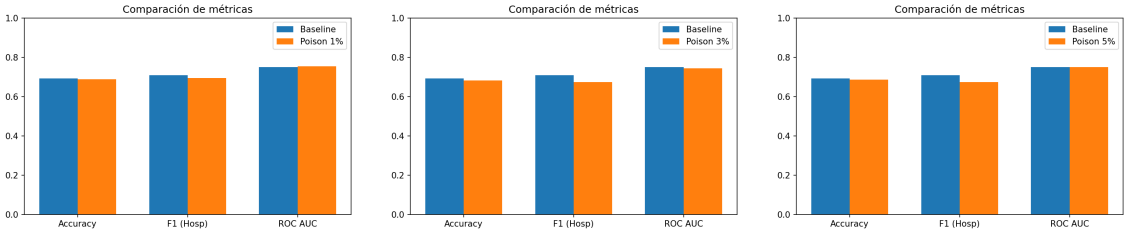
(b) Envenenamiento al 3 %

(c) Envenenamiento al 5 %

Figura 15: Dataset de covid19: Comparativa importancias de características

En todos los casos, la característica *num_casos* mantiene su posición como aquella con el peso relativo de importancia más alto, incluso aumenta con el envenenamiento al 1 % y 3 %, manteniéndose alta al 5 %. En contraste, la característica *num_uci* y *grupo_edad_80+* muestran una disminución en su importancia relativa. Finalmente, se muestra la característica *provincia*, que corresponde al dummy *provincia_SS*.

Las gráficas comparativas para todos los modelos entrenados sin y con datos envenenados se muestran en las figuras 16 y 17, y en la cuadro 7.

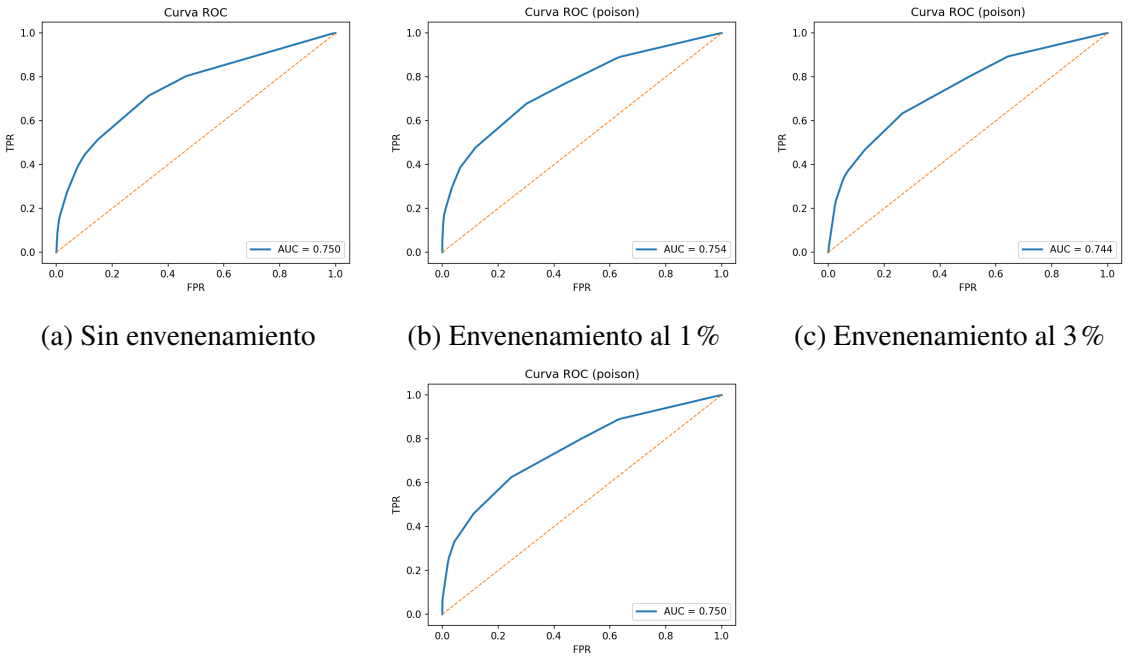


(a) Envenenamiento al 1 % (b) Envenenamiento al 3 % (c) Envenenamiento al 5 %

Figura 16: Dataset de covid19: Comparativa de métricas de rendimiento. El *F1 (Hosp)* mostrado corresponde a la clase positiva (hospitalizado).

Tasa de Envenenamiento	Accuracy	ROC	AP
0 % (modelo sin envenenamiento)	0,6926	0,750	0,755
1 %	0,6879	0,754	0,768
3 %	0,6819	0,744	0,747
5 %	0,6864	0,750	0,760

Cuadro 7: Dataset de covid19: Comparativa del rendimiento de los modelos con distintas tasas de envenenamiento



(a) Sin envenenamiento (b) Envenenamiento al 1 % (c) Envenenamiento al 3 %

(d) Envenenamiento al 5 %

Figura 17: Dataset de covid19: Comparativa de métrica ROC

La figura 16 presenta la comparación de las métricas *Accuracy*, *F1 (Hosp)* y *ROC*

AUC entre el modelo *baseline* y los escenarios de envenenamiento al 1 %, 3 % y 5 %. Se observa que, en todos los casos, los valores de las métricas disminuyen levemente respecto al modelo entrenado con datos sin envenenamiento, con la variación más visible en la métrica *F1 (Hosp)*.

Por su parte, el cuadro 7, resume las métricas de rendimiento de los modelos en función de la tasa de envenenamiento, mostrando una tendencia general de disminución en todas las métricas evaluadas e incorpora la métrica *AP* (Average Precision).

Posteriormente, se incorporan las gráficas del comportamiento de la curva ROC en forma comparativa, donde se observa cambios sutiles en la forma de las curvas, frente a las diferentes tasas de envenenamiento.

El cuadro 8 presenta un resumen de las métricas de rendimiento de los modelos en función de la matriz de confusión normalizada.

Tasa de Envenenamiento	**VP**	**VN**	**FP**	**FN**
0 % (sin envenenar)	0,72	0,67	0,28	0,33
1 %	0,68	0,70	0,32	0,30
3 %	0,63	0,74	0,37	0,26
5 %	0,62	0,75	0,38	0,25

Cuadro 8: Dataset de covid19: Comparativa del rendimiento según matriz de confusion normalizada

Finalmente, la figura 18 muestra la comparación de las fronteras de decisión entre el modelo *baseline* a la izquierda y el modelo entrenado con mayor porcentaje de datos envenenados, es decir, con un 5 % de envenenamiento a la derecha.

En ambos gráficos se representan visualmente las regiones de clasificación, donde la zona en verde claro corresponde a la predicción de la clase 0 (No hospitalizado), mientras que la zona en rosa corresponde a la predicción de la clase 1 (Hospitalizado). Los puntos representan los casos reales, distinguiendo entre puntos naranja que son los positivos (hospitalizados) y puntos azules que son negativos (no hospitalizados).

La comparación permite observar diferencias en las áreas donde se concentran los falsos positivos (FP) y falsos negativos (FN), siendo estas más pronunciadas en el modelo envenenado, con un incremento en la proporción de FN (19,59 % frente a 14,81 %) y una

reducción en FP (11,77 % frente a 15,93 %) respecto al baseline.

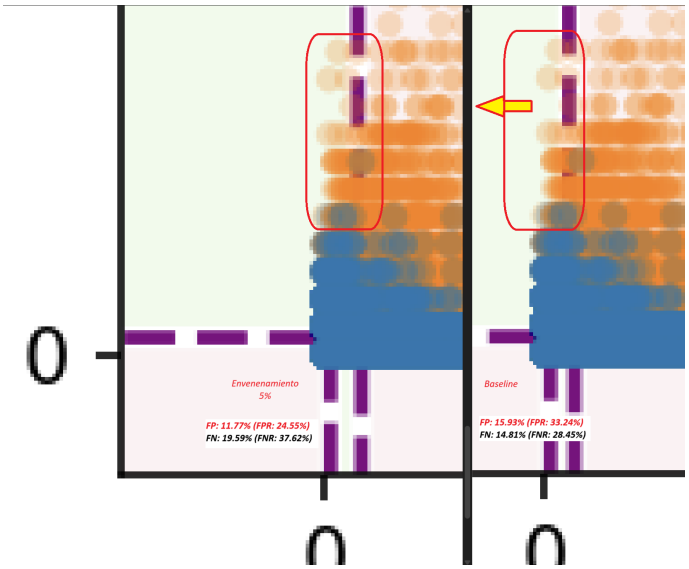


Figura 18: Dataset de covid19: Comparativa de fronteras de falsos positivos y falsos negativos del modelo entrenado con datos envenenados (5 % de envenenamiento).

5. Discusión

En esta sección se presenta el análisis y discusión de los resultados obtenidos en los experimentos realizados con los datasets de criminalidad y covid19, abordando la interpretación de los hallazgos y sus implicaciones en el contexto de la seguridad de los modelos de aprendizaje automático.

5.1. Dataset de criminalidad

5.1.1. Análisis exploratorio de datos

Durante el análisis exploratorio de los datos (EDA) del dataset de criminalidad, se observaron varios aspectos relevantes que permiten comprender mejor la naturaleza de los datos y su distribución.

Uno de los primeros aspectos es el relacionado a la extensión de las dimensiones que se produce por el one-hot encoding, que transforma las variables categóricas en múltiples variables binarias. Este proceso incrementó significativamente el número de columnas del dataset, pasando de 4 a 44 columnas. Este aumento en la dimensionalidad puede tener

implicaciones en el rendimiento del modelo, ya que un mayor número de características puede llevar a un fenómeno conocido como la *maldición de la dimensionalidad*, donde el espacio de características se vuelve tan grande que los datos se vuelven escasos, dificultando la generalización del modelo. Sin embargo, en este caso, el uso de árboles de decisión ayuda a mitigar este problema, ya que estos modelos son capaces de manejar bien conjuntos de datos con alta dimensionalidad y seleccionar automáticamente las características más relevantes durante el proceso de entrenamiento.

Por otra parte, como no se encontraron valores nulos o faltantes en el dataset, se considera que esta situación tiene explicación por la propia condición agregada del dataset, que correspondería a registros de tipo "resumen o totales" de las variables. En este sentido, la ausencia de valores nulos puede interpretarse como que no existen combinaciones de variables sin datos registrados, lo que es consistente con la naturaleza agregada del dataset.

Esta misma condición, de que los datos estén agregados, se contempló para analizarlos agrupados por sus respectivos criterios de agregación.

Respecto a los resultados obtenidos al agruparlos por año se puede ver, en la figura 2, que la cantidad de delitos se distribuye de forma heterogénea dentro del periodo 2010 al 2023. En el año 2023 se registran 2.464.759 de delitos, la mayor cantidad de delitos dentro del rango total de años. Por otra parte, con 1.766.779 delitos, el menor registro de delitos es en el año 2020. Si bien, el dataset no da información adicional que pudiera explicar este comportamiento una contextualización de la situación de España podría darnos algunas pistas, como por ejemplo que en el año 2020 se vivió el confinamiento por el COVID-19 lo que sería consistente con observar una disminución en los delitos, así como un aumento en el año 2023, consistente con el término del período de confinamiento.

Luego, los resultados obtenidos al agrupar los datos por comunidad autónoma se muestran en la figura 3. En ella, se observa que la distribución de delitos es también heterogénea, pero con bastante diferencia entre las tres comunidades autónomas que acumulan la mayor y menor cantidad de delitos. Para analizar este comportamiento, se calcula el porcentaje de delitos por comunidad autónoma y se muestran en el cuadro 9. En este cuadro se observa que las tres primeras comunidades autónomas mencionadas, juntas, explican el 54,86 % de los delitos. Por otra parte, las tres últimas, explican menos del 1 % de

la cantidad de delitos cometidos por comunidades autónomas.

Comunidades Autónomas	Mayor % delitos	Menor % delitos
Cataluña	20,49 %	–
Madrid	17,70 %	–
Andalucía	16,67 %	–
La Rioja	–	0,43 %
Ceuta	–	0,22 %
Melilla	–	0,22 %

Cuadro 9: Dataset de covid19: Los 3 porcentajes mayores y menores de delitos.

Nuevamente, el dataset no da otras características que nos permitiera relacionar causas, pero si se incluye un enfoque demográfico de España, las tres comunidades autónomas que concentran el mayor porcentaje de delitos, se corresponden con las comunidades autónomas más pobladas de España, y al contrario, las últimas tres comunidades autónomas que menor cantidad de delitos presentan son también las que se registran como las menos pobladas.

Finalmente, al agrupar los datos por categoría de delitos, los resultados se muestran en la figura 4. En ella se observa que la categoría de delito que mayor cantidad de ellos registra, es la categoría de delitos contra el *Patrimonio*, que abarca la cantidad de 22.996.509 delitos y representa el 76,52 %. Si bien no hay información adicional, se puede inferir que los delitos contra la propiedad suelen ser los más frecuentes por la gran diversidad de modalidades por las cuales pueden llevarse a cabo, muchos de ellos con penas bajas o inexistentes y la falta de consentimiento o participación que se requiere de la persona contra la que se comete el delito. Estas, pueden ser algunas razones de aporte experiencial que podrían explicar este comportamiento.

Cabe señalar que, si bien el objetivo del presente experimento no es desarrollar el mejor modelo predictivo para el fenómeno de la criminalidad, sí resulta necesario realizar un análisis exploratorio, al menos en forma breve, que permita comprender los datos en su contexto. En este sentido, el EDA presentado aporta la caracterización mínima indispensable sobre la distribución temporal, territorial y categórica de los delitos, constituyendo la base sobre la cual se construye y justifica el uso del modelo de regresión con árboles

de decisión en este trabajo.

5.1.2. Evaluación de la vulnerabilidad del modelo y diseño del ataque

La evaluación de las vulnerabilidades del modelo entrenado con datos sin envenenar, arrojó como resultado dos enfoques complementarios: por un lado, el análisis de la importancia de las características, que permitió identificar las variables más determinantes en la predicción; y, por otro, el análisis de los puntos que desestabilizan al modelo por su alto error.

Los resultados del análisis de la importancia de las características identificaron las cinco más relevantes, es decir, aquellas que el modelo considera son las más importantes para reducir el error (MSE) en la predicción de la criminalidad.

La característica codificada *Categoría_Patrimonio*, que corresponde a los delitos contra el patrimonio, es la que más reduce el error, por tanto, el árbol separa con mucha eficacia los valores objetivos cuando distingue si el delito pertenece a ella o no.

Le siguen las características codificada *Comunidad_Cataluña*, *Comunidad_Madrid* y *Comunidad_Comunitat Valenciana*, que a su vez, son las comunidades autónomas con mayor cantidad de delitos como se observó en el análisis exploratorio.

Luego, la característica codificada *Año_2020* es la variable temporal que más reduce el error, lo que indica que el modelo ha aprendido a identificar patrones temporales en los datos, aunque su importancia es menor en comparación con las otras características.

Del resultado presentado en la figura 5 se puede observar que la importancia de estas características es bastante concentrada, ya que solo las tres primeras características codificadas más importantes acumulan más del 90 % de la importancia total del modelo. Este resultado es consistente con el análisis exploratorio dado que, las características que en las gráficas de distribución de los datos mostraron mayor concentración en pocas categorías (tipo de delito y comunidad) son las que más reducen el error cuadrático en el modelo. Por el contrario, la variable temporal que es más equilibrada en su distribución, aporta una importancia marginal. Esto refuerza la coherencia entre el comportamiento del dataset y la explicación del modelo.

Realizado el análisis del segundo enfoque, relacionado a los puntos que desestabilizan al modelo por su alto error, se elaboró el cuadro 4. En ella se puede identificar las

cinco comunidades autónomas que presentan mayor MAE en la predicción de criminalidad. La comunidad autónoma que presenta mayor MAE es *Aragón*, con un valor de 12.038,58, que es casi el doble del segundo valor más alto, que corresponde a *Madrid* con 6.853,16. Esta diferencia significativa sugiere que el modelo tiene dificultades particulares para predecir la criminalidad en *Aragón*, lo que podría deberse a factores específicos de esta comunidad autónoma que no están bien capturados por las características del modelo, o por falta de datos representativos. Si bien, las comunidades autónomas de *Madrid* y *Comunidad Valenciana* también presentan un MAE alto y, adicionalmente, se corresponden con ser características consideradas importantes por el modelo, sus valores son considerablemente menores que el de *Aragón*, lo que indica que el modelo tiene un mejor desempeño en estas regiones en comparación con *Aragón*.

En dicho cuadro, se observa que la media de *Aragón* fue de μ =3.600,90 y la desviación estándar de $\sigma = 8{,}842{,}31$. De ahí se obtiene el umbral de $\mu + 2\sigma = 21.285{,}52$. En el cuadro también fueron registrados los valores máximos, tanto de *Aragón* con 39.806 y del dataset completo con 431.028, todo ello con el objetivo de tener una referencia clara del rango empírico de los datos. De esta forma, se observa que la magnitud de la perturbación obtenida en promedio fue de 10.802,71 que cumple el objetivo de mantener los valores envenenados dentro del rango empírico para que el ataque no de alertas tempranas de outliers.

5.1.3. Análisis comparativo

5.1.3.a Comparación de las estructuras de los árboles

En la figura 6 se puede observar el árbol sin envenenamiento que sirve de modelo baseline para realizar las comparaciones.

Este modelo inicia su nodo raíz con la característica codificada *Categoría_Patrimonio*, lo que indica que es la característica más importante para efectuar el primer split y es confirmado según el análisis de importancia de características. El árbol realiza su primera división en función de si el delito pertenece o no a esta categoría, lo que indica que el modelo ha aprendido a identificar este patrón como el más importante para separar con eficacia los valores objetivo. Para hacer esta separación utiliza el umbral de 0,5, que es el

valor estándar para variables binarias.

Cuando se compara esta característica en el árbol envenenado al 5 % de la figura 7, se observa que este también inicia su nodo raíz con la característica codificada *Categoría_Patrimonio*, lo que indica que en general la característica *Categoría* sigue siendo la más importante para efectuar el primer split, incluso después de introducir datos envenenados. Inclusive al revisar los resultados de los cambios en las estructura de los árboles en todos los experimentos (disponibles para consulta en el cuadro 13 del anexo A), esta característica se mantiene estable como la característica más importante en el nodo raíz, lo que sugiere que el envenenamiento no ha afectado la importancia relativa de esta característica en el modelo, aunque ya se pueden observar cambios en la cantidad de ejemplos y en el error cuadrático pero que se mantienen en el mismo orden de magnitud. Todos estos cambios incipientes implican que este nodo no ha sido absolutamente inmune a los ataques de envenenamiento.

A partir del segundo nivel del árbol se observan cambios en la estructura del árbol lo que comienza a evidenciar, más notoriamente, el deterioro de este. Así al comparar los cambios del nodo 1, el árbol sin envenenamiento utiliza la característica codificada *Comunidad_Cataluña* para efectuar el split, mientras que el árbol envenenado al 1 % utiliza la característica codificada *Comunidad_Madrid*, y los envenenados al 3 % y 5 % utilizan la característica codificada *Categoría_Contra La Libertad*, reflejando una clara inestabilidad en la decisión. Puesto que comienza a variar entre las características más generales *Comunidad* y *Categoría*.

Este cambio indica que el envenenamiento ha afectado la selección de características en este nodo, generando un deterioro estructural que podría considerarse fuerte ya que cambia totalmente la lógica del modelo, lo que podría tener implicaciones para la capacidad del modelo para generalizar correctamente. Esta alteración queda reafirmada con las métricas de rendimiento obtenidas, que se presentan más adelante.

Otro cambio importante a considerar es el observado en el nodo 8, en este nodo los experimentos sin envenenamiento y con el 1 % (con reemplazo) mantienen valores bajos y estables del *Value*, valores $\approx$ 14-16, pero al 1 % de envenenamiento da un salto importante a 15.465,9 para luego volver a valores más pequeños de 59 en el modelo envenenado al 5 %. Estos cambios de los valores en tan diferentes magnitudes también son observables

en otros nodos como los del 12 al 14. Este es un ejemplo de la evidencia obtenida respecto a la distorsión introducida (un envenenamiento mínimo hace que un nodo pase a predecir cifras que ya son consideradas outliers internos, es decir, se inducen los outliers).

Como punto particular de interés obtenido de la observación del desarrollo de esta experimentación es la observación de que no siempre un mayor porcentaje de datos envenenados significa mayor deterioro del modelo. Eso se pudo observar, por ejemplo en el nodo 8 con la explicación del párrafo anterior. Así el valor más alto de *Value* en el modelo envenenado al 1 % es 15.465,9, mientras que en el modelo envenenado al 5 % el valor es 59. Esto indica que el envenenamiento no siempre tiene un efecto lineal con el porcentaje de datos envenenados, por tanto, el incremento del deterioro podría depender más de qué registros específicos son alterados y cómo impactan en los split del árbol. Es decir, un cambio pequeño en registros considerados "críticos" por el modelo pueden desestabilizar mucho más que un cambio mayor pero en registros menos influyentes.

5.1.3.b Comparación de los rendimientos de los modelos

En el cuadro 6 se presentan las métricas de rendimiento de los modelos entrenados con diferentes tasas de envenenamiento, incluyendo el modelo sin envenenamiento que sirve como referencia (baseline).

Dentro de las observaciones de estos resultados, en primer lugar, se puede señalar que el modelo entrenado con datos sin envenenamiento (0 %), presenta un buen ajuste con un R^2 de 0,8415. Esto nos indica que el modelo es capaz de explicar aproximadamente el 84,15 % de la variabilidad en los datos. Además, el modelo muestra valores de error moderados (MAE = 4.315,89, MSE = 1.94×10^8). Cuando se introduce un 1 % de envenenamiento con reemplazo, las métricas se mantienen prácticamente inalteradas, lo que indica que este tipo de perturbación no logra degradar de manera significativa el desempeño.

Sin embargo, con un 1 % de envenenamiento se observa un deterioro considerable: el MAE aumenta en torno a un 11 %, el MSE se incrementa en más de un 30 % y el coeficiente de determinación, R^2, cae de 0,8415 hasta 0,6256. Este resultado refleja la vulnerabilidad del modelo ante pequeñas modificaciones dirigidas, capaces de alterar nodos clave en la estructura del árbol y de generar predicciones anómalas.

De forma paradójica, el escenario del 3 % se registra una recuperación del desempeño en algunas de las métricas, donde el R^2 asciende a 0,8555 y el MSE disminuye ligeramente, en torno a un 3,5 %, respecto al MSE del modelo entrenado sin datos envenenados. Este comportamiento puede interpretarse como un efecto de regularización accidental, en el que la presencia de ruido obliga al árbol a simplificar sus divisiones y mejora transitoriamente la capacidad de generalización.

Finalmente, con un 5 % de envenenamiento se alcanza un punto de colapso. El error absoluto medio aumenta más de un 40 %, el MSE se multiplica por 2,5 respecto al modelo base y el R^2 desciende de 0,8415 a 0,5047 (una disminución *aprox*. del 40 %), lo que evidencia una pérdida evidente de la capacidad explicativa.

En conjunto, los resultados confirman que el árbol de decisión es especialmente sensible a inyecciones específicas de datos envenenados, pudiendo mostrar tanto comportamientos de degradación inmediata como aparentes mejoras espurias antes de colapsar en niveles más altos de envenenamiento.

5.2. Dataset de covid19

5.2.1. Análisis exploratorio de datos

De los resultados del análisis exploratorio de datos (EDA) del dataset de covid19, se pueden destacar los siguientes aspectos relevantes:

Se identificaron 6.386 registros con valores nulos, lo que corresponde al 2,1 % de los registros que contenían un valor nulo en la columna *provincia*. En lugar de eliminarlos, se imputaron a una categoría “Desconocido” con el fin de preservar la muestra, mantener la neutralidad en la distribución territorial y asegurar la consistencia con la codificación one-hot utilizada en el modelo.

Si bien, el porcentaje de registros con valores nulos es bajo, y por tanto el riesgo de distorsión también lo es, se analizó que eliminarlos implicaría perder más de seis mil registros que contienen información válida en las demás variables. Por otra parte, al tratarse de *provincia*, la opción de imputar por media o moda se consideró que carecía de sentido (no hay un “promedio” de provincias), por tanto, el valor “Desconocido” se consideró como la forma más semánticamente coherente de reflejar que esa información está

ausente.

Respecto a la distribución de clases, tal como se observa en los resultados de la figura 9, la adaptación del dataset permitió obtener una distribución de la variable objetivo *hospitalizado* con cierto balance entre las dos clases: un 52,1 % de los registros corresponden a casos con al menos una hospitalización, mientras que el 47,9 % no presentan hospitalización. Esta proporción cercana al equilibrio es un aspecto positivo para el modelado, ya que reduce el riesgo de sesgo en el entrenamiento de los clasificadores y permite que las métricas de desempeño (Accuracy, F1-score, Recall, etc.) reflejen de manera más fiel la capacidad del modelo para diferenciar ambas clases. Asimismo, la ligera mayoría de casos hospitalizados garantiza que la clase positiva, de mayor interés analítico, se encuentre adecuadamente representada. Este balance permite que el modelo disponga de ejemplos suficientes para aprender patrones asociados a la hospitalización sin necesidad de técnicas de rebalanceo adicionales.

La matriz de correlación, de las variables numéricas, presentada en la figura 10 evidencia varios patrones relevantes que se discuten a continuación.

En primer lugar, se extraen de esa matriz las asociaciones entre las variables clínicas que son mostradas en el cuadro 10. Todas ellas son asociaciones positivas, con coeficientes de correlación que van desde r =0,22 hasta r =0,28. Estas relaciones son coherentes desde un punto de vista epidemiológico, ya que a mayor incidencia se espera un aumento en alguna proporción en los indicadores de gravedad.

Variables	*num_casos*	*num_uci*
num_uci	0,28	–
num_def	0,28	0,22

Cuadro 10: Dataset de covid19: Correlaciones entre variables clínicas

Respecto de la correlación entre las variables predictoras y la variable objetivo *hospitalizado*, esta mantiene correlaciones positivas bajas a moderadas con las variables clínicas. En contraste, las variables temporales presentan escasa relación con la hospitalización, con correlaciones negativas muy bajas. Estas correlaciones se resumen en el cuadro 11.

Hospitalizado		
Variables	**Correlación Positiva**	**Correlación Negativa**
num_casos	0,20	–
num_uci	0,16	–
num_def	0,15	–
anio	–	-0,12
mes	–	-0,05
semana	–	-0,04

Cuadro 11: Dataset de covid19: Correlaciones con la variable objetivo *hospitalizado*

Este resultado confirma que la hospitalización no depende linealmente de un único factor, sino de la combinación de distintos indicadores. Estos resultados permiten concluir que la dimensión temporal aporta información limitada en comparación con las variables clínicas para explicar la hospitalización.

5.2.2. Evaluación de la vulnerabilidad del modelo y diseño del ataque

La evaluación de las vulnerabilidades del modelo entrenado con datos sin envenenar, arrojó como resultado dos enfoques complementarios: por un lado, el análisis de la importancia de las características, que permitió identificar las variables más determinantes en la predicción; y, por otro, la elaboración de un mapa de vulnerabilidades, en el que se localizaron los nodos y categorías que desestabilizan al modelo por presentar mayores niveles de impureza.

Los resultados del análisis de la importancia de las características identificaron las cinco más relevantes, es decir, aquellas cinco que el modelo considera más influyentes para reducir la impureza (índice Gini) en la clasificación de la variable objetivo.

La variable con mayor importancia es *num_num_casos*, que por sí sola explica más del 56 % de la reducción total de impureza, lo que indica que el modelo se apoya fuertemente en esta característica para realizar sus predicciones de hospitalización.

Le siguen las variables *num_num_def* y la categoría codificada *cat_grupo_edad_80+*, ambas con una contribución relevante y de magnitud similar entre ellas, lo que evidencia que la mortalidad y la pertenencia al grupo etario de 80 años o más constituyen fac-

tores significativos para la clasificación. En un nivel ligeramente inferior se encuentra *num_num_uci*, que mantiene una influencia moderada. Finalmente, la variable temporal *anio* presenta una importancia residual, confirmando que la dimensión temporal aporta muy poco al desempeño del modelo en comparación con las variables clínicas.

En resumen, solo las cuatro primeras características (*num_num_casos*, *num_num_def*, *cat_grupo_edad_80+* y *num_num_uci*) concentran un valor de importancia del modelo de 0,97, esto es que pueden explicar más del 97 % de la reducción total de impureza del modelo, lo que indica una alta concentración de la capacidad explicativa del modelo en un conjunto reducido de variables.

Con un segundo enfoque, se analiza el mapa de vulnerabilidades lo que permite identificar los puntos críticos del árbol donde una manipulación de los datos tendría mayor impacto en la clasificación. Destaca de manera evidente la variable *num_casos*, la cual concentra los nodos con mayor *attack_score*, lo que refleja que el modelo depende de manera significativa de esta característica para sus divisiones principales. Esta concentración sugiere que pequeñas perturbaciones sobre *num_casos* pueden inducir cambios estructurales relevantes en la predicción, aumentando la fragilidad del modelo.

Asimismo, se observa que variables como *num_def*, *num_uci* y *grupo_edad* aparecen como vulnerabilidades secundarias, con valores de *attack_score* más bajos pero todavía presentes en varios nodos. Esto indica que, aunque su influencia es menor que la de *num_casos*, podrían ser utilizadas en ataques focalizados para degradar la precisión del clasificador.

Un hallazgo adicional es la presencia de la variable *provincia*, cuya importancia había resultado marginal en el análisis de importancias globales, pero que en el mapa muestra vulnerabilidades puntuales. Este contraste evidencia que una variable con baja importancia general puede, sin embargo, generar inestabilidad local en determinados nodos del árbol, lo que amplía el espectro de posibles puntos de ataque.

En conjunto, el análisis refuerza la idea de que la vulnerabilidad del modelo no depende únicamente de las características globalmente más importantes, sino también de aquellas que, en nodos específicos, concentran altos niveles de impureza y muestras, convirtiéndose en puntos sensibles para un ataque de envenenamiento de datos.

Los aportes para el análisis de vulnerabilidad de ambos enfoques pueden señalarse co-

mo: en el caso del análisis de importancia de características aportó una visión global sobre las variables que estructuran el modelo, y el mapa de vulnerabilidades permitió identificar con mayor precisión los puntos débiles específicos, asociados a nodos y categorías con mayor índice de impureza.

Por tanto, el análisis de importancia de características como el mapa de vulnerabilidades permitieron evidenciar la dependencia del modelo de clasificación respecto de un conjunto reducido de variables, destacando especialmente *num_casos*. Esta concentración de dependencia en pocas variables implica que la calidad de las etiquetas de entrenamiento resulta crítica para la estabilidad del clasificador, ya que cualquier inconsistencia entre los predictores y la clase objetivo altera de manera directa la lógica de decisión. Este factor fue el punto clave para definir el diseño del ataque, pues identificó que la manipulación de la etiqueta de salida constituía un punto de vulnerabilidad particularmente sensible para este modelo y, por tanto, un ataque de tipo *label-flipping* sobre la variable objetivo *hospitalizado* era apropiado.

5.2.3. Análisis comparativo

5.2.3.a Comparación de las estructuras de los árboles

En la figura 13 se puede observar el árbol sin envenenamiento que sirve de modelo baseline para realizar las comparaciones.

Este modelo inicia su nodo raíz con la característica *num_num_casos*, lo que confirma que es la variable más importante para efectuar el primer split. Este resultado es consistente con el análisis de importancia de características, ya que *num_num_casos* resulta ser el predictor principal de la necesidad de hospitalización en el conjunto de datos. Para realizar esta separación se utiliza el umbral de 5,5, dividiendo entre escenarios con muy pocos casos frente a aquellos con un mayor volumen de casos, lo que refleja el patrón aprendido por el modelo para discriminar entre hospitalización y no hospitalización.

En el segundo nivel del árbol, cuando la característica *num_num_casos* tiene valores bajos y el flujo continua por la rama izquierda, la característica codificada que se utiliza para el siguiente split es la *cat_grupo_edad_80+*. Esta división captura la vulnerabilidad especial de los adultos mayores, lo que coincide con la lógica epidemiológica de la en-

fermedad. En contraste, en la rama derecha (cuando *num_num_casos* es mayor), el árbol utiliza la variable *num_num_def*, lo que sugiere que la severidad del cuadro clínico se convierte en el factor determinante para la clasificación.

En los nodos de niveles más profundos se observa la aparición de otras características como *num_anio* y el dummy *cat_provincia_SS*, que si bien tienen menor importancia global, contribuyen a refinar las predicciones en subconjuntos de datos específicos. Este comportamiento también es visible en el análisis de importancia de características, donde el peso de *num_num_casos* es seguido por variables asociadas a desenlaces graves como *num_num_def* y *num_num_uci*.

En la figura 14 se pudo observar la estructura resultante del árbol entrenado con el mayor porcentaje de datos envenenados, esto es con datos envenenados al 5 %. Al comparar este modelo envenenado con el sin envenenar, se observa que el modelo envenenado también inicia su nodo raíz con la característica *num_num_casos*, lo que indica que esta característica sigue siendo la más importante para reducir los niveles de impureza y efectuar el primer split, incluso después de introducir datos envenenados.

Inclusive al revisar todos los experimentos, en el cuadro comparativo 14, del anexo A, se observa que esta característica se mantiene estable como la característica más importante en el nodo raíz, lo que sugiere que el envenenamiento no ha afectado la importancia relativa de esta característica en el modelo, aunque ya es posible observar cambios en el índice Gini y en la elección del umbral, cambiando de 5,5 a 7,5 en el experimento del 3 %. Esto sugiere que aunque se mantienen los valores en el mismo orden de magnitud las divisiones son suavemente menos puras, por tanto, todos estos cambios incipientes implican que este nodo no ha sido absolutamente inmune a los ataques de envenenamiento de datos.

Siguiendo con la comparativa, a partir del segundo nivel del árbol se observan cambios más notorios en la estructura, lo que comienza demostrar visiblemente el deterioro de este. Así al comparar los cambios del nodo 1, el árbol sin envenenamiento y con envenenamiento del 3 % utilizan la característica codificada *cat_grupo_edad_80+* para efectuar el split, mientras que el árbol envenenado al 1 % y 5 % utilizan la característica *num_num_casos*, reflejando una clara inestabilidad en la decisión.

Este cambio indica que el envenenamiento ha afectado la selección de características

en este nodo, generando un deterioro estructural que podría considerarse fuerte ya que cambia totalmente la lógica del modelo, lo que podría tener implicaciones para la capacidad del modelo para generalizar correctamente. Esta alteración queda reafirmada con las métricas de rendimiento obtenidas, que se presentan más adelante.

No obstante este cambio, una observación interesante es comparar la cantidad de ejemplos y el índice Gini en este nodo 1, ya que a pesar de haber hecho un cambio estructural fuerte consistente en el cambio de la característica con la cual se hace el split, la cantidad de ejemplos se mantiene igual y el índice Gini también se mantiene en valores similares, es decir, los modelos sin envenenamiento y con envenenamiento del 1 % mantienen estos valores similares entre si. Los modelos envenenados al 3 % y 5 % también mantienen estos valores similares entre si.

Otros cambios de interés son los observados en algunos nodos como el 6 que cambia la característica de selección del split, alternando entre *num_num_uci*, *num_num_semana* y *num_num_casos*, lo que indica una mayor inestabilidad en la estructura del árbol y pérdida de consistencia. Los nodos como el 8, 9, 11, 13, 27 y 28, cambian directamente la clase predicha, mostrando un claro deterioro en la capacidad clasificatoria.

Adicionalmente, al igual que se observó con el experimento de regresión, los cambios observados en el experimento de clasificación también evidencian que no siempre un mayor porcentaje de datos envenenados significa mayor deterioro del modelo. Por ejemplo, en el nodo 1, el cambio estructural más fuerte se observa en los modelos envenenados al 1 % y 5 %, mientras que el modelo al 3 % mantiene la misma característica que el modelo sin envenenamiento. Esto indica que el envenenamiento no siempre tiene un efecto lineal con el porcentaje de datos envenenados, por tanto, este segundo experimento confirmaría que el incremento del deterioro podría depender más de qué registros específicos son alterados y cómo impactan en los split del árbol. En algunos casos, un envenenamiento mínimo puede impactar directamente en registros considerados críticos por el árbol, provocando una alteración más significativa que la introducida con porcentajes mayores de envenenamiento en registros menos influyentes. Este comportamiento evidencia que la estabilidad del árbol depende no solo de la magnitud del envenenamiento, sino también de la posición estratégica de los registros alterados en relación con las divisiones clave.

La figura 15 muestra la comparación de la importancia de las características entre los

modelos envenenados al 1 %, 3 % y 5 % y el modelo *baseline*.

Es importante destacar que el enfoque de la comparativa es diferente a lo realizado hasta ahora, ya que cuando se realizó el análisis de importancia de características tal como muestra la figura 11, se identificaron las cinco características más importantes del modelo sin envenenamiento, cuyas cuatro primeras coinciden en todos los modelos de este experimento. Pero la última, *num_anio*, no se mantiene en los modelos envenenados, siendo sustituida por *provincia*.

Finalmente, podemos observar que en todos los casos, la variable *num_num_casos* se mantiene como el factor más influyente, aunque con variaciones en su peso relativo: aumenta levemente en el modelo envenenado al 1 %, se refuerza en el envenenado al 3 % y si bien en el modelo envenenado al 5 % disminuye comparativamente con el peso relativo obtenido en el modelo envenenado al 3 %, sigue siendo mayor que en el modelo baseline. Este comportamiento refleja que, en niveles bajos de perturbación, el modelo tiende a redistribuir parte de la importancia hacia otras variables, en particular las características codificadas *cat_grupo_edad_80+* y *num_num_def*, mientras que en niveles altos la estructura se hace progresivamente más dependiente de *num_num_casos*. En conjunto, estos resultados evidencian un patrón de reacomodo inicial seguido de una concentración en un único predictor, lo que reduce la diversidad de señales utilizadas por el árbol y aumenta su vulnerabilidad.

5.2.3.b Comparación de los rendimientos de los modelos

Los resultados comparativos de la figura 16 y las curvas ROC en la figura 17 permiten profundizar en el análisis del impacto del envenenamiento.

En primer lugar, se observa que la *Accuracy* desciende ligeramente en todos los escenarios de envenenamiento. Esta métrica alcanza una disminución del 0,89 % en el escenario de datos envenenados al 5 %. No obstante, el mayor cambio se observa en el escenario de datos envenenados al 3 %, con un descenso del *Accuracy* del 1,5 % del rendimiento global. La métrica *AP* muestra una mayor variabilidad: alcanza un valor superior al baseline en los escenarios de envenenamiento al 1 % (0,768) y 5 % (0,760), pero disminuye en el de 3 % de envenenamiento (0,747). En cuanto a la métrica *ROC AUC*, los valores se mantienen cercanos al baseline (0,750), con incrementos mínimos en el modelo envenenado

al 1 % (0,754), disminuciones en el modelo envenenado al 3 % (0,744) y recuperación en el envenenado al 5 % (0,750).

Las curvas ROC confirman que la capacidad discriminativa del clasificador entre clases positivas y negativas se conserva relativamente estable, sin desviaciones significativas respecto al modelo sin envenenamiento. Este comportamiento, en conjunto con las variaciones de *AP*, indica que el envenenamiento afecta de forma selectiva a las métricas, impactando con mayor fuerza aquellas sensibles al equilibrio entre clases (como *F1 (Hosp)* y *Accuracy*), mientras que otras asociadas a la discriminación global (*ROC AUC*) mantienen valores estables.

En términos generales, estos resultados sugieren que, si bien el envenenamiento no provoca un colapso inmediato del rendimiento, sí introduce inestabilidad en la consistencia de las métricas, lo que constituye una forma de degradación silenciosa pero crítica para la robustez del modelo. Por otra parte, estos resultados sugieren que el efecto del envenenamiento no es estrictamente lineal y puede inducir fluctuaciones que, en determinados contextos, incluso aparentan mejorar la capacidad predictiva del modelo.

Del análisis anterior, también podemos comprobar la necesidad de utilizar múltiples métricas para evaluar el impacto del envenenamiento, ya que cada una refleja diferentes aspectos del rendimiento, y por tanto, el análisis de métricas como las tasas de falsos positivos y falsos negativos, que se discuten a continuación, resultan cruciales para comprender las implicaciones prácticas del deterioro inducido por el envenenamiento.

El análisis comparativo de la matriz de confusión normalizada se muestra en el cuadro 8, y nos permite observar cómo el envenenamiento afecta de manera diferenciada a los aciertos y errores del modelo. En el modelo sin envenenamiento, la proporción de verdaderos positivos (VP = 0,72) y verdaderos negativos (VN = 0,67) se mantiene relativamente equilibrada, con tasas de falsos positivos (FP = 0,28) y falsos negativos (FN = 0,33) que reflejan un desempeño consistente con las métricas globales. Sin embargo, con la introducción del envenenamiento se aprecia un patrón progresivo: los VP descienden hasta 0,62 en el modelo envenenado al 5 %, mientras que los VN aumentan hasta 0,75. Esta dinámica implica que el clasificador tiende a volverse más conservador en la identificación de casos positivos, lo que conlleva una reducción de falsos negativos (FN = 0,25 en el modelo envenenado al 5 %), pero a costa de un incremento en los falsos positivos

(FP = 0,38).

Este resultado conecta con las variaciones observadas en las métricas globales. La disminución en *Accuracy* y *F1* se explica por el aumento de falsos positivos, que deteriora el balance entre ambas métricas. En contraste, la estabilidad relativa de la métrica *ROC AUC* se entiende porque el modelo mantiene una buena capacidad de discriminación global entre clases, aun cuando redistribuye los errores entre falsos positivos y falsos negativos. Finalmente, el comportamiento descrito también es coherente con el análisis de importancias: la creciente dependencia de *num_casos* hace que el modelo refuerce divisiones en torno a este predictor, pero con menor capacidad de generalización, amplificando las confusiones en escenarios de envenenamiento más severo.

Para una mejor comprensión de este fenómeno, la figura 18 presenta una comparación visual de las fronteras de decisión entre el modelo sin envenenamiento y el modelo con mayor porcentaje de datos envenenados (modelo envenenado al 5 %), ya que este escenario permite observar con mayor claridad el impacto del envenenamiento. En la figura mencionada anteriormente, el gráfico de la derecha corresponde al modelo entrenado con datos sin envenenar, el modelo baseline. El gráfico de la izquierda corresponde al modelo entrenado con datos envenenados al 5 %.

Tasa de Envenenamiento	FP	FPR	FN	FNR
0 % (sin envenenar)	15,93 %	32,24 %	14,81 %	28,65 %
5 %	11,77 %	24,53 %	19,59 %	37,62 %

Cuadro 12: Dataset de covid19: Métricas de error y tasas condicionales en las fronteras de decisión, calculadas sobre el conjunto de test.

En el caso del modelo baseline, las fronteras se mantienen más regulares, con un nivel intermedio de falsos positivos y negativos como puede verse en el cuadro 12 que resume las métricas clave para ambos modelos. Bajo el escenario del 5 % de envenenamiento, la frontera se desplaza y genera una redistribución de los errores: disminuyen los falsos positivos (FP ↓), pero aumentan los falsos negativos (FN ↑). Este incremento de FN se aprecia en la zona resaltada en la parte superior de la figura, donde varios puntos naranjas (clase positiva) quedan ahora clasificados en la región verde claro (clase negativa). Este desplazamiento evidencia visualmente cómo el envenenamiento afecta la estructura de las

regiones de clasificación y, en consecuencia, la distribución de los errores.

Este patrón confirma lo observado en la matriz de confusión (Cuadro 8), donde el clasificador se vuelve más conservador al identificar casos positivos, priorizando la reducción de predicciones erróneas de la clase negativa a costa de una mayor omisión de casos positivos reales, reduciendo los falsos positivos pero aumentando los falsos negativos.

En términos prácticos, este comportamiento implica un riesgo crítico, ya que la degradación inducida por el envenenamiento no solo reduce la capacidad de generalización, sino que además incrementa los errores más costosos en aplicaciones clínicas, como los falsos negativos.

6. Conclusiones

El objetivo más importante que este TFM ha pretendido explorar, mediante la experimentación empírica, es observar el comportamiento de un modelo basado en árboles de decisión cuando se expone a un ataque de envenenamiento de datos. El objetivo general ha sido alcanzado al demostrar que ataques simples de envenenamiento deterioran tanto métricas (Accuracy, FPR/FNR) como la lógica interna (raíz, reglas, Gini), cumpliendo el objetivo de evidenciar la degradación del modelo y de medir los cambios. Las conclusiones relacionadas con los objetivos específicos se exponen a continuación.

6.1. Evidencias sobre la vulnerabilidad, modos de envenenamiento y medida del deterioro

En ambos tipos de experimentos se observaron modificaciones significativas: alteraciones en las características seleccionadas para los *splits*, cambios en los umbrales de decisión, variaciones en el índice Gini y en la distribución de ejemplos por nodo, así como modificaciones en las predicciones de clase. Estos resultados confirman que el envenenamiento de datos puede inducir cambios estructurales profundos, alterando la lógica de decisión y la importancia relativa de las características.

La comparación entre modelos sin envenenar y modelos envenenados permitió cuantificar este deterioro a través de métricas de rendimiento.

En el experimento de regresión, incluso con solo un 1 % de envenenamiento, se evidenció un deterioro notable: el MAE aumentó alrededor de un 11 %, el MSE más de un 30 % y el coeficiente de determinación cayó aproximadamente un 25 %. En cambio, en clasificación, el impacto inicial fue menor: con un 1 % de envenenamiento la *Accuracy* descendió en torno al 0,6 %, mientras que los indicadores ROC y AP mostraron, incluso, una ligera mejora respecto al modelo entrenado sin datos envenenados.

Estos resultados reflejan que el efecto del envenenamiento no es lineal. En regresión, el mayor deterioro global se observó con el 5 %, pero entre los modelos envenenados al 1 % y 3 % hubo variaciones no monotónicas, con ligeras mejoras intermedias en MAE y MSE. Una tendencia similar se observó en clasificación, donde el peor desempeño general

se registró con el 3 % de envenenamiento. Este comportamiento sugiere que el impacto depende más de la ubicación estratégica de los registros manipulados que de la magnitud absoluta del envenenamiento.

Además, se constató que una medida global (como la *Accuracy* o *ROC*) no es siempre una métrica suficiente en clasificación. El análisis de la matriz de confusión y de las fronteras de decisión mostró que el envenenamiento desplaza los límites del modelo, volviéndolo más conservador con los positivos: se reducen los falsos positivos pero aumentan los falsos negativos. En un contexto clínico, este efecto es crítico, pues implica un mayor riesgo de no detectar pacientes que requieren hospitalización (para los efectos prácticos del experimento realizado en este TFM).

Los resultados obtenidos son consistentes con la literatura revisada en el marco teórico: los árboles de decisión son sensibles a modificaciones en los datos y presentan cambios apreciables incluso con tasas de envenenamiento bajas (1 %, 3 %, 5 %, es decir, menores al 10 %). Asimismo, se confirma que los ataques de envenenamiento de datos pueden alterar la estructura, alterando la lógica de decisión y redistribuir la importancia de las características.

Finalmente, se destaca que la transparencia de los árboles de decisión (su explicabilidad) ha sido clave para diseñar los ataques, hacer visibles sus efectos y medirlos con precisión. Por ello, también se demuestra que esa misma explicabilidad constituye también una vulnerabilidad al permitir identificar y explotar puntos críticos del modelo. Identificar las características más importantes, los nodos más vulnerables y los problemas que tenía el modelos para predecir, facilitó el diseño de los ataques más convenientes para los modelos usados (ataques featured-based y label-flipping). Finalmente, la comprensión del funcionamiento del modelo y el acceso al dataset, permitió la selección de registros más convenientes para el envenenamiento, maximizando su impacto incluso con tasas bajas.

Con todo lo anterior, se evidencia la necesidad de proteger los modelos antes incluso de su entrenamiento, es decir, garantizar la integridad de los datos desde su origen y toma importancia las recomendaciones de seguridad que se presentan más adelante.

6.2. Evaluación de los riesgos en contextos críticos

Los experimentos realizados demuestran que un modelo basado en árboles de decisión sometido a envenenamiento de datos puede sufrir cambios estructurales y deterioro en su rendimiento. La alteración de su lógica de decisión incrementa el riesgo de conclusiones erróneas o sesgadas, con consecuencias potencialmente graves en aplicaciones prácticas, lo que resalta la necesidad de controlar el entorno de entrenamiento, especialmente en sectores críticos como la salud, la seguridad o la defensa, donde las decisiones automatizadas pueden tener consecuencias directas sobre la vida de las personas.

No obstante, más allá de estas consecuencias más extremas, la manipulación de modelos predictivos en el sector público puede tener efectos sociales y económicos de gran alcance. Entre ellos se incluyen errores en la planificación presupuestaria, reasignaciones inadecuadas de recursos o restricciones en la provisión de servicios, con consecuencias para la población que pueden ir desde el aumento impositivo por la presión de aumentar la recaudación estatal o deterioro en la calidad de vida por reducción de prestaciones esenciales, debido a insuficiencia de los recursos públicos.

Ejemplos concretos ilustran estos riesgos: un modelo de predicción criminal envenenado podría conducir a una distribución ineficiente de recursos policiales; un modelo de apoyo al diagnóstico médico manipulado podría derivar en diagnósticos incorrectos y tratamientos inapropiados; y, en defensa, un sistema de predicción de amenazas comprometido podría inducir decisiones estratégicas equivocadas, adjudicación de licitaciones erróneas o incluso elección de objetivos tácticos inadecuados, con posibles pérdidas humanas.

6.3. Recomendaciones de seguridad

De la revisión de la literatura realizada en este TFM se evidencia una brecha importante: la ausencia de defensas específicas para árboles de decisión frente a ataques de envenenamiento. Esta vulnerabilidad pone de relieve la urgencia de investigar y diseñar estrategias de mitigación adaptadas a este tipo de modelos. Si bien existen enfoques generales para enfrentar ataques adversariales, estos resultan insuficientes para abordar las particularidades de los árboles de decisión o prevenir el envenenamiento desde su origen.

A continuación, se presentan recomendaciones preventivas organizadas en forma de pautas con sus reflexiones asociadas:

- **Recomendación 1: Identificar y perfilar al proveedor de los datos.**
  Reflexiones: ¿Es interno o externo a la organización? ¿Ofrece garantías o certificaciones? ¿Es reconocido y de confianza? ¿Qué medidas aplica para preservar la integridad de los datos?

- **Recomendación 2: Evitar usar datos sin verificar.**
  Reflexiones: Es esencial establecer mecanismos de validación en todas las fases del proceso. ¿Qué auditorías se realizan? ¿Cómo se transmiten y almacenan los datos? ¿Se aplican pruebas de integridad y técnicas de detección de anomalías?

- **Recomendación 3: Comprobar reproducibilidad con estudios previos.**
  Reflexiones: ¿Existen investigaciones previas con esos datos? ¿Se replican los resultados? ¿Coincide la estructura de los datos con lo descrito en la literatura? ¿Qué metodologías se han usado para validar su calidad?

- **Recomendación 4: Utilizar múltiples fuentes de datos.**
  Reflexiones: La diversificación reduce la probabilidad de que todas sean comprometidas. ¿Se integran varias fuentes? ¿Cómo se validan entre sí? ¿Qué mecanismos detectan inconsistencias?

- **Recomendación 5: Contrastar los datos con conocimiento previo del dominio.**
  Reflexiones: ¿Son coherentes los datos con la información ya conocida? ¿Se detectan anomalías o patrones inusuales? ¿Coinciden cantidad de registros y dimensionalidad con lo esperado?

- **Recomendación 6: Monitorear y auditar continuamente el modelo.**
  Reflexiones: ¿Existen umbrales de alerta para anomalías? ¿Se mantienen registros de actividad? ¿Qué acciones se ejecutan tras la detección de cambios?

Finalmente, resulta fundamental sensibilizar a los equipos de desarrollo y operaciones sobre la importancia de la seguridad en todo el ciclo de vida de los datos y los modelos. Incluir estas medidas desde las fases iniciales del desarrollo contribuye a fortalecer la resiliencia frente a ataques adversariales.

6.4. Contribución y coherencia con los ODS.

Coherente con el apartado de *Objetivos de Desarrollo Sostenible* de la Introducción, los resultados permiten *aterrizar* la contribución a los ODS en términos verificables:

- **ODS 3 (Salud y bienestar).** Se observó que el envenenamiento desplaza las fronteras de decisión y vuelve al clasificador más conservador con la clase positiva, reduciendo FP pero aumentando FN. En un contexto de salud, este patrón eleva el riesgo de *no detectar* pacientes que requieren atención, aun cuando las métricas globales puedan aparentar estabilidad. Los resultados *evidencian* que pequeñas manipulaciones en el entrenamiento pueden traducirse en decisiones sanitarias subóptimas, reforzando la necesidad de controlar el entorno de entrenamiento y de fundamentar la toma de decisiones en modelos explicables.

- **ODS 9 (Industria, innovación e infraestructura).** Se observó que incluso tasas bajas de envenenamiento (1–5 %) alteran la lógica interna del árbol (raíz, reglas, Gini), lo que permite *delimitar* requisitos de infraestructura metodológica *previos* a cualquier uso operativo: control del entorno de entrenamiento, trazabilidad del origen de los datos y sus transformaciones, etc. Al identificar puntos de fallo y *dónde* se producen los cambios estructurales, el estudio aporta criterios concretos para fortalecer la infraestructura de datos y experimentación que sostiene la innovación responsable.

- **ODS 16 (Paz, justicia e instituciones sólidas).** En sectores como seguridad y defensa, los cambios estructurales inducidos por envenenamiento incrementan el riesgo de conclusiones erróneas o sesgadas, con impacto potencial en la asignación de recursos y en decisiones estratégicas. Además, la transparencia propia de los árboles, también expone puntos críticos susceptibles de explotación, lo que subraya la necesidad de trazabilidad de datos y decisiones, así como de procedimientos de auditoría de reglas y explicaciones. Estas implicaciones se alinean con instituciones más transparentes y con mayor rendición de cuentas.

En suma, los ODS no quedan como marco declarativo, sino que orientan *requisitos de diseño* y *prácticas operativas* derivadas de la evidencia empírica.

7. Limitaciones y futuras líneas de investigación

Las principales limitaciones de este TFM se relacionan con el número reducido de *datasets*, la elección de un único algoritmo base (árbol individual) de ML sin ensembles, los tipos de ataques fueron simples comparados con otros que pueden considerarse de mayor complejidad y la limitada variedad de porcentajes de envenenamiento evaluados (1 %, 3 % y 5 %). Estas limitaciones condicionan la generalización de los resultados.

Asimismo, se optó por no implementar técnicas con componentes aleatorias, como *cross-validation*, *grid search* o *random search*. Si bien estas estrategias habrían permitido optimizar hiperparámetros y posiblemente mejorar el rendimiento de los modelos, también habrían introducido variabilidad en los resultados, dificultando la comparación directa entre los distintos escenarios de envenenamiento.

Futuras líneas de investigación podrían orientarse hacia:

- Realizar un mayor número de experimentos, incluyendo variaciones más amplias en los porcentajes de envenenamiento, para analizar si los patrones de deterioro se mantienen o emergen nuevos comportamientos.

- Incorporar técnicas de optimización como *cross-validation*, *grid search* y *random search*, evaluando si mejoran el rendimiento o el uso de ensembles para analizar robustez relativa de los modelos frente a ataques de envenenamiento.

- Extender la aplicación de ataques a otros tipos de modelos, como redes neuronales o máquinas de soporte vectorial, empleando métodos explicables que permitan comprender cómo los ataques afectan su estructura y desempeño.

- Explorar enfoques de aprendizaje federado o descentralizado que reduzcan la dependencia de un único conjunto centralizado de datos, disminuyendo así la efectividad de los ataques. En paralelo, estudiar el uso de técnicas generativas (*GANs*) para crear datos sintéticos que contribuyan a entrenar modelos más resilientes.

- Desarrollar un *framework* de defensa preventiva específico para árboles de decisión, orientado a proteger la integridad de los datos desde el origen, evitando la manipulación antes de que se materialice un ataque.

Referencias

Alcántara Suárez, E. J. (2023). *Análisis de La Aplicación de Machine Learning En Sistemas de Defensa* [Tesis de maestría, Universitat Oberta de Catalunya]. https://hdl.handle.net/10609/147218.

Alruwaili, E., & Moulahi, T. (2025). Prevention of Data Poisonous Threats on Machine Learning Models in E-Health. *ACM Transactions on Computing for Healthcare* Advance online publication https://doi.org/10.1145/3728369.

Armada. (2021). Revista General de Marina, noviembre 2021. https://armada.defensa.gob.es/archivo/rgm/2021/11/RGMNoviembre2021cap07.pdf.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, *58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Bertsimas, D., & Dunn, J. (2017). Optimal classification trees. *Machine Learning*, *106*(7), 1039-1082. https://doi.org/10.1007/s10994-017-5633-9

Biggio, B., & Roli, F. (2018). Wild Patterns: Ten Years after the Rise of Adversarial Machine Learning. *Pattern Recognition*, *84*, 317-331. https://doi.org/10.1016/j.patcog.2018.07.023

Bishop, C. M. (2009). *Pattern Recognition and Machine Learning*. Springer Science + Business Media.

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (2017). *Classification and Regression Trees*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315139470

Calzavara, S., Cazzaro, L., & Vettori, M. (2025). Timber! Poisoning Decision Trees. *Proceedings of the 3rd IEEE Conference on Secure and Trustworthy Machine Learning (SaTML 2025)*. https://doi.org/10.1109/SaTML59763.2025.10992468

Carnerero-Cano, J. (2023). *Indiscriminate Data Poisoning against Supervised Learning: General Attack Formulations, Robust Defences, and Poisonability* [Tesis doctoral, Imperial College London].

https://rissgroup.org/phd-thesis-indiscriminate-data-poisoning-against-supervised-learning-general-attack-formulations-robust-defences-and-poisonability/.

Chang, J. Y., & Im, E. G. (2020). Data Poisoning Attack on Random Forest Classification Model. *Proceedings of Smart Media & Applications 2020 (SMA 2020)*, 1-6 Awarded Best Paper (Gold).

Chen, H., Zhang, H., Boning, D., & Hsieh, C.-J. (2019). *Robust Decision Trees Against Adversarial Examples* (Preprint). arXiv. http://arxiv.org/abs/1902.10660.

Cinà, A. E., Grosse, K., Demontis, A., Vascon, S., Zellinger, W., et al. (2024). Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning. *ACM Computing Surveys*, *55*(294), 1-39. https://doi.org/10.1145/3585385

Cordero, A., Bertomeu-Gonzalez, V., Segura, J. V., Morales, J., Álvarez-Álvarez, B., Escribano, D., Rodríguez-Manero, M., Cid-Alvarez, B., García-Acuña, J. M., González-Juanatey, J. R., & Martínez-Mayoral, A. (2024). Árboles de Clasificación Obtenidos Mediante Inteligencia Artificial Para La Predicción de Insuficiencia Cardiaca Tras El Síndrome Coronario Agudo. *Medicina Clínica*, *163*(4), 167-174. https://doi.org/10.1016/j.medcli.2024.01.040

Cuenca, D. P., & Medina, D. P. (2023). *IA en la seguridad y delincuencia* [Tesis de maestría, Universitat Oberta de Catalunya]. https://hdl.handle.net/10609/149212.

Cuesta Calvo, R., Maudes Raedo, J., Díez-Pastor, J.-F., & Arjona, I. (2018). Predicción de delincuencia con datos públicos. *XVIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2018)*, 214 https://sci2s.ugr.es/caepia18/proceedings/docs/CAEPIA2018_paper_214.pdf.

Drews, S., Albarghouthi, A., & D'Antoni, L. (2020). Proving Data-Poisoning Robustness in Decision Trees. [Preprint] arXiv. https://doi.org/10.48550/arXiv.1912.00981

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* (2nd ed.). John Wiley & Sons.

El País. (2025). La Policía Nacional Deja de Usar Veripol, Su IA Estrella Para Detectar Denuncias Falsas. *19-03-2025*

https://elpais.com/tecnologia/2025-03-19/la-policia-nacional-deja-de-usar-veripol-su-ia-estrella-para-detectar-denuncias-falsas.html.

Fantozzi, P., & Naldi, M. (2024). The Explainability of Transformers: Current Status and Directions. *Computers*, *13*(4), 92. https://doi.org/10.3390/computers13040092

Gómez, J. E. L., Pedraza, J. L. P., & Martínez, A. P. (2023). Experiencia de empleado: Modelo de gestión de personas en la economía del conocimiento. *Logos Guardia Civil, Revista Científica del Centro Universitario de la Guardia Civil*, (1), 169-194.

Gómez Mármol, F., Ruipérez-Valiente, J. A., Nespoli, P., Martínez Pérez, G., Rivera Pinto, D., Larriva Novo, X. A., Álvarez-Campana, M., Villagrá González, V., Maestre Vidal, J., Rodriguez López, F. A., Páramo Castrillo, M., Rojo Lacal, J. I., & García-Abril Alonso, R. (2021). 51 COBRA: Cibermaniobras adaptativas y personalizables de simulación hiperrealista de APTs y entrenamiento en ciberdefensa usando gamificación. *Universidad de Castilla-La Mancha*, 227-230 https://hdl.handle.net/10578/28661.

Grimmelikhuijsen, S., & Meijer, A. (2022). Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response. *Perspectives on Public Management and Governance*, *5*(3), 232-242. https://doi.org/10.1093/ppmgov/gvac008

Hastie, T., Tibshirani, Robert & Friedman, H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd. ed.). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Lewis, D. A., Blum, G., & Modirzadeh, N. K. (2016). War-Algorithm Accountability [Preprint] arXiv. https://doi.org/1609.04667

Lyu, L., Yu, H., & Yang, Q. (2020). Threats to Federated Learning: A Survey. [Preprint] arXiv. https://doi.org/10.48550/arXiv.2003.02133

Mienye, I. D., & Jere, N. (2024). A Survey of Decision Trees: Concepts, Algorithms, and Applications. *IEEE Access*, *12*, 86716-86727. https://doi.org/10.1109/ACCESS.2024.3416838

Ministerio de Asuntos Económicos y Transformación Digital. (2022). *Agenda España Digital 2026.* (Sitio web). https://espanadigital.gob.es/.

Ministerio de Sanidad. (s.f.). *Ministerio de Sanidad - Profesionales - Áreas - Excelencia Clínica* (Sitio Web). https://www.sanidad.gob.es/areas/calidadAsistencial/excelenciaClinica/.

Ministerio para la Transformación Digital y de la Función Pública. (2024). *Estrategia de Inteligencia Artificial 2024* (Sitio web). https://portal.mineco.gob.es/ca-es/comunicacion/Pagines/20240514-Gobierno-aprueba-Estrategia-IA-2024.aspx.

Molina, O. A. (2024). *Regulación de La Inteligencia Artificial En La Sanidad Española* [Tesis Doctoral]. Universidad Complutense de Madrid. https://hdl.handle.net/20.500.14352/117121.

Monforte, C. C. (2023). La Inteligencia Artificial y la Investigación de delitos. *Logos Guardia Civil, Revista Científica del Centro Universitario de la Guardia Civil*, (1), 61-84.

Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E. C., & Roli, F. (2017). *Towards Poisoning of Deep Learning Algorithms with Back-gradient Optimization.* (Preprint). arXiv. http://arxiv.org/abs/1708.08689.

Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. The MIT Press.

Nijssen, S. (2008). Bayes optimal classification for decision trees. *Proceedings of the 25th International Conference on Machine Learning (ICML '08)*, (pp. 696-703). https://doi.org/10.1145/1390156.1390244

Oates, B. J. (2006). *Researching Information Systems and Computing*. SAGE Publications.

Ocaña, G. S. (2024). Contribución de la Inteligencia Artificial a la Seguridad Ciudadana. *Revista Veritas de Difusão Científica*, *5*(3), 2882-2900. https://doi.org/10.61616/rvdc.v5i3.382

Oksuz, A. C., Halimi, A., & Ayday, E. (2024). Autolycus: Exploiting Explainable AI (XAI) for Model Extraction Attacks against Interpretable Models. *Proceedings*

on Privacy Enhancing Technologies, *2024*(4), 684-699. https://doi.org/10.56553/popets-2024-0137

Ortiz de Zárate Alcarazo, L. (2022). Explicabilidad (de la inteligencia artificial). *Eunomía. Revista en Cultura de la Legalidad*, 22, 328-344. https://doi.org/10.20318/eunomia.2022.6819

OWASP. (2024). *OWASP Top 10 for LLM Applications 2025* (Sitio Web). https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/.

Pandya, R., & Pandya, J. (2015). C5. 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. *International Journal of Computer Applications*, *117*(16), 18-21. https://doi.org/10.5120/20639-3318

Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). SoK: Security and Privacy in Machine Learning. *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, IEEE.(pp. 399-414). https://doi.org/10.1109/EuroSP.2018.00035

Phillips, P. J., Hahn, C., Fontana, P., Yates, A., Greene, K. K., Broniatowski, D. A., & Przybocki, M. A. (2021, septiembre). *Four Principles of Explainable Artificial Intelligence* (NIST 8312). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8312

Pinto, I., Olazarán, Á., Jurío, D., de la Osa, B., Sainz, M., Oscoz, A., Ballaz, J., Gorricho, J., Galar, M., & Andonegui, J. (2024). *Improving Diabetic Retinopathy Screening Using Artificial Intelligence: Design, Evaluation and before-and-after Study of a Custom Development* (Preprint). arXiv. https://doi.org/10.48550/arXiv.2412.14221.

Policia Nacional. (2018). *Detalle Nota de Prensa.* (Comunicado de Prensa). https://www.policia.es/_es/comunicacion_prensa_detalle.php?ID=4433&idiomaActual=es.

Preece, A., Harborne, D., Braines, D., Tomsett, R., & Chakraborty, S. (2018). *Stakeholders in Explainable AI* (Preprint). arXiv. https://doi.org/10.48550/arXiv.1810.00184.

Quijano-Sánchez, Lara, Liberatore, Federico, Camacho-Collados, José & Camacho-Collados, Miguel. (2018). Applying automatic text-based detection of deceptive language to police reports: Extracting behavioral patterns from a multi-step classification

model to understand how we lie to the police. *Knowledge-Based Systems*, *149*, 155-168. https://doi.org/10.1016/j.knosys.2018.03.010

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81-106. https://doi.org/10.1007/BF00116251

Ramirez, M. A., Kim, S.-K., Hamadi, H. A., Damiani, E., Byon, Y.-J., Kim, T.-Y., Cho, C.-S., & Yeun, C. Y. (2022). *Poisoning Attacks and Defenses on Artificial Intelligence: A Survey* (Preprint). arXiv. https://doi.org/10.48550/arXiv.2202.10276.

Roldán Tudela, J. M. (2017). Desafíos Éticos En El Uso Militar de La Inteligencia Artificial. En *Documentos de Seguridad y Defensa 79: La Inteligencia Artificial, Aplicada a La Defensa* (pp. 142-143). Instituto Español de Estudios Estratégicos, Ministerio de Defensa. https://dialnet.unirioja.es/servlet/libro?codigo=731297.

Rudin, C. (2019). *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead* (Preprint). arXiv. https://doi.org/10.48550/arXiv.1811.10154.

Russell, S. J., & Norvig, P. (2004). *Inteligencia Artificial: Un Enfoque Moderno* (2.ª ed.). Pearson Educación.

Salzberg, S. L. (1994). C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning*, *16*(3), 235-240. https://doi.org/10.1007/BF00993309

Subdirección General de Planificación y Gestión de Infraestructuras y Medios de Seguridad. (s.f.). *Sistema Automático de Identificación Dactilar (SAID) - ABIS* (Sitio web). https://fondoseuropeosparaseguridad.interior.gob.es/es/detalle/proyecto/SISTEMA-AUTOMATICO-DE-IDENTIFICACION-DACTILAR-SAID-ABIS/?

Sun, T., Chen, H., Hu, G., & Zhao, C. (2025). Explainability-Based Knowledge Distillation. *Pattern Recognition*, *159*, 111095. https://doi.org/10.1016/j.patcog.2024.111095

Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2018). Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation. *Proceedings of the 2018 AAAI/ACM*

Conference on AI, Ethics, and Society, (pp. 303-310). https://doi.org/10.1145/3278721.3278725

Unión Europea. (2024). *Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de 13 de junio de 2024 por el que se establecen normas armonizadas en materia de inteligencia artificial y se modifican determinados actos legislativos de la Unión* (Diario Oficial de la Unión Europea, L 168). https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32024R1689.

Unión Europea. (2025). *Reglamento (UE) 2025/327 del Parlamento Europeo y del Consejo, de 11 de febrero de 2025, relativo al Espacio Europeo de Datos de Salud, y por el que se modifican la Directiva 2011/24/UE y el Reglamento (UE) 2024/2847* (Diario Oficial de la Unión Europea). http://data.europa.eu/eli/reg/2025/327/oj/spa.

Vassilev, A. (2025). *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* (NIST AI 100-2e2025). National Institute of Standards and Technology. Gaithersburg, MD. https://doi.org/10.6028/NIST.AI.100-2e2025

Wan, Y., Qu, Y., Ni, W., Xiang, Y., Gao, L., & Hossain, E. (2023). *Data and Model Poisoning Backdoor Attacks on Wireless Federated Learning, and the Defense Mechanisms: A Comprehensive Survey* (Preprint). arXiv. https://doi.org/10.48550/arXiv.2312.08667.

Acrónimos

- **IA** - Inteligencia Artificial
- **ML** - Machine Learning
- **TFM** - Trabajo de Fin de Máster
- **DNN** - Deep Neural Networks
- **CNN** - Convolutional Neural Networks
- **RNN** - Recurrent Neural Networks
- **SVM** - Support Vector Machines
- **ODS** - Objetivos de Desarrollo Sostenible
- **XAI** - Explainable Artificial Intelligence
- **EDA** - Exploratory Data Analysis
- **LIME** - Local Interpretable Model-agnostic Explanations
- **SHAP** - SHapley Additive exPlanations
- **MALE** - Mando de Apoyo Logístico del Ejército
- **DIAD** - Dirección de Adquisiciones
- **CART** - Classification and Regression Trees
- **CRISP-DM** - Cross Industry Standard Process for Data Mining

Anexos

A. Anexo: Cambios en la estructura en los árboles, según porcentaje de envenenamiento

Experimento de Regresión

Nodo / Nivel	0 %	1 % (cr)	1 %	3 %	5 %
Raíz (nodo 0)					
Característica:	PATRIMONIO	PATRIMONIO	PATRIMONIO	PATRIMONIO	PATRIMONIO
Threshold:	0,5	0,5	0,5	0,5	0,5
Samples:	2.688	2.688	2.714	2.768	2.822
Squared error:	1.530374e+09	1.535982e+09	1.651157e+09	1.478535e+09	1.535246e+09
Value:	9.221,463170	9.295,952009	9.455,829772	9.234,269870	9.196,032247
Nodo 1					
Característica:	CATALUÑA	CATALUÑA	MADRID	C. LA LIBERTAD	C. LA LIBERTAD
Threshold:	0,5	0,5	0,5	0,5	0,5
Samples:	2.454	2.454	2.485	2.529	2.585
Squared error:	3.032758e+07	3.059474e+07	3.038398e+07	3.095521e+07	2.847380e+07
Value:	2.325,817848	2.353,250611	2.319,868813	2.361,159747	2.314,753965
Nodo 2					
Característica:	MADRID	MADRID	CATALUÑA	MADRID	CATALUÑA
Threshold:	0,5	0,5	0,5	0,5	0,5
Samples:	2.323	2.323	2.369	2.286	2.346
Squared error:	2.662868e+07	2.692440e+07	2.713773e+07	2.947659e+07	2.677087e+07
Value:	2.092,433922	2.121,413689	2.125,354158	2.073,790901	2.016,586104
Nodo 3					
Característica:	C. VALENCIANA	C. VALENCIANA	C. VALENCIANA	C. VALENCIANA	C. VALENCIANA
Threshold:	0,5	0,5	0,5	0,5	0,5
Samples:	2.201	2.201	2.248	2.180	2.231
Squared error:	2.269439e+07	2.301989e+07	2.460622e+07	2.604955e+07	2.400772e+07
Value:	1.872,619718	1.903,205816	1.921,445285	1.904,798624	1.849,499328
Nodo 4					
Característica:	Hoja	Hoja	Hoja	Hoja	Hoja
Samples:	2.084	2.084	2.128	2.070	2.127
Squared error:	1.955595e+07	1.991353e+07	2.145241e+07	2.321450e+07	2.122172e+07
Value:	1.658,158829	1.690,462092	1.711,578477	1.728,533333	1.680,198402
Nodo 5					
Característica:	Hoja	Hoja	Hoja	Hoja	Hoja
Samples:	117	117	120	110	104
Squared error:	6.318478e+07	6.318478e+07	6.590214e+07	6.781291e+07	6.841146e+07
Value:	5.692,589744	5.692,589744	5.643,083333	5.221,790909	5.312,028846
Nodo 6					
Característica:	L. ESPECIAL	L. ESPECIAL	C. LA LIBERTAD	L. ESPECIAL	ADMÓN. PÚBLICA
Threshold:	0,5	0,5	0,5	0,5	0,5
Samples:	122	122	121	106	115

Nodo / Nivel	0 %	1 % (cr)	1 %	3 %	5 %
Squared error:	8.100897e+07	8.100897e+07	5.904551e+07	8.729090e+07	6.932718e+07
Value:	6.058,098361	6.058,098361	5.913,677686	5.549,292453	5.258,069565
Nodo 7					
Característica:	Hoja	Hoja	Hoja	Hoja	Hoja
Samples:	111	111	110	95	102
Squared error:	8.505852e+07	8.505852e+07	5.249850e+07	9.344338e+07	7.428021e+07
Value:	6.657,018018	6.657,018018	4.958,454545	6.189,926316	5.920,578431
Nodo 8					
Característica:	Hoja	Hoja	Hoja	Hoja	Hoja
Samples:	11	11	11	11	13
Squared error:	7.024793e+01	7.024793e+01	2.414592e+07	5.915702e+01	3.569941e+02
Value:	14,454545	14,454545	15.465,909091	16,545455	59,923077
Nodo 9					
Característica:	C. LA LIBERTAD	C. LA LIBERTAD	L. ESPECIAL	CATALUÑA	CATALUÑA
Threshold:	0,5	0,5	0,5	0,5	0,5
Samples:	131	131	116	243	239
Squared error:	7.782585e+07	7.782585e+07	8.012710e+07	3.677989e+07	3.575088e+07
Value:	6.464,374046	6.464,374046	6.292,327586	5.064,555556	5.241,539749
Nodo 10					
Característica:	L. ESPECIAL	L. ESPECIAL	O. I. PENALES	MADRID	MADRID
Threshold:	0,5	0,5	0,5	0,5	0,5
Samples:	120	120	105	231	228
Squared error:	7.360916e+07	7.360916e+07	8.396076e+07	3.133401e+07	3.109157e+07
Value:	5.577,241667	5.577,241667	6.949,952380	4.512,277056	4.750,736842
Nodo 11					
Característica:	Hoja	Hoja	Hoja	Hoja	Hoja
Samples:	108	108	93	217	216
Squared error:	7.802882e+07	7.802882e+07	8.957711e+07	2.823898e+07	2.847932e+07
Value:	6.190,342593	6.190,342593	7.721,806452	4.090,336406	4.388,365741
Nodo 12					
Característica:	Hoja	Hoja	Hoja	Hoja	Hoja
Samples:	12	12	12	14	12
Squared error:	1.909889e+03	1.909889e+03	3.422041e+04	3.377490e+07	3.320314e+07
Value:	59,333333	59,333333	968,083333	11.052,357143	11.273,416667
Nodo 13					
Característica:	Año_2023	Año_2023	Año_2023	Año_2023	Año_2023
Threshold:	0,5	0,5	0,5	0,5	0,5
Samples:	11	11	11	12	11
Squared error:	2.158062e+07	2.158062e+07	6.727273e+01	2.271570e+07	.384264e+07
Value:	16.142,181818	16.142,181818	15,000000	15.695,916667	15.414,545455
Nodo 14					
Característica:	Hoja	Hoja	Hoja	Hoja	Hoja
Samples:	10	10	10	11	10
Squared error:	1.806820e+07	1.806820e+07	3.000000e+01	1.901311e+07	.934884e+07
Value:	15.424,200000	15.424,200000	13,000000	15.002,636364	14.623,800000
Nodo 15					
Característica:	Hoja	Hoja	Hoja	Hoja	Hoja
Samples:	1	1	1	1	1
Squared error:	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Value:	23.322,000000	23.322,000000	35,000000	23.322,000000	23.322,000000
Nodo 16					

Nodo / Nivel	0 %	1 % (cr)	1 %	3 %	5 %
Característica:	CATALUÑA	CATALUÑA	CATALUÑA	CATALUÑA	ATALUÑA
Threshold:	0.5	0.5	0.5	0.5	0.5
Samples:	234	234	229	239	237
Squared error:	1.153338e+10	1.151655e+10	1.269010e+10	1.100694e+10	1.182010e+10
Value:	81.537,333333	82.105,307692	86.891,912664	81.962,702929	84.251,324895
Nodo 17					
Característica:	MADRID	MADRID	MADRID	MADRID	MADRID
Threshold:	0,5	0,5	0,5	0,5	0,5
Samples:	221	221	215	229	223
Squared error:	6.913130e+09	6.915767e+09	7.812134e+09	7.458010e+09	6.973622e+09
Value:	64.513,461538	65.114,846154	68.356,576744	69.095,292576	66.215,156951
Nodo 18					
Característica:	C. VALENCIANA	C. VALENCIANA	C. VALENCIANA	C. VALENCIANA	C. VALENCIANA
Threshold:	0,5	0,5	0,5	0,5	0,5
Samples:	211	211	202	217	210
Squared error:	4.248984e+09	4.266296e+09	4.270792e+09	4.425107e+09	3.401863e+09
Value:	52.950,194313	53.580,080569	52.797,000000	55.732,124424	51.018,271429
Nodo 19					
Característica:	Hoja	Hoja	Hoja	Hoja	Hoja
Samples:	199	199	190	203	199
Squared error:	3.366742e+09	3.395661e+09	3.308353e+09	3.394973e+09	2.474092e+09
Value:	45.023,135678	45.691,005025	44.360,205263	46.590,975369	43.495,944724
Nodo 20					
Característica:	Hoja	Hoja	Hoja	Hoja	Hoja
Samples:	12	12	12	14	11
Squared error:	5.564751e+08	5.564751e+08	5.380945e+08	5.818063e+08	6.430892e+08
Value:	184.407,250000	184.407,250000	186.379,583333	188.278,785714	187.104,000000
Nodo 21					
Característica:	Año_2020	Año_2020	Año_2020	Año_2020	Año_2020
Threshold:	0.5	0.5	0.5	0.5	0.5
Samples:	10	10	13	12	13
Squared error:	7.766844e+08	7.766844e+08	6.236246e+08	6.787625e+08	6.761729e+08
Value:	308.498,400000	308.498,400000	310.128,461538	310.745,916667	311.703,307692
Nodo 22					
Característica:	Hoja	Hoja	Hoja	Hoja	Hoja
Samples:	9	9	12	11	12
Squared error:	3.377862e+08	3.377862e+08	2.721071e+08	2.889980e+08	3.098009e+08
Value:	315.745,444444	315.745,444444	315.699,583333	316.879,636364	317.405,666667
Nodo 23					
Característica:	Hoja	Hoja	Hoja	Hoja	Hoja
Samples:	1	1	1	1	1
Squared error:	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Value:	243.275,000000	243.275,000000	243.275,000000	243.275,000000	243.275,000000
Nodo 24					
Característica:	Año_2020	Año_2020	Año_2020	Año_2020	Año_2020
Threshold:	0.5	0.5	0.5	0.5	0.5
Samples:	13	13	14	10	14
Squared error:	1.395103e+09	1.395103e+09	1.300110e+09	.659159e+09	1.300110e+09
Value:	370.943,153846	370.943,153846	371.541,714286	376.626,400000	371.541,714286
Nodo 25					
Característica:	Año_2023	Año_2023	Año_2021	Año_2021	Año_2021

Nodo / Nivel	0 %	1 % (cr)	1 %	3 %	5 %
Threshold:	0,5	0,5	0,5	0,5	0,5
Samples:	12	12	13	9	13
Squared error:	1.064323e+09	1.064323e+09	9.829011e+08	1.129440e+09	9.829011e+08
Value:	376.807,250000	376.807,250000	377.000,769231	385.076,666660	377.000,769231
Nodo 26					
Característica:	Hoja	Hoja	Hoja	Hoja	Hoja
Samples:	11	11	12	8	12
Squared error:	8.695199e+08	8.695199e+08	7.984521e+08	7.231714e+08	7.984521e+08
Value:	371.878,090909	371.878,090909	381.527,250000	392.875,875000	381.527,250000
Nodo 27					
Característica:	Hoja	Hoja	Hoja	Hoja	Hoja
Samples:	1	1	1	1	1
Squared error:	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Value:	431.028,000000	431.028,000000	322.683,000000	322.683,000000	322.683,000000
Nodo 28					
Característica:	Hoja	Hoja	Hoja	Hoja	Hoja
Samples:	1	1	1	1	1
Squared error:	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Value:	300.574,000000	300.574,000000	300.574,000000	300.574,000000	300.574,000000

Cuadro 13: Dataset de criminalidad: Comparación de nodos seleccionados del árbol de decisión bajo distintos niveles de envenenamiento.

Experimento de Clasificación

Nodo / Nivel	0 %	1 %	3 %	5 %
Raíz (nodo 0)				
Característica:	`num_casos`	`num_casos`	`num_casos`	`num_casos`
Threshold:	5,5	5,5	7,5	7,5
Samples:	240.297	240.297	240.297	240.297
Gini:	0,499139	0,499195	0,499344	0,499450
Class:	1	1	1	1
Nodo 1				
Característica:	`grupo_edad_80+`	`num_casos`	`grupo_edad_80+`	`num_casos`
Threshold:	0,5	0,5	0,5	0,5
Samples:	144.419	144.419	163.815	163.815
Gini:	0,480516	0,480518	0,486700	0,485962
Class:	0	0	0	0
Nodo 2				
Característica:	`num_casos`	`num_def`	`num_casos`	`num_def`
Threshold:	0,5	0,5	0,5	0,5
Samples:	94.537	11.342	107.712	11.342
Gini:	0,447912	0,422616	0,461821	0,422616
Class:	0	1	0	1
Nodo 3				
Característica:	`num_def`	`num_uci`	`num_def`	`num_uci`
Threshold:	0,5	0,5	0,5	0,5
Samples:	5.757	7.412	5.757	7.412
Gini:	0,394457	0,114764	0,394457	0,114764

Nodo / Nivel	0 %	1 %	3 %	5 %
Class:	1	1	1	1
Nodo 4				
Característica:	Hoja	Hoja	Hoja	Hoja
Samples:	4.314	6.586	4.314	6.586
Gini:	0,166736	0,000000	0,166736	0,000000
Class:	1	1	1	1
Nodo 5				
Característica:	Hoja	Hoja	Hoja	Hoja
Samples:	1.443	826	1.443	826
Gini:	0,315313	0,495310	0,315313	0,495310
Class:	0	0	0	0
Nodo 6				
Característica:	num_uci	num_semana	num_casos	num_semana
Threshold:	0,5	17,5	2,5	17,5
Samples:	88.780	3.930	101.955	3.930
Gini:	0,430254	0,364747	0,449478	0,364747
Class:	0	0	0	0
Nodo 7				
Característica:	Hoja	Hoja	Hoja	Hoja
Samples:	83.604	1.367	52.522	1.367
Gini:	0,410902	0,430133	0,372080	0,430133
Class:	0	0	0	0
Nodo 8				
Característica:	Hoja	Hoja	Hoja	Hoja
Samples:	5.176	2.563	49.433	2.563
Gini:	0,415010	0,321122	0,493014	0,321122
Class:	1	0	0	0
Nodo 9				
Característica:	num_casos	num_casos	num_casos	num_casos
Threshold:	2,5	2,5	2,5	2,5
Samples:	49.882	133.077	56.103	152.473
Gini:	0,499193	0,469312	0,498525	0,478099
Class:	1	0	1	0
Nodo 10				
Característica:	num_casos	grupo_edad_80+	num_casos	grupo_edad_80+
Threshold:	0,5	0,5	0,5	0,5
Samples:	32.366	79.303	32.366	79.303
Gini:	0,496269	0,422711	0,496269	0,422711
Class:	0	0	0	0
Nodo 11				
Característica:	Hoja	Hoja	Hoja	Hoja
Samples:	5.585	52.522	5.585	52.522
Gini:	0,447078	0,372080	0,447078	0,372080
Class:	1	0	1	0
Nodo 12				
Característica:	Hoja	Hoja	Hoja	Hoja
Samples:	26.781	26.781	26.781	26.781
Gini:	0,485165	0,485165	0,485165	0,485165
Class:	0	0	0	0
Nodo 13				
Característica:	anio	grupo_edad_80+	anio	grupo_edad_80+

Nodo / Nivel	0 %	1 %	3 %	5 %
Threshold:	2020,5	0,5	2020,5	0,5
Samples:	17.516	53.774	23.737	73.170
Gini:	0,462452	0,499446	0,469704	0,499950
Class:	1	0	1	0
Nodo 14				
Característica:	Hoja	Hoja	Hoja	Hoja
Samples:	3.740	36.258	5.052	49.433
Gini:	0,354142	0,485029	0,393463	0,494128
Class:	1	0	1	0
Nodo 15				
Característica:	Hoja	Hoja	Hoja	Hoja
Samples:	13.776	17.516	18.685	23.737
Gini:	0,479638	0,467233	0,482346	0,481018
Class:	1	1	1	1
Nodo 16				
Característica:	`num_def`	`num_def`	`num_def`	`num_def`
Threshold:	0,5	0,5	0,5	0,5
Samples:	95.878	95.878	76.482	76.482
Gini:	0,419453	0,420845	0,392750	0,392750
Class:	1	1	1	1
Nodo 17				
Característica:	`num_uci`	`num_uci`	`num_uci`	`num_uci`
Threshold:	0,5	0,5	0,5	0,5
Samples:	69.031	69.031	53.263	53.263
Gini:	0,463829	0,464426	0,446376	0,446376
Class:	1	1	1	1
Nodo 18				
Característica:	`grupo_edad_80+`	`grupo_edad_80+`	`grupo_edad_80+`	`grupo_edad_80+`
Threshold:	0,5	0,5	0,5	0,5
Samples:	61.180	61.180	46.640	46.640
Gini:	0,479454	0,479724	0,466691	0,466691
Class:	1	1	1	1
Nodo 19				
Característica:	Hoja	Hoja	Hoja	Hoja
Samples:	46.090	46.090	35.527	35.527
Gini:	0,496305	0,496174	0,489660	0,489660
Class:	1	1	1	1
Nodo 20				
Característica:	Hoja	Hoja	Hoja	Hoja
Samples:	15.090	15.090	11.113	11.113
Gini:	0,343586	0,349169	0,305621	0,305621
Class:	1	1	1	1
Nodo 21				
Característica:	`num_casos`	`num_casos`	`num_casos`	`num_casos`
Threshold:	13,5	13,5	22,5	22,5
Samples:	7.851	7.851	6.623	6.623
Gini:	0,191691	0,198850	0,166994	0,166994
Class:	1	1	1	1
Nodo 22				
Característica:	Hoja	Hoja	Hoja	Hoja
Samples:	3.450	3.450	3.767	3.767

Nodo / Nivel	0 %	1 %	3 %	5 %
Gini:	0,268011	0,282009	0,219596	0,219596
Class:	1	1	1	1
Nodo 23				
Característica:	Hoja	Hoja	Hoja	Hoja
Samples:	4.401	4.401	2.856	2.856
Gini:	0,124287	0,124287	0,090703	0,090703
Class:	1	1	1	1
Nodo 24				
Característica:	`provincia_SS`	`provincia_SS`	`provincia_SS`	`provincia_SS`
Threshold:	0,5	0,5	0,5	0,5
Samples:	26.847	26.847	23.219	23.219
Gini:	0,224869	0,229820	0,200212	0,200212
Class:	1	1	1	1
Nodo 25				
Característica:	`num_def`	`num_def`	`provincia_GI`	`provincia_GI`
Threshold:	1,5	1,5	0,5	0,5
Samples:	26.016	26.016	22.533	22.533
Gini:	0,207571	0,212722	0,181610	0,181610
Class:	1	1	1	1
Nodo 26				
Característica:	Hoja	Hoja	Hoja	Hoja
Samples:	14.815	14.815	22.077	22.077
Gini:	0,285347	0,289309	0,167805	0,167805
Class:	1	1	1	1
Nodo 27				
Característica:	Hoja	Hoja	Hoja	Hoja
Samples:	11.201	11.201	456	456
Gini:	0,086268	0,093222	0,499529	0,499529
Class:	1	1	0	0
Nodo 28				
Característica:	`anio`	`anio`	`anio`	`anio`
Threshold:	2021,5	2021,5	2021,5	2021,5
Samples:	831	831	686	686
Gini:	0,499739	0,499837	0,499996	0,499996
Class:	1	1	0	0
Nodo 29				
Característica:	Hoja	Hoja	Hoja	Hoja
Samples:	373	373	279	279
Gini:	0,273588	0,280747	0,275382	0,275382
Class:	1	1	1	1
Nodo 30				
Característica:	Hoja	Hoja	Hoja	Hoja
Samples:	458	458	407	407
Gini:	0,371703	0,371703	0,392179	0,392179
Class:	0	0	0	0

Cuadro 14: Dataset de covid19: Comparación de nodos seleccionados del árbol de decisión bajo distintos niveles de envenenamiento.

B. Anexo: Representación matemática de los árboles de decisión

Desde una perspectiva más formal, un árbol de decisión se construye dividiendo el conjunto de datos en subconjuntos de forma recursiva. En cada nodo, el algoritmo selecciona un atributo que **maximiza una métrica de ganancia**, de forma que la partición resultante sea lo más "pura" posible.

Las **métricas de ganancia** se utilizan para evaluar que tan homogéneos o "puros" son los subconjuntos de datos resultantes después de una división del árbol de decisión.

Si se consideran los modelos de árboles de decisión CART, ese encuentra una representación formal de un árbol de regresión definido en el trabajo de Kevin Murphy como (Murphy, 2022):

$$f(x;\theta) = \sum_{j=1}^{J} w_j \mathbb{I}(x \in R_j) \tag{9}$$

donde:

- $f(x;\theta)$ es la función de predicción del modelo
- w_j es el valor asignado a la hoja j
- $\mathbb{I}(x \in R_j)$ es una función indicadora que toma el valor 1 si la entrada x pertenece a la región R_j, y 0 en caso contrario.
- R_j es el conjunto de nodos hoja del árbol
- J es el número total de nodos hoja en el árbol

Esta misma representación se aplica a árboles de clasificación, donde en lugar de asignar un valor continuo, se asigna una clase discreta a cada hoja del árbol.

De este modo, la diferencia se produce en el tipo de salida que se espera del modelo:

- Para clasificación, w_j es una clase discreta asignada a la hoja j.
- Para regresión, w_j es un valor continuo que representa la media de los valores de salida en la hoja j.

No obstante, algunos autores prefieren utilizar una notación más explícita para los árboles de decisión, donde se define la función de predicción como (Nijssen, 2008):

$$f(x) = \arg\max_{c} P(y = c \mid x; \theta) \tag{10}$$

C. Anexo: Métricas comunes en árboles de decisión

Índice de Gini

Esta métrica es comúnmente utilizada en algunos algoritmos de árboles de decisión para clasificación y regresión (Classification and Regression Trees - CART).

Se puede volver al trabajo de Kevin Murphy para observar la representación estándar del índice Gini para un nodo i (Murphy, 2022):

$$\text{Gini}_i = 1 - \sum_{c=1} \hat{\pi}_{ic}^2 \tag{11}$$

donde:

- $\hat{\pi}_{ic}$ es la proporción de ejemplos en el nodo i que pertenecen a la clase c.
- $1 - \hat{\pi}_{ic}$ es la proporción de ejemplos que no pertenecen a la clase c.

El índice de Gini toma valores entre 0 y 1, donde 0 indica que todos los ejemplos en el nodo pertenecen a la misma clase (pureza máxima) y 1 indica una distribución uniforme entre las clases (impureza máxima).

Entropía

Otra de las métricas más comunes es la **impureza de la entropía** (o impureza de la información), que mide la incertidumbre de un conjunto de datos, representada como (Duda et al., 2001):

$$i(N) = -\sum_j P(w_j) \log_2 P(w_j) \tag{12}$$

donde:

- $i(N)$ es la impureza de la entropía del nodo N.
- $P(w_j)$ es la proporción de ejemplos en el nodo que pertenecen a la clase j.
- j es el número de clases posibles.

Ganancia de información

La ganancia de información, es otra métrica usada como criterio como por ejemplo modelos ID3 y C4.5, y se obtiene con base a la medida de la entropía, tratada anteriormente.

Esta métrica mide la reducción de la incertidumbre al dividir un conjunto de datos en función de un atributo, para este caso A y un conjunto de datos S (número de ejemplos en el nodo padre). Matemáticamente, se define como (Mienye & Jere, 2024):

$$IG(S,A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \tag{13}$$

donde:

- $IG(S,A)$ es la ganancia de información al dividir el conjunto de datos S según el atributo A.
- $H(S)$ es la entropía del conjunto de datos original o padre (número de ejemplos en el nodo padre) S.
- S_v es el subconjunto de datos que tiene el valor v (número de ejemplos en el nodo hijo) para el atributo A.
- $H(S_v)$ es la entropía del subconjunto de datos hijo.

Otras medidas de pureza

Hay métricas menos comunes que las anteriores, pero también son útiles:

- **Índice de Clasificación Errónea**: Mide la proporción de ejemplos mal clasificados.
- **Índice de Deviación**: Se utiliza principalmente en árboles de regresión y se basa en minimizar la desviación o log-verosimilitud.
- **Varianza**: Mide la dispersión de los valores de salida en nodos hoja.

D. Anexo: Variantes en árboles de decisión

Dentro de los llamados árboles axis-aligned univariados se encuentran los referentes clásicos (Duda et al., 2001):

- **ID3** (Iterative Dichotomiser 3): Introducido por Quinlan en 1986, utiliza la ganancia de información como criterio de división y detiene el crecimiento cuando las instancias quedan perfectamente separadas (Quinlan, 1986) Es uno de los primeros algoritmos de árboles de decisión.

- **C4.5**: También desarrollado por Quinlan, es una evolución directa del ID3, que introduce la métrica gain ratio, la gestión explícita de valores ausentes y permite la poda post-crecimiento de árboles para evitar el sobreajuste (Salzberg, 1994).

- **C5.0**: Una versión más avanzada y optimizada de C4.5, que incluye mejoras en la velocidad y la eficiencia (Pandya & Pandya, 2015).

- **CART** (Classification and Regression Trees): Introducido por Breiman et al., en 1986, que impone divisiones binarias, emplea el índice de Gini (clasificación) o la varianza (regresión) y utiliza la poda de complejidad-coste para equilibrar sesgo y varianza (Breiman et al., 2017).

La literatura revisada amplía esta taxonomía en dos direcciones, por un lado, surgen los árboles multivariados u oblicuos, que permiten hiperplanos de división orientados arbitrariamente en el espacio de características y permiten realizar divisiones en múltiples dimensiones simultáneamente, como el algoritmo **CHAID** (Chi-squared Automatic Interaction Detector) (Mienye & Jere, 2024); por otro, los métodos de optimización global como los Optimal Classification Trees (OCT), que formulan la inducción mediante mixed-integer optimization, garantizando árboles de tamaño mínimo para una precisión dada y demostrando que la interpretabilidad no tiene por qué sacrificar rendimiento (Bertsimas & Dunn, 2017).

Las revisiones sistemáticas más recientes coinciden en clasificar los árboles por:

- **Estrategia de partición**: (univariante vs. multivariante)

- **Función objetivo**: (impureza, distancia, error reducido)
- **Técnica de generalización**: (poda estadística, complejidad-coste, recocido simulado, MIO)
- **Contexto de uso**: Con despliegues masivos en diagnóstico médica, gestión logística y de inventarios, análisis crediticio y detección de intrusiones cibernéticas debido a su transparencia y a la facilidad de generar explicaciones de tipo contrafactual.

E. Anexo: Variantes de Ataques de Envenenamiento de Datos

Listado de ejemplos agrupados de ataques de envenenamiento de datos en diferentes tipos de modelos (Ramirez et al., 2022):

- **Ataque de manipulación de etiquetas (Label-Flipping Attacks):** Consiste en la alteración maliciosa de las etiquetas en los datos de entrenamiento, lo que puede realizarse de forma aleatoria o específica para reducir la precisión general o causar una clasificación errónea de una clase específica, respectivamente.

- **Ataques de máquinas de vectores de soporte (SVM) (Attacks on Support Vector Machines):** Estos ataques aprovechan el conocimiento previo sobre los datos de entrenamiento, los datos de validación y los hiper parámetros del algoritmo de aprendizaje SVM para maximizar la función objetivo basada en la tasa de error del clasificador.

- **Ataques a algoritmos de agrupamiento (Clustering Algorithms):** Se enfocan en interferir con el proceso de agrupamiento, a menudo insertando pequeñas muestras envenenadas entre dos clusters existentes para crear conflictos en sus límites de decisión, lo que puede llevar a una clasificación incorrecta de los datos. Este modelo solo es aplicable en escenarios de caja blanca, donde el atacante tiene acceso a los datos de entrenamiento y al modelo.

- **Ataques mediante optimización de gradiente de redes neuronales(Gradient Optimization in NN):** Utilizan la optimización de retrogradiente para realizar ataques de envenenamiento en modelos de aprendizaje profundo, lo que les permite abordar problemas de multi clases y ofrecer una generalización adecuada en diversos modelos de aprendizaje.

- **Ataques mediante GAN (Generative Adversarial Network):** Proponen el uso de GAN para generar datos de envenenamiento que maximicen el error del clasificador objetivo y logren ser indetectables. Se busca un equilibro entre detectabilidad y eficacia del ataque.

- **Ataques de Envenenamiento basados en características (Feature-Based Poisoning Attacks):** Crean muestras de entrenamiento envenenadas que son indistinguibles de las muestras originales para la inspección visual humana, preservando así la privacidad y mostrando una alta resistencia a los métodos de defensa existentes.

- **Ataques a sistemas de detección por multitud (Attacks on Crowd-Sensing Systems):** Se basan en la creación de interferencias con los datos recopilados mediante la inyección de datos falsos, aprendiendo de intentos de ataque pasados para mejorar progresivamente.

- **Ataques a modelos de agregación de datos (Attacks on Data Aggregation Models):** Se centran en manipular los resultados agregados en la salida del modelo de agregación, minimizando los parámetros de agregación del modelo y maximizando el error de los resultados agregados.

- **Ataques misceláneos:(Miscellaneous attacks)** Incluyen ataques a análisis de componentes principales (PCA) y ataques dirigidos a mecanismos de defensa específicos.

F. Anexo: Proyectos de IA en sectores críticos

Sector Defensa Nacional

Proyecto COBRA :

El Proyecto Cobra es una iniciativa española orientada al desarrollo de ciber maniobras adaptativas para la simulación realista de Amenazas Persistentes Avanzadas (APT) y el entrenamiento en ciberdefensa mediante técnicas de gamificación. Con participación de instituciones como la Universidad de Murcia, la Universidad Politécnica de Madrid e Indra, y validación en entornos militares del Mando Conjunto del Ciberespacio, este proyecto emplea inteligencia artificial con aprendizaje adaptativo para generar escenarios dinámicos, personalizados según el desempeño del usuario. Además, incorpora sistemas de telemetría y biometría para ajustar los entrenamientos en función de capacidades individuales.(Gómez Mármol et al., 2021)

Proyecto SOPRENE :

El programa I+D+i SOPRENE (Sostenimiento Predictivo basado en Redes Neuronales) ha sido fundamental en el contexto de la importancia del mantenimiento inteligente para la Armada española, destacando el mantenimiento predictivo (PdM) como una estrategia clave. Lanzado en noviembre de 2019 y concluido en febrero de 2021, SOPRENE desarrolló un demostrador tecnológico para la predicción de averías mediante Inteligencia Artificial (IA). Actualmente, SOPRENE predice fallos en propulsores de BAM y generadores de F-100, con planes de incluir más equipos. Sus capacidades predictivas se están integrando en ATAVIA, una aplicación residente en el CESADAR en tierra. Sin embargo, para superar las limitaciones de comunicación y latencia en buques navegando, se ha propuesto el programa MAPRE. MAPRE busca “miniaturizar” los algoritmos de SOPRENE para llevar la predicción a bordo de las unidades en tiempo real, sincronizándola con los sistemas en tierra, lo que es especialmente importante para submarinos S-80 (Armada, 2021).

Sector Seguridad Pública

Proyecto VeriPol

VeriPol, es una herramienta basada en técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático, capaz de identificar falsos reportes con una precisión superior al 91 %.

VeriPol se integraba con el sistema policial SIDENPOL y permitía entender patrones de engaño en los textos, proporcionando apoyo en la toma de decisiones y desalentando la presentación de reportes falsos. Se presentó como la primera herramienta validada con documentos reales, superando a modelos previos que usaban textos ficticios.(Quijano-Sánchez, Lara et al., 2018).

Según la nota de prensa del 27 de octubre de 2018 (Policia Nacional, 2018), la Policía Nacional de España implementó VeriPol en ese año, habiéndola probado en 2015 y luego en un estudio piloto en junio de 2017 en Málaga y Murcia. No obstante, en marzo de 2025 aparece un artículo en el Diario el País que anuncia que la aplicación dejó de estar operativa en octubre de 2024 (El País, 2025), en medio de cuestionamientos de la transparencia del sistema, lo que no ha podido comprobarse en esta investigación.

No obstante, la herramienta VeriPol es un ejemplo de cómo la IA puede ser utilizada para mejorar la eficiencia y efectividad de las investigaciones policiales, al mismo tiempo que plantea desafíos éticos y de transparencia que deben ser abordados.

Proyecto ABIS:

El Sistema Automático de Identificación Biométrico (ABIS) es una herramienta antes conocida como Sistema de Identificación Automática Dactilar (SAID), que permite la identificación de personas a través de sus huellas dactilares e imágenes faciales recogidas de escenarios de delitos, ya que es capaz de gestionar imágenes de reseñas dactilares, palmares, falanges, y otras; además de gestionar las imágenes faciales. Esta herramienta es utilizada por las Fuerzas y Cuerpos de Seguridad del Estado (Cuerpo Nacional de Policía y Guardia Civil) para la identificación de personas, tanto en el ámbito nacional como internacional.

Se registra que el coste de contribución de la UE a este proyecto es de 9.848.790,39

euros, financiado con fondos europeos y tiene por fecha de término el 31 de diciembre de 2025, conforme publica en la página web del Ministerio del Interior (Subdirección General de Planificación y Gestión de Infraestructuras y Medios de Seguridad, s.f.).

Sector Salud

Proyecto NaIA-RD:

El proyecto NaIA-RD, es una herramienta de IA creada a la medida por el Hospital Universitario de Navarra (HUN) de España, para asistir en el cribado de la Retinopatía Diabética (RD), que es la principal causa de pérdida de visión entre la población en edad de trabajar en países desarrollados. Esta herramienta fue implementada en julio de 2020 e integrada en el Sistema de información hospitalaria (HIS) del HUN, siendo utilizada desde entonces para el cribado rutinario de la RD (Pinto et al., 2024).

Proyecto Árboles de clasificación obtenidos mediante IA

Si bien no es un proyecto propiamente tal, se hace mención especial de la investigación sobre árboles de decisión de clasificación para la predicción de insuficiencia cardiaca tras el síndrome coronario agudo sobre pacientes provenientes de dos centros españoles entre 2006 y 2017. Este estudio resalta el valor de la IA para identificar variables relacionadas con la insuficiencia cardiaca, y sugiere que los árboles de decisión pueden ser una herramienta útil para la predicción de esta condición en pacientes con síndrome coronario agudo (Cordero et al., 2024).

Fe de errata

Se presenta el siguiente fe de erratas para subsanar errores de transcripción detectados en la memoria ya depositada. Las correcciones no alteran los resultados ni las conclusiones del trabajo.

Página 76

1. **Cuadro 15 (Matriz de confusión normalizada)**

   Dice:

Tasa de Envenenamiento	VP	VN	FP	FN
0 % (sin envenenar)	0,72	0,67	0,28	0,33
1 %	0,68	0,70	0,32	0,30
3 %	0,63	0,74	0,37	0,26
5 %	0,62	0,75	0,38	0,25

Cuadro 15: Dataset de covid19: Comparativa del rendimiento según matriz de confusion normalizada

   Debe decir:

Tasa de Envenenamiento	VP	VN	FN	FP
0 % (sin envenenar)	0,72	0,67	0,28	0,33
1 %	0,68	0,70	0,32	0,30
3 %	0,63	0,74	0,37	0,26
5 %	0,62	0,75	0,38	0,25

Cuadro 16: Dataset de covid19: Comparativa del rendimiento según matriz de confusion normalizada

   Corrección: se ha intercambiado el orden de las columnas FN/FP.

2. **Descripción de figura de fronteras de decisión**

   Dice:

   Finalmente, la figura 18 muestra la comparación de las fronteras de decisión entre el modelo *baseline* a la izquierda y el modelo entrenado con mayor porcentaje de datos envenenados, es decir, con un 5 % de envenenamiento a la derecha.

   Debe decir:

   Finalmente, la figura 18 muestra la comparación de las fronteras de decisión entre el modelo *baseline* a la **derecha** y el modelo entrenado con mayor porcentaje de datos envenenados, es decir, con un 5 % de envenenamiento a la **izquierda**.

   Corrección: se invirtieron las posiciones izquierda/derecha.

Página 92

1. **Texto explicativo de la matriz de confusión**

   Dice:

   En el modelo sin envenenamiento, la proporción de verdaderos positivos (VP = 0,72) y verdaderos negativos (VN = 0,67) se mantiene relativamente equilibrada, con tasas de falsos positivos (FP = 0,28) y falsos negativos (FN = 0,33) que reflejan un desempeño consistente con las métricas globales. Sin embargo, con la introducción del envenenamiento se aprecia un patrón progresivo: los VP descienden hasta 0,62 en el modelo envenenado al 5 %, mientras que los VN aumentan hasta 0,75. Esta dinámica implica que el clasificador tiende a volverse más conservador en la identificación de casos positivos, lo que conlleva una reducción de falsos negativos (FN = 0,25 en el modelo envenenado al 5 %), pero a costa de un incremento en los falsos positivos (FP = 0,38).

   Debe decir:

   En el modelo sin envenenamiento, la proporción de verdaderos positivos (VP = 0,72) y verdaderos negativos (VN = 0,67) se mantiene relativamente equilibrada, con tasas de falsos positivos (FP = **0,33**) y falsos negativos (FN = **0,28**) que reflejan un desempeño consistente con las métricas globales. Sin embargo, con la introducción del envenenamiento se aprecia un patrón progresivo: los VP descienden hasta 0,62 en el modelo

envenenado al 5 %, mientras que los VN aumentan hasta 0,75. Esta dinámica implica que el clasificador tiende a volverse más conservador en la identificación de casos positivos, lo que conlleva una reducción de **falsos positivos** (**FP** = 0,25 en el modelo envenenado al 5 %), pero a costa de un incremento en los **falsos negativos** (**FN** = 0,38).

Corrección: valores numéricos y falsos positivos/negativos invertidos.

Página 93

1. **Cuadro 17 (Métricas de error y tasas condicionales en fronteras de decisión)**

   Dice:

Tasa de Envenenamiento	FP	FPR	FN	FNR
0 % (sin envenenar)	15,93 %	32,24 %	14,81 %	28,65 %
5 %	11,77 %	24,53 %	19,59 %	37,62 %

Cuadro 17: Dataset de covid19: Métricas de error y tasas condicionales en las fronteras de decisión, calculadas sobre el conjunto de test.

   Debe decir:

Tasa de Envenenamiento	FP	FPR	FN	FNR
0 % (sin envenenar)	15,93 %	**33,24** %	14,81 %	**28,45** %
5 %	11,77 %	**24,55** %	19,59 %	37,62 %

Cuadro 18: Dataset de covid19: Métricas de error y tasas condicionales en las fronteras de decisión, calculadas sobre el conjunto de test.

   Corrección: ajustes menores en decimales (FPR y FNR).