9

Empresas desarrollando cumplimiento de requisitos

Guía 9. Precisión

Reglamento Europeo de Inteligencia Artificial

Financiado por la Unión Europea
NextGenerationEU

GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

Plan de Recuperación, Transformación y Resiliencia

España | digital 2026

Esta guía ha sido desarrollada en el marco del desarrollo del piloto español de sandbox regulatorio de IA, en colaboración entre los participantes, asistencias técnicas, potenciales autoridades nacionales competentes y el grupo asesor de expertos del sandbox.

La guía tiene como objetivo servir de apoyo introductorio a la normativa europea de Inteligencia Artificial y sus obligaciones aplicables. Si bien **no tiene carácter vinculante ni sustituye ni desarrolla la normativa aplicable, proporciona recomendaciones prácticas** alineadas con los requisitos regulatorios, a la espera de que se aprueben las normas armonizadas de aplicación para todos los Estados miembros.

El presente documento está sujeto a un **proceso permanente de evaluación y revisión**, con actualizaciones periódicas conforme al desarrollo de los estándares y las distintas directrices publicadas desde la Comisión Europea, y será actualizada una vez se apruebe el Ómnibus digital que modifica el Reglamento de Inteligencia Artificial.

Entre las referencias técnicas relevantes actualmente en desarrollo, destaca el **prEN 18229-2 "AI Trustworthiness Framework – Part 2: Accuracy and Robustness"**, que servirá de base para la evaluación de la precisión y la solidez de los sistemas de IA una vez sea adoptado como norma armonizada en el contexto del cumplimiento del Reglamento Europeo de Inteligencia Artificial.

Fecha de versión: 10 de diciembre de 2025

Contenido general

Índice detallado

1. Preámbulo

1.1 Análisis previo y relación de los artículos

Como hemos indicado en la introducción, la precisión en el Reglamento Europeo de IA está claramente relacionada y presente con otros ámbitos del propio sistema de IA. Es dentro del artículo 15, precisión, solidez y ciberseguridad, del Reglamento Europeo de IA, donde se aborda de manera específica los requerimientos que debe cumplir el sistema de IA, en los aspectos relativos a la precisión.

Este artículo establece los requisitos que deben cumplirse en materia de tres aspectos fundamentales *"Precisión, solidez y ciberseguridad"*. Los requisitos de ciberseguridad y solidez son tratados de manera específica en sus guías.

En esta guía vamos a realizar énfasis en los párrafos de dicho artículo que orientados específicamente a la precisión en IA.

Por ello, indicaremos una serie de medidas destinadas a que los sistemas de IA no degraden sus especificaciones de rendimiento y exactitud una vez puestos en marcha, durante todo su ciclo de vida.

Los sistemas de IA no deben de presentar problemas de funcionamiento (compatibilidad con antiguas librerías que usen o datos que procesen) ni de calidad en cuanto a la precisión de estos, conforme se usan en el tiempo. Para ello, deberán respetar niveles mínimos de precisión y/o métricas asociadas concretas a la tarea, preestablecidas, garantizando así que esta sea consistente durante todo el ciclo de vida.

Además, el Reglamento Europeo de IA indica (Art. 15, precisión, solidez y ciberseguridad Párrafo 2):

Las instrucciones sobre cómo obtener los niveles precisión propuestos, cómo usarlos e interpretarlos (guía de transparencia), sus umbrales y métricas asociadas a ella deben documentarse según la guía de documentación técnica, ver el apartado 5.

Dentro del análisis del articulado que se definen en esta guía, se puede resumir en los siguientes puntos:

- Analizar y establecer la relación entre el ciclo de vida del sistema de IA de alto riesgo y la precisión, en los puntos críticos del ciclo de vida. Este aspecto, se revisa en el apartado 4.1 Precisión y ciclo de vida. En este apartado se aborda como aspectos del ciclo de vida del sistema de IA, pueden influir en la precisión final del sistema, y como éstos deben de considerarse, con el objetivo de que la precisión sea consistente a lo largo este.
- Acorde al modelo de nuestro sistema de IA, seleccionar las métricas más adecuadas para la medida de la precisión, en el apartado 4.2, se aborda como se deben seleccionar dichas métricas.
- Además, estas métricas incluirán la selección de una función objetivo que será utilizada para alcanzar durante el entrenamiento la precisión indicada.
- Una vez seleccionadas las métricas y sus valores acorde a la finalidad prevista, el proveedor debe garantizar la precisión a lo largo del ciclo de vida de manera *consistente*

tal y como establece el Reglamento Europeo de IA. En el apartado 4.3, se proporcionan las medidas que entendemos que ayudan a cubrir ese requisito específico.

- Finalmente, todo el proceso debe estar acompañado de una correcta documentación acorde tanto a la documentación técnica (ver la propia guía de documentación técnica en detalle) como de las medidas propuestas en esta guía que deben ser adecuadamente documentadas. Abordamos en el apartado 5, cómo se propone realizar esta acción.

1.2 Contenido de los artículos en el Reglamento de IA

1.3 Objetivo del documento

Los sistemas de IA de alto riesgo lo son, fundamentalmente, debido a los riesgos potenciales para la salud, la seguridad y los derechos fundamentales que su utilización representa. Así se indica en varios puntos del Reglamento de IA, así como a lo largo de las guías que acompañan este sandbox. Una manera clave de poder mitigar al máximo esos riesgos es específicamente con la **precisión** de este sistema de IA; a través de la precisión del sistema, obtenemos una medida cuantitativa de la relación entre la **finalidad prevista** de este y su desempeño, desde el diseño hasta su funcionamiento tras la puesta en marcha del sistema de AI.

A lo largo de la guía se desarrollan una serie de medidas organizativas y técnicas destinadas, primero a seleccionar y evaluar las métricas de precisión para el sistema de inteligencia artificial. A continuación, se procede a aplicar los controles de calidad del modelo, que ayudan a verificar y validar las razones que llevan a utilizar tales métricas. También trataremos aspectos complementarios que son clave para poder implantar el sistema, pues van más allá del resultado inmediato del modelo para un marco de referencia y pueden tener implicaciones de discriminación, sesgo o imprecisión.

Es importante notar que la aplicación de los conceptos de precisión en un sistema de IA para cumplir con los requerimientos del Reglamento Europeo de IA exige, por parte de los proveedores, un conocimiento del estado del arte de estas técnicas relacionadas con la precisión y de cómo pueden aplicarse a su sistema de IA, acorde con su finalidad prevista. Por ello, acompañado por los mecanismos e información proporcionados en esta guía, los proveedores deben mantenerse atentos no solo a la evolución de la normativa, sino también a la evolución siempre rápida del estado del arte.

1.4 ¿Cómo leer esta guía?

El proceso de alcanzar la precisión adecuada del modelo pasa por una serie de pasos que hemos entendido que ayudan a cubrir los requisitos que establece el artículo 15, precisión, solidez y ciberseguridad del Reglamento Europeo de IA. Ver apartado del contenido del artículo y correspondencia del artículo con los apartados de la guía.

1.5 ¿A quién está dirigido?

Dar cabida a todas las cuestiones detalladas en esta guía es responsabilidad del proveedor del sistema de inteligencia artificial de alto riesgo, quien deberá tomar las medidas adecuadas

aquí propuestas, tanto organizativas como técnicas. Todo ello, con el objetivo de garantizar que se cumpla con los requerimientos de precisión del sistema.

Dentro de su ámbito de aplicación, el responsable del despliegue del sistema también tiene responsabilidades que se materializarán en medidas concretas, de nuevo organizativas y técnicas. La guía indicará, en cada caso, cuales le son de aplicación y el alcance de estas.

1.6 Casos de uso utilizados en la guía

A lo largo de la guía se utilizarán dos casos de uso a modo de ejemplo de cómo elaborar la documentación técnica. Estos ejemplos estarán centrados únicamente en el proveedor ya que es el responsable de generar y conservar la documentación. La descripción detallada de los casos de uso utilizados podrá encontrarse en la Guía práctica y ejemplos para entender el Reglamento de IA.

Nota: Siempre que se ponga un ejemplo, se hará de manera ilustrativa. Proveedor y responsable del despliegue han de considerar la aplicación de todas las medidas indicadas en esta guía.

Los casos de uso se han seleccionado atendiendo a su capacidad para explicar la información y procedimientos detallados en esta guía.

Los casos de uso seleccionados en este caso para la elaboración de la guía son:

- **Detección de denuncias falsas.**
- **Sistema de promoción de empleados.**

2. Introducción

2.1 ¿Qué es la precisión para IA?

En el contexto del Reglamento Europeo de IA, se indica, en su considerando (66) que la precisión está entre los requisitos fundamentales para la mitigación de los riesgos asociados al sistema de IA:

> **AI Act**
>
> (66)
>
> Deben aplicarse a los sistemas de IA de alto riesgo requisitos referentes a la gestión de riesgos, la calidad y la pertinencia de los conjuntos de datos utilizados, la documentación técnica y la conservación de registros, la transparencia y la comunicación de información a los responsables del despliegue, la supervisión humana, la solidez, la precisión y la ciberseguridad. Dichos requisitos son necesarios para mitigar de forma efectiva los riesgos para la salud, la seguridad y los derechos fundamentales […].

Aclaración sobre la nomenclatura del **inglés**: En esta guía se ha traducido el término en inglés accuracy, genéricamente, por precisión, siguiendo la traducción aportada por el Reglamento Europeo en español. Sin embargo, se han **mantenido excepciones** cuando el término accuracy forma parte de un **término compuesto** (ejemplo, balanced accuracy) en cuyo caso se hace **referencia al término compuesto en inglés** para evitar confusión y mantener así una única nomenclatura internacional como aquella usada en la mayoría de los métodos y herramientas y librerías. Es importante notar que, en ocasiones, cuando se habla del término de accuracy para referirse a la **noción de performance**, como medida de calidad del modelo más genérico, en inglés, en IA, se suele sobreentender nos referimos a accuracy, precisión, recall, etc. o la métrica asociada en particular al problema a resolver. En estos casos se ha mantenido la traducción usual en español para no causar confusión. En definitiva: accuracy se tradujo por exactitud, precisión por precisión, y performance (aunque se suele traducir por rendimiento o eficiencia fuera de campos de la IA como high performance computing o ingeniería del software), **en esta guía se ha traducido por precisión.**

Podemos considerar que la precisión, entendida como hemos indicado en la nota previa, nos permite acotar, delimitar y conocer el comportamiento del sistema de IA acorde a su finalidad prevista, los conjuntos de datos con los que va a trabajar, y la relación de la precisión con el **resto** de los requisitos, entre otros: ciberseguridad, solidez, transparencia, gobierno del dato,

supervisión. Además, la precisión es una métrica fundamental dentro del sistema de gestión de calidad que rodea al sistema de IA de alto riesgo.

El Reglamento Europeo de IA indica también, en su considerando (74) la necesidad que el nivel de precisión esté al día de los avances y estado del arte del campo y que este nivel sea mantenido durante todo el ciclo de vida del sistema de IA, desde su puesta en el mercado hasta su retirada.

AI Act

(74)

Los sistemas de IA de alto riesgo deben funcionar de manera uniforme durante todo su ciclo de vida y presentar un nivel adecuado de precisión, solidez y ciberseguridad, a la luz de su finalidad prevista y con arreglo al estado actual de la técnica generalmente reconocido. Se anima a la Comisión y las organizaciones y partes interesadas pertinentes a que tengan debidamente en cuenta la mitigación de los riesgos y las repercusiones negativas del sistema de IA. El nivel previsto de los parámetros de funcionamiento debe declararse en las instrucciones de uso que acompañen a los sistemas de IA. Se insta a los proveedores a que comuniquen dicha información a los responsables del despliegue de manera clara y fácilmente comprensible, sin malentendidos ni afirmaciones engañosas.

Cómo podemos comprobar en el considerando anterior, el Reglamento Europeo de IA se puede entender una clara relación entre la precisión y la transparencia, al relacionar ambas con la información al responsable del despliegue.

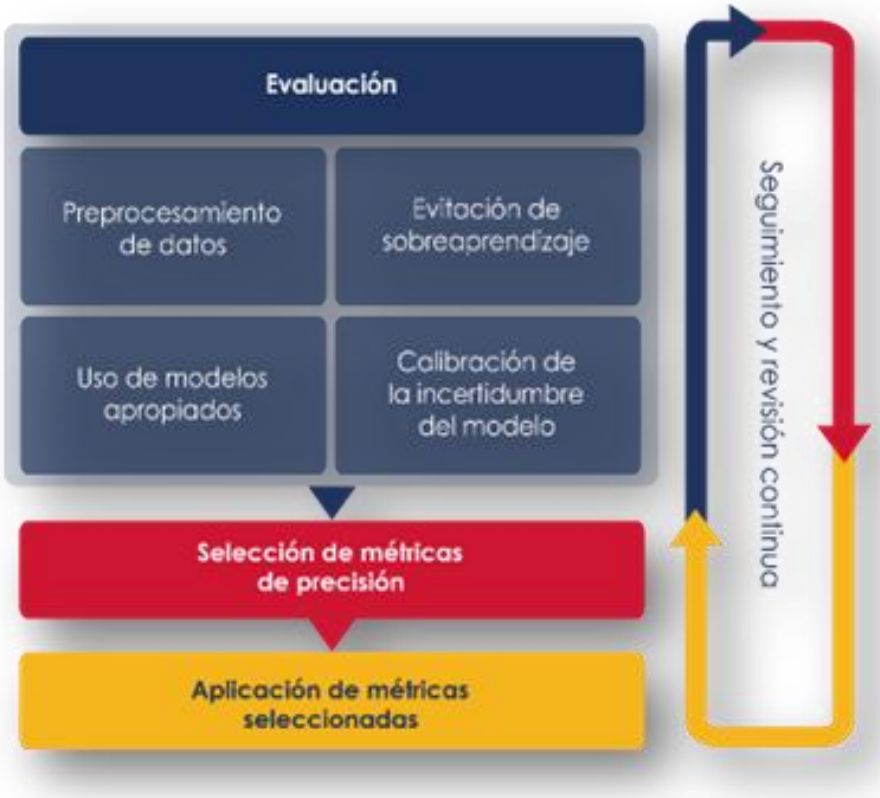
Por otro lado, la precisión se encuentra muy presente a lo largo del Anexo IV del Reglamento Europeo de IA, que explica cómo debe abordarse la documentación técnica. Sirva de ejemplo el siguiente extracto del párrafo 3 del Anexo IV, relativo a la relación entre documentación técnica y precisión.

AI Act

Anexo IV.3 – Documentación técnica a que se refiere el artículo 11, apartado 1

Información detallada acerca de la supervisión, el funcionamiento y el control del sistema de IA, en particular con respecto a sus capacidades y limitaciones de funcionamiento, incluidos los niveles de precisión para las personas o colectivos de personas específicos en relación con los que está previsto que se utilice el sistema y el nivel general de precisión esperado en relación con su finalidad prevista; […].

El proceso para alcanzar la precisión puede ser visualizado como:



3. Reglamento de Inteligencia Artificial

La puesta en servicio o la utilización de sistemas de IA de alto riesgo debe supeditarse al cumplimiento de determinados requisitos obligatorios, entre los cuales está el de precisión. Estos requisitos tienen como objetivo garantizar que los sistemas de IA de alto riesgo disponibles en la Unión o cuyos resultados de salida se utilicen en la Unión no representen riesgos inaceptables para intereses públicos importantes, reconocidos y protegidos por el Derecho de la Unión.

En este apartado se incluye los artículos referentes a la generación de precisión del Reglamento 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024 (Reglamento Europeo de Inteligencia Artificial) y se detalla en que secciones de esta guía se abordan los diferentes elementos de dichos artículos.

3.1 Análisis previo y relación de los artículos

Como hemos indicado en la introducción, la precisión en el Reglamento Europeo de IA está claramente relacionada y presente con otros ámbitos del propio sistema de IA. Es dentro del artículo 15, precisión, solidez y ciberseguridad, del Reglamento Europeo de IA, donde se aborda de manera específica los requerimientos que debe cumplir el sistema de IA, en los aspectos relativos a la precisión.

Este artículo establece los requisitos que deben cumplirse en materia de tres aspectos fundamentales *"Precisión, solidez y ciberseguridad"*. Los requisitos de ciberseguridad y solidez son tratados de manera específica en sus guías.

En esta guía vamos a realizar énfasis en los párrafos de dicho artículo que orientados específicamente a la precisión en IA.

Por ello, indicaremos una serie de medidas destinadas a que los sistemas de IA no degraden sus especificaciones de rendimiento y exactitud una vez puestos en marcha, durante todo su ciclo de vida.

Los sistemas de IA no deben de presentar problemas de funcionamiento (compatibilidad con antiguas librerías que usen o datos que procesen) ni de calidad en cuanto a la precisión de estos, conforme se usan en el tiempo. Para ello, deberán respetar niveles mínimos de precisión y/o métricas asociadas concretas a la tarea, preestablecidas, garantizando así que esta sea consistente durante todo el ciclo de vida.

Además, el Reglamento Europeo de IA indica (Art. 15, precisión, solidez y ciberseguridad Párrafo 2):

Las instrucciones sobre cómo obtener los niveles precisión propuestos, cómo usarlos e interpretarlos (guía de transparencia), sus umbrales y métricas asociadas a ella deben documentarse según la guía de documentación técnica, ver el apartado 5.

Dentro del análisis del articulado que se definen en esta guía, se puede resumir en los siguientes puntos:

- Analizar y establecer la relación entre el ciclo de vida del sistema de IA de alto riesgo y la precisión, en los puntos críticos del ciclo de vida. Este aspecto, se revisa en el apartado 4.1 Precisión y ciclo de vida. En este apartado se aborda como aspectos del ciclo de vida del sistema de IA, pueden influir en la precisión final del sistema, y como éstos deben de considerarse, con el objetivo de que la precisión sea consistente a lo largo este.
- Acorde al modelo de nuestro sistema de IA, seleccionar las métricas más adecuadas para la medida de la precisión, en el apartado 4.2, se aborda como se deben seleccionar dichas métricas.
- Además, estas métricas incluirán la selección de una función objetivo que será utilizada para alcanzar durante el entrenamiento la precisión indicada.
- Una vez seleccionadas las métricas y sus valores acorde a la finalidad prevista, el proveedor debe garantizar la precisión a lo largo del ciclo de vida de manera *consistente* tal y como establece el Reglamento Europeo de IA. En el apartado 4.3, se proporcionan las medidas que entendemos que ayudan a cubrir ese requisito específico.
- Finalmente, todo el proceso debe estar acompañado de una correcta documentación acorde tanto a la documentación técnica (ver la propia guía de documentación técnica en detalle) como de las medidas propuestas en esta guía que deben ser adecuadamente documentadas. Abordamos en el apartado 5, cómo se propone realizar esta acción.

3.2 Contenido de los artículos en el Reglamento de IA

Aclarar, que dentro del artículo 15 en el que se trata de precisión, solidez y ciberseguridad, se habla específicamente de precisión exclusivamente en los puntos uno, dos y tres.

AI Act

Art.15 – Precisión, solidez y ciberseguridad

1. Los sistemas de IA de alto riesgo se diseñarán y desarrollarán de modo que alcancen un nivel adecuado de precisión, solidez y ciberseguridad y funcionen de manera uniforme en esos sentidos durante todo su ciclo de vida.

2. Para abordar los aspectos técnicos sobre la forma de medir los niveles adecuados de precisión y solidez establecidos en el apartado 1 y cualquier otro parámetro de rendimiento pertinente, la Comisión, en cooperación con las partes interesadas y organizaciones pertinentes, como las autoridades de metrología y de evaluación comparativa, fomentará, según proceda, el desarrollo de parámetros de referencia y metodologías de medición.

3. En las instrucciones de uso que acompañen a los sistemas de IA de alto riesgo se indicarán los niveles de precisión de dichos sistemas, así como los parámetros pertinentes para medirla.

4. Los sistemas de IA de alto riesgo serán lo más resistentes posible en lo que respecta a los errores, fallos o incoherencias que pueden surgir en los propios

sistemas o en el entorno en el que funcionan, en particular a causa de su interacción con personas físicas u otros sistemas. Se adoptarán medidas técnicas y organizativas a este respecto.

La solidez de los sistemas de IA de alto riesgo puede lograrse mediante soluciones de redundancia técnica, tales como copias de seguridad o planes de prevención contra fallos.

Los sistemas de IA de alto riesgo que continúan aprendiendo tras su introducción en el mercado o puesta en servicio se desarrollarán de tal modo que se elimine o reduzca lo máximo posible el riesgo de que los resultados de salida que pueden estar sesgados influyan en la información de entrada de futuras operaciones (bucles de retroalimentación) y se garantice que dichos bucles se subsanen debidamente con las medidas de reducción de riesgos adecuadas.

5. Los sistemas de IA de alto riesgo serán resistentes a los intentos de terceros no autorizados de alterar su uso, sus resultados de salida o su funcionamiento aprovechando las vulnerabilidades del sistema.

Las soluciones técnicas encaminadas a garantizar la ciberseguridad de los sistemas de IA de alto riesgo serán adecuadas a las circunstancias y los riesgos pertinentes.

Entre las soluciones técnicas destinadas a subsanar vulnerabilidades específicas de la IA figurarán, según corresponda, medidas para prevenir, detectar, combatir, resolver y controlar los ataques que traten de manipular el conjunto de datos de entrenamiento («envenenamiento de datos»), o los componentes entrenados previamente utilizados en el entrenamiento («envenenamiento de modelos»), la información de entrada diseñada para hacer que el modelo de IA cometa un error («ejemplos adversarios» o «evasión de modelos»), los ataques a la confidencialidad o los defectos en el modelo.

3.3 Correspondencia del articulado con los apartados de la guía

En la tabla dispuesta a continuación se detalle en que secciones de esta guía se abordan los diferentes elementos de dicho artículo:

Artículo Reglamento	Requerimiento Reglamento	Sección guía
15.1	Nivel adecuado de precisión	Apartado 4.1 y Apartado 4.3
15.1	Precisión consistente a lo largo del ciclo de vida	Apartado 4.1
15.2	Aspectos técnicos de la precisión	Apartado 5
15.3	Instrucciones y niveles de precisión pertinentes	Apartado 5

4. ¿Cómo abordar los requisitos?

4.1 Precisión y ciclo de vida

Establecer y medir la precisión del sistema de IA es un proceso que abarca todo el ciclo de vida de este, tal y como hemos indicado. Existen, no obstante, puntos durante el ciclo de vida donde establecer la precisión del sistema acorde a su finalidad prevista es crítico. A continuación, para completar el proceso de trabajo descrito, presentamos estos puntos críticos, cómo estos afectan a la precisión, y cómo abordarlos.

4.1.1 Preprocesamiento de los datos

La precisión de un modelo dependerá de la calidad de los datos de entrenamiento y, por tanto, de su preprocesado. Además de las medidas de la guía de Datos, cuestiones relevantes para tener en cuenta en el preprocesamiento de los datos para garantizar la precisión del modelo son:

- Los datos usados para probar el modelo de *machine learning* deben ser representativos de la finalidad prevista que se pretende dar al sistema (ver Guía de Datos).
- Los datos de entrenamiento del modelo deben estar libres de sesgo de muestreo. Además, los datos de entrenamiento para una tarea particular no necesariamente son extensibles a otras tareas diferentes. Debe ponerse especial atención al dividir bases de datos no balanceadas para asegurar que se mantienen distribuciones similares entre datos de entrenamiento, validación y evaluación (ver Guía de Datos).
- Cuando se preprocesen datos de forma diferente, con el objetivo de encontrar la precisión adecuada a la finalidad prevista, no se podrá atribuir la diferencia en precisión al algoritmo siendo evaluado (el modelo con el objetivo o tarea final, "*downstream algorithm*"). Por tanto, para comparar la precisión entre varios modelos deberán usarse idénticas maneras de preprocesado de datos , y el mismo conjunto de evaluación (*test*) deberá ser usado para comparar dos modelos, de forma que los conjuntos de validación y evaluación nunca contendrán muestras que estén incluidas también en el conjunto de entrenamiento. Ver ISO/IEC TS 4213:2022, *Information technology – Artificial intelligence – Assessment of machine learning classification performance* [116].
- Las tareas de preentrenamiento del modelo, si difieren de las tareas del modelo final, deberán documentarse de forma precisa y reproducible igualmente, y se velará porque coincidan tanto en fase de entrenamiento, como de validación y de puesta en producción.

Más información sobre sesgos en sistemas de IA en ISO/IEC TR 24027 y en Guía de Datos.

Ejemplo - Detección de denuncias falsas

Durante la fase de diseño del sistema, se establece que la precisión seleccionada para el sistema debe de garantizarse para un amplio abanico de denuncias, y que la precisión no debe verse afectada por elementos de categorización de la denuncia en sí misma como dirección o pertenencia a minorías.

Para ello, se toma la decisión, tal y como se indica en este apartado, de realizar un trabajo sobre esos conjuntos de datos, con el objetivo de disponer de una base de datos que permita a lo largo de todo el ciclo de vida definir, establecer y documentar la métrica de precisión elegida. Para ello:

- Se recaban denuncias de todo el territorio nacional en castellano. La recopilación es representativa en equivalente proporción por cada territorio, para evitar sesgos de tamaño de población.
- Sumado al punto anterior, para mitigar el sesgo de muestreo, se realiza una mezcla aleatoria de las denuncias, que garantice una distribución uniforme a lo largo de los conjuntos de entrenamiento, prueba y validación.
- El preprocesado requerido para el proceso de anonimización, es idéntico para todos los orígenes territoriales y conjuntos de datos. Para ello, se desarrolla una pieza de software específico de preprocesado única para todos los territorios y que será utilizada para remitir los datos. De esta manera el equipo de diseño, desarrollo e implantación ya recibe datos anonimizados.

Todo este proceso, es documentado en cada uno de los pasos indicando las decisiones tomadas y la motivación.

4.1.2 Sobreaprendizaje (*Overfitting*)

Los algoritmos generativos se entrenan optimizando parámetros del modelo de aprendizaje de manera que se maximiza la probabilidad de acierto, sobre la muestra de los datos de entrenamiento disponibles y acorde a las categorías de clasificación establecidas, mientras que los discriminativos optimizan sus parámetros para maximizar la exactitud de la clasificación [116].

Para evitar el común problema del sobreaprendizaje se tomarán medidas de acorde con el tipo de modelo y finalidad prevista. En general:

- Todos los hiperparámetros deben ser reportados en los procesos de entrenamiento/prueba y validación del modelo, así como sus valores para cada modelo. El sesgo por selección de hiperparámetros debe tenerse en cuenta cuando se comparan modelos, pues diferentes modelos tienen diferentes capacidades de ajuste, por lo que el nivel de sobreaprendizaje durante entrenamiento puede diferir entre algoritmos, especialmente en aprendizaje profundo.
- Ninguna información del conjunto de datos de prueba debe ser usada cuando se ajusten los hiperparámetros, esto típicamente lleva a sobreestimar las métricas de precisión con optimismo. Cuando la información de etiqueta es necesaria para un ajuste o tuning, típicamente se usan las de un conjunto de datos separado, el conjunto de validación, el cual es disjunto del de test. Este desafío se puede enfocar, por ejemplo, con validación cruzada anidada, donde un bucle exterior

mide la precisión de la predicción mientras que el bucle interior ajusta los hiperparámetros de los modelos individuales. De esta manera, los métodos pueden elegir la configuración óptima de parámetros, para los modelos predictivos se eligen en el bucle exterior.

- Idealmente, se deben de utilizar herramientas para la automatización de la búsqueda de hiperparámetros, por ejemplo, *Hyperband, Hyperopt u Optuna*[1] (comúnmente usados en aprendizaje por refuerzo).

Ejemplo - Detección de denuncias falsas

El sistema de gestión de riesgos del proveedor para el sistema de IA ha establecido que existe un riesgo para los derechos y libertades si el sistema está sobre ajustado (*overfitting*) respecto a los sets de entrenamiento. Un *overfitting* puede causar que denuncias verdaderas sean procesadas como falsas llevando a un procedimiento de tramitación diferente, por lo que este *overfitting* pueda afectar al proceso final. De esta manera, el sobreajuste al conjunto de datos de entrenamiento, prueba y validación puede proporcionar una medida excelente de la métrica de precisión elegida, pero cuando el sistema se encuentra en producción, provocar resultados inadecuados para su finalidad prevista.

Para ello y siguiendo las indicaciones presentes en esta guía, el proveedor establece que:

- Los hiperparámetros de entrenamiento se van a almacenar asociados al código del modelo de IA, para que estos estén siempre accesibles e identificables, de tal manera que el equipo de desarrollo pueda consultar un histórico de estos y estudiar la posibilidad de *overfitting*.
- Los conjuntos de entrenamiento, prueba y validación se organizan y estructuran de manera que todos tienen una combinación aleatoria de orígenes, para que los orígenes de las denuncias anónimas ya clasificadas se encuentres uniformemente distribuidos en la muestra. Para cada denuncia se genera un identificador único (a través del cálculo de un *hash* 256 de su contenido en texto). Este sistema se utiliza para que los conjuntos de entrenamiento, prueba y validación sean completamente disjuntos y no contengan ninguna muestra en común.
- Cuando se está realizando el entrenamiento se aplica la indicación propuesta en este apartado de dos bucles de programación, que además llevan la cuenta del origen de los datos, para poder detectar intercambio de datos entre conjuntos. Así no solo se garantiza el ajuste de hiperparámetros si no que se garantiza adecuadamente que los conjuntos son disjuntos.
- Dentro del *framework* de trabajo para el sistema de IA, el equipo de desarrollo ha utilizado, combinado con el proceso descrito en los puntos anteriores, una librería específica del *framework* para el ajuste automatizado de los hiperparámetros.

[1] Las palabras y descripciones **destacadas** se corresponden con términos que son desarrollados en el glosario (ver 7.4).

4.1.3 Uso de modelos apropiados

La precisión de un modelo tendrá relevancia con respecto a cómo se posiciona con respecto al estado del arte en modelos de su categoría y/o con modelos de referencia base. Es decir, deberá compararse y establecer en contexto, las métricas de precisión seleccionadas, junto con aquellos modelos de base con los que está relacionado.

- Por reproducibilidad y sostenibilidad de la precisión del modelo, se deberá señalar o reportar estudios sobre la eliminación de los componentes y elementos (ablación en inglés) del modelo, para justificar la composición y complejidad de los componentes y elementos del modelo necesarios para alcanzar la precisión garantizada.
- En ellos, se usarán modelos de referencia base apropiados (*Baselines*). *Baselines* triviales tales como aquellos que siempre predicen la clase mayoritaria son útiles para calibrar la interpretación de la métrica (ver guía de Transparencia) pero no deben ser el único punto de comparación (ISOIEC TS 4213_2)[116].

Ejemplo - Sistema de promoción de empleados

A lo largo del diseño del sistema de IA, y considerando su finalidad prevista, el proveedor del sistema analiza los modelos de referencia base dentro del estado del arte. El objetivo de este análisis es seleccionar uno que sirva tanto de punto de partida para el diseño e implementación del sistema de IA, como también para saber cómo éste se posiciona dentro del contexto de sistemas en la misma categoría.

Para el análisis, el proveedor tiene en cuenta la finalidad prevista de establecer un mecanismo para la promoción de empleados que será el elemento último de decisión en mecanismos de promoción.

Como modelo base seleccionado para esta tarea se opta por un modelo de espacio vectorial o VSM (*vector space model*) donde los empleados serán descritos a través de las variables (coordenadas) establecidas para su evaluación. Este modelo base trivial será utilizado junto con el resto de las técnicas indicadas en la guía, para la selección de la métrica o métricas de precisión adecuadas y para la evaluación de la precisión del sistema de IA.

El proveedor registra las motivaciones y evaluaciones realizadas sobre los candidatos de modelo de base, así como la indicación de la elección final de éste mismo.

Nota Importante: El carácter técnico de esta guía hace que los ejemplos sean específicos de los casos de uso. Esto implica que las propuestas son específicas para los modelos considerados como ejemplo, y no una solución general para otros tipos de modelo, o incluso modelos de la misma tipología. Los ejemplos deben considerarse demostrativos de la operativa para obtener una conclusión técnica pero nunca como demostrativos de dicha conclusión en sí misma. Cada proveedor deberá, acorde a esta guía, establecer la medida técnica concreta para su tipo de sistema de IA y su finalidad prevista.

4.1.4 Incertidumbre y precisión

La precisión de un modelo tiene asociada un nivel de certeza o confianza que no siempre acompaña a las salidas de un modelo, si éstas no están calibradas o tienen en cuenta la incertidumbre de este. El sistema de IA puede proporcionar el resultado de su operación acompañado de un indicativo, por ejemplo, en porcentaje (o en cualquier caso normalizado) de la certeza que otorga a dicho resultado. La incertidumbre de un modelo, entendida como confianza de sus resultados, debe por tanto documentarse para garantizar la precisión y facilitar su supervisión y transparencia, según la guía de solidez.

4.2 Evaluando la precisión

Las métricas de precisión aplicables garantizadas por el proveedor en la documentación deben reflejar y ser una señal de la calidad del sistema. De otro modo, cuando no se puedan garantizar estas métricas de precisión establecidas mínimas, el proveedor proveerá mecanismos para notificar al humano de que se requiere su supervisión (según guía de supervisión humana).

En cada paso del ciclo de vida de un sistema de IA de alto riesgo, desde su diseño, desarrollo y validación, hasta su introducción al mercado, se debe establecer un nivel de precisión acorde a la finalidad prevista del sistema, estableciendo unas medidas adecuadas.

Dependiendo de la fase del ciclo de vida las medidas deberán de ser realizadas por proveedor (diseño, desarrollo) o responsable del despliegue (durante su puesta en producción o funcionamiento), evitando la degradación durante su ciclo de vida.

Proveedor

El proveedor debe tomar las siguientes medidas organizativas para asegurar que el sistema de inteligencia artificial de alto riesgo tiene un nivel de precisión adecuado. Es responsabilidad del proveedor realizar la selección la métrica (o métricas) de precisión más adecuada, desde la concepción y diseño del sistema de IA. Este proceso de selección debe tener en cuenta dos aspectos importantes:

- Las métricas deben estar relacionadas y motivadas por la finalidad prevista.
- Las métricas deben estar motivadas y permitir mitigar los riesgos para los derechos y libertades de las personas físicas, la salud y los daños a la propiedad/medio ambiente identificados en el análisis de riesgos.

Una vez establecidos los requisitos funcionales (normalmente especificados por la funcionalidad que el responsable del despliegue solicitará en los casos de uso típicos del sistema) y no funcionales (especificados por el personal técnico proveedor como características técnicas de implementación o métricas cuantificables) del sistema, una batería de pruebas significativos y representativos del dominio de aplicación. Mediante estas. se deberá reflejar para entradas compatibles, salidas esperadas dentro de los posibles rangos de salida de datos contemplados siempre en relación directa con la finalidad prevista y con la mitigación de riesgos.

La batería de pruebas puede incluir, entre otros:

- Aquellas pruebas de unidad y de stress acumulados tras el desarrollo del modelo (por ejemplo, siguiendo metodologías como *Test Driven Development*).
- Test de integración para el encale de los requisitos funcionales y no funcionales con la finalidad prevista, y que tengan como pieza fundamental la precisión establecida.
- Seguimiento de la evolución de la precisión del modelo cuando está en producción.
- La degradación del modelo podrá ser monitorizada con paneles de monitorización y herramientas de visualización de la precisión. En la guía indicaremos otras posibles métricas de calidad del modelo asociadas como son las funciones objetivo. Este aspecto está relacionado con la información presentada en la guía de solidez.

Como complemento a estas medidas organizativas, el proveedor deberá alinear la precisión del sistema de IA con las siguientes medidas técnicas:

- Proveedores, diseñadores y desarrolladores deberán proporcionar documentación detallada de instalación y puesta en marcha (ver guía de guía de documentación técnica) para usar el sistema, indicando dependencias y otros requisitos o potenciales casos especiales que el responsable del despliegue debiera tener en cuenta para poder tratar con todos los tipos y formatos de datos admitidos por el sistema).
- Se deberá indicar el proceso completo necesario para la puesta a punto de datos que requieran un preprocesamiento anteriormente para ser usados como entrada al sistema provisto (guía de datos).
- Deberá especificar con detalle suficiente cualquier post procesamiento necesario para poder interpretar la precisión del modelo de manera clara y concisa, y así poder notificar al proveedor cuando no sea el caso.
- Deberán proporcionar (según guía de documentación técnica) precisiones sobre el tipo, formato y la cantidad de datos de entrada y salida esperados, y los requisitos de éstos durante su uso en las etapas de entrenamiento, validación y testeo del sistema para obtener el nivel de precisión documentado, con ejemplos de uso de ejecución del sistema para su comprensión y transparencia hacia el responsable del despliegue.
- Para que la precisión reportada sea accionable y útil, el proveedor deberá incluir los rangos posibles de los parámetros configurables y tanto de los datos de entradas como de salidas, así como las medidas de latencia esperadas para obtener una precisión esperada (ver guía de solidez).

Para poder reproducir el comportamiento del modelo en cuanto a sus métricas de precisión, se recomienda usar documentación y formateado del modelo en el estándar abierto para interoperabilidad de aprendizaje automático ONNX.

Para detallar el proceso de selección de métricas de precisión y como éstas pueden garantizar los requerimientos específicos del artículo 15, precisión, solidez y ciberseguridad, del Reglamento Europeo de IA tanto en su definición como a lo largo del ciclo de vida, como tal, de manera consistente:

- Los apartados 4.2.1 y 4.2.2 abordan la selección de las métricas propuestas tanto propias de precisión, como aquellas relativas a la función objetivo, que como se ha indicado, es muy relevante para el concepto de precisión. Ambos apartados tienen su reflejo en los anexos 7.1 Métricas de precisión y 7.2 Funciones Objetivo, respectivamente.
- En el apartado 4.2.3 se proporcionan aspectos generales y complementarios relacionados con las métricas seleccionadas en los apartados anteriores y que deben tenerse en cuenta.
- Ambos aspectos están directamente relacionados el apartado anterior 4.1, donde se ha indicado la relación de la precisión con hitos críticos en el ciclo de vida.

Las medidas técnicas a aplicar por el proveedor se traducirán en controles basados en métricas de precisión del modelo, que se relacionarán y establecerán con la finalidad prevista del mismo, y con el objetivo de que la precisión siempre esté destinada a garantizarla, pero con el foco principal en la mitigación de los riesgos para los derechos y libertades de las personas físicas que han sido localizados en el sistema de gestión de riesgos (ver guía de gestión de riesgos).

4.2.1 Selección de métricas de precisión

El proveedor decidirá las métricas de precisión relevantes o KPIs de control de calidad (según guía de gestión de calidad) del sistema a medir y evaluar, y elegirá un mecanismo de almacenar e informa un registro histórico o log (según guía de registros) con los valores de éstas en el tiempo de uso del sistema, con el fin de monitorizar (ver guía de supervisión humana) la precisión y rendimiento del sistema. Como vemos, la relación de la precisión es transversal totalmente a los aspectos solicitados por el Reglamento Europeo de IA y su relación con el resto de las guías del sandbox.

Para proporcionar a los proveedores de una referencia de métricas de precisión, ver el anexo 7.1 Métricas de precisión, donde se presentan una lista de las métricas de precisión aplicables, para una tipología de modelos. El anexo ofrece una clasificación no exhaustiva, que pueda servir para ayudar en la selección de la métrica.

Para el proceso de selección de la métrica de precisión adecuada al sistema de IA recomendamos al proveedor, añadido a la vertiente fuertemente técnica de la selección, tener en consideración dos aspectos fundamentales:

- La finalidad prevista del sistema, considerando aquello que el sistema va a realizar y el objetivo de este.
- Los riesgos encontrados detectados en el sistema de gestión de riesgos en relación con las decisiones del sistema y que deban mitigados.

Como medida técnica es importante disponer de un repositorio centralizado donde gestionar la información de métricas asociadas al modelo en cualquier punto de su ciclo de vida con cambios asociados para poder trazar los causantes de pérdidas de precisión de este, junto con actores identificados responsables del mismo. El proveedor debe dotar al sistema de las capacidades que permitan al responsable del despliegue poder acceder y monitorizar las diferentes métricas de precisión y métricas asociadas al modelo, así como las variables protegidas, siempre de acuerdo con la guía de datos, para poder observar y reportar posibles

cambios en el sistema durante su uso, potenciales riesgos que puedan surgir, o sesgos que puedan emerger, en el sistema de IA alto riesgo.

Ejemplo - Detección de denuncias falsas

El proveedor, considerando que la finalidad prevista del sistema de IA es clasificar entre denuncias falsas y ciertas, con un grado de probabilidad, en una primera aproximación para seleccionar una métrica de precisión, se considera que se está realizando una clasificación binaria. Se ha identificado también, un riesgo de que no haya capacidad discriminatoria y no categorizar adecuadamente una denuncia verdadera, causando que la persona denunciante no tenga el trato adecuado.

Con ambos criterios, se establece que una de las métricas de precisión seleccionadas para este sistema de IA es la utilización del área bajo la curva ROC (ver Anexo 7.1Métricas de precisión). Un valor de 1 representa una clasificación perfecta de la denuncia y un valor de 0,5 significa que la valoración tiene una capacidad discriminatoria nula. De esta manera, a través de la selección de esta métrica de precisión a lo largo de todo el ciclo de vida se puede medir como el sistema de detección de denuncias falsas se comporta en el intervalo 1 (clasificación perfecta) y 0,5 (discriminación nula).

4.2.2 Selección de la función objetivo

Como hemos comentado en la introducción a este apartado, para optimizar la precisión del modelo, las funciones objetivo a optimizar (minimizar o maximizar), vendrán determinadas por los problemas concretos y función a aprender por el sistema de IA y, por tanto, directamente relacionados con la finalidad prevista. Éstas sirven para monitorizar el proceso de aprendizaje automático de modelos.

El proveedor del sistema deberá seleccionar una función objetivo que permita, del mismo modo que la métrica de precisión, alcanzar la finalidad prevista. En el anexo 7.1, se presentan una serie de funciones objetivo-clasificadas por modelos, que pueden ser utilizadas como base para el sistema de IA.

4.2.3 Dimensiones complementarias a la precisión

Como dimensiones complementarias, que respalden la precisión, se consideran los siguientes puntos:

- Una vez implementado el cómputo de las métricas de precisión, se deberá monitorizar (según guía de supervisión humana), durante el ciclo de vida de funcionamiento del sistema, que las nociones pertinentes al caso de uso, y la precisión del modelo global no varíen o se deterioren significativamente.
- Tanto métricas de sesgos, como de imparcialidad y explicabilidad, tendrán un papel especialmente relevante en una fase del ciclo de vida del sistema de IA (ver guía de datos, transparencia y supervisión humana).
- Tanto las herramientas de juicio equitativo (*fairness*) como de explicabilidad no etiquetadas como propietarias son de código abierto y preferibles por el principio de transparencia hacia IA responsable (ver guía de transparencia). Pueden encontrar más herramientas por considerar para alcanzar sistemas de IA responsable en [15].

Los aspectos relativos a estos aspectos complementarios a la precisión se presentan en mayor detalle en el anexo 7.3.

Un aspecto complementario relevante, y que es responsabilidad del proveedor, está relacionado con las capacidades que se le dota al sistema de IA para el responsable del despliegue. Para facilitar la comunicación de las métricas y buen funcionamiento del sistema al responsable del despliegue, el responsable del despliegue deberá a tener acceso a:

- Una interfaz de obtención de métricas de precisión y rendimiento del sistema que permita notificar deficiencias de este, potenciales errores de funcionamiento. Se entiende, por tanto, que la disponibilidad de dicha interfaz será responsabilidad del proveedor, que estará obligado a dotar al sistema de dichas capacidades, con las instrucciones adecuadas (ver guía de documentación técnica) y con las capacidades de supervisión y transparencia adecuadas (ver guías de supervisión y transparencia respectivamente).
- Inspeccionar los datos tanto de entrada como de salida.

Ejemplo - Sistema de promoción de empleados

Siguiendo las indicaciones establecidas en esta guía, el proveedor del sistema de promoción de empleados ha seleccionado las métricas de su sistema de IA. La selección la realiza basándose en proporcionar una medida de la precisión adecuada a la finalidad prevista establecida para el sistema de IA y por diseño y en seleccionar también una función objetivo también adecuada para la finalidad prevista, con ello:

- Por un lado, selecciona *Discounted Cumulative Gain* como métrica para evaluar la precisión del sistema. Está métrica está considerada para evaluar la relevancia de los resultados obtenidos en relación con una consulta o *query*. Dada la finalidad prevista, establece que su consulta se construye en base a los parámetros configurados para la promoción.
- Por otro lado, dentro de las funciones objetivo (ver anexo 7.2), descarta la utilización de *Pointwise* o *Pairwise* por simplificar demasiado su enfoque al aproximar su sistema de AI a un modelo de regresión o una clasificación binaria respectivamente, que no están alineados con la finalidad prevista del sistema de IA y que podría incurrir en excluir candidatos válidos. Se opta por tanto por la utilización de una función objetivo de tipo *Listwise* que permite maximizar la métrica de precisión seleccionada.

Todo el proceso de análisis y selección, tanto de la métrica de precisión, como de la función objetivo son documentados, explicando motivaciones y enfoques, así como todos los resultados que de su aplicación se obtienen durante las fases de diseño y validación del sistema de IA. Todo ello, se incorporará a la documentación técnica tal y como establece la correspondiente guía.

Responsable del despliegue

El responsable del despliegue del sistema IA de alto riesgo, deberá conocer, además de tener personal capacitado para entender durante la operación, el nivel de precisión del sistema. Por tanto, la precisión del modelo no será completa si no se provee transparencia de cómo se produce y cómo se computan las métricas de precisión y métricas de rendimiento relacionadas.

Ejemplo - Sistema de promoción de empleados

Una de las empresas que ha contratado la utilización de sistema de AI para la promoción de empleados, siguiendo las indicaciones presentes en este apartado para el responsable del despliegue, decide que aquel personal que va a operar con el sistema de IA reciba una formación que le permita entender, a nivel de operación, el concepto de precisión del sistema y la relación de esta con la finalidad prevista (disponer de un indicador para la promoción de empleados).

Para ello consultan con una empresa especializada en cursos de formación, y siguiendo las instrucciones proporcionadas por el proveedor del sistema de IA sobre precisión, añaden a la formación de uso del sistema unas horas dedicadas a proporcionar adicionalmente la siguiente información:

- El concepto de precisión a alto nivel y su relación con los resultados del sistema dentro de su operativa.
- Una explicación a alto nivel de cómo funciona la métrica de precisión seleccionada por el proveedor y su relación con los paneles informativos del sistema.

Esta formación, es complementada con un plan periódico de actualización, para garantizar que se tiene presente el concepto en la operación del sistema.

La transparencia, por tanto, dotará de calidad al modelo, en cuanto al grado de disponibilidad de información sobre el sistema de IA y la manera en que la información es comunicada a las partes relevantes de acuerdo con sus objetivos y saber hacer (en referencia a ISO/IEC 25059, *Software engineering – Systems and software Quality Requirements and Evaluation (73) – Quality model for AI systems*). Dado el contexto de la guía en la que nos encontramos, uno de los aspectos clave para esa información de transparencia, será la información asociada a las métricas de precisión seleccionadas (para más detalle en transparencia consultar la guía de transparencia).

El responsable del despliegue debe capacitar a los miembros de su organización que interactúen con el sistema de IA de Alto riesgo para:

- Saber utilizar e interpretar visualizaciones de monitoreo de la precisión relacionadas con todas las métricas, acorde a la finalidad prevista y su explicabilidad e incertidumbres asociadas en cualquier momento del ciclo de vida del sistema.
- Saber usar la salida del sistema, identificando requisitos que podrían ser necesarios para usarlo como entrada en otro sistema de IA. (Ver ISO de Sesgos, SC42_N1011 ISOIEC_TS_4213_2 [116]).
- Poder inspeccionar tanto datos de entrada como de salida y tener permiso para corregir y/o notificar al proveedor de potenciales datos o salidas erróneas, o la ausencia de estas.
- Disponer de conocimiento sobre las interfaces que proveen transparencia sobre el funcionamiento, salida y métricas de evaluación del modelo de IA, y su interpretación de acuerdo con la explicabilidad del modelo [35] para interpretarla.
- Responsabilizarse de entender los potenciales sesgos e imparcialidad que puedan comprometer la precisión del modelo. Por ejemplo, conocer los conceptos de discriminación algorítmica, sesgos e imparcialidad, para detectar posibles problemas o potenciales degradaciones de la salida del modelo. Así como, la

degradación de la precisión de forma no favorable para una clase minoritaria o (grupo de) variables protegidas.

En línea con el ejemplo anteriormente indicado, sobre el sistema de IA para la promoción de empleados, la empresa responsable del despliegue del mismo debe de asegurarse que todos los puntos anteriores son adecuadamente trasladados al personal de recursos humanos.

4.3 Garantizando la precisión

Las métricas de precisión aplicables garantizadas por el proveedor en la documentación técnica deben reflejar y ser una señal de la calidad del sistema. De otro modo, cuando no se puedan garantizar estas métricas de precisión establecidas mínimas, el proveedor facilitará al responsable del despliegue mecanismos para notificar al humano se requiere su supervisión (según guía de supervisión humana). Del mismo modo, el responsable del despliegue estará obligado a conocer dicha información y disponer de procesos internos para poderlos llevar a cabo.

En el apartado anterior hemos indicado cómo seleccionar las métricas de precisión para el sistema de IA y como estas se relacionan con otras acciones complementarias. En este apartado vamos a proporcionar la información necesaria para que las métricas seleccionadas, se mantengan de manera consistente a lo largo del ciclo de vida, tal y como establece el Reglamento Europeo de IA.

Proveedor

El proveedor del sistema de IA es el principal garante de que la precisión se mantenga de manera consistente a lo largo del ciclo de vida de IA, con la aplicación de las medidas que lo garanticen.

4.3.1 Medidas técnicas

Las **medidas técnicas** que el proveedor debe implementar en relación con los inventarios de métricas de precisión y funciones objetivo mencionadas en el apartado anterior (ver apartado 4.2) se identifican con las siguientes acciones:

- Para que la precisión reportada por el sistema de IA sea accionable y útil, dicha documentación debe incluir los rangos posibles de los parámetros configurables, así como los datos de entrada y salida del sistema, y las medidas de latencia esperadas para alcanzar la precisión deseada (ver Guía de Solidez).
- La salida del sistema, idealmente, deberá acompañarse de una medida de incertidumbre asociada a la precisión de dicha salida.
- El proveedor elegirá un mecanismo para almacenar todas las precisiones comunicadas al responsable del despliegue en un registro histórico, conforme a la Guía de Registros, con el fin de permitir la trazabilidad de las métricas de precisión a lo largo del tiempo de uso del sistema y así monitorear su rendimiento (según la Guía de Solidez).
- Se proporcionará al responsable del despliegue una interfaz gráfica que permita observar las métricas de precisión en el tiempo y detectar incoherencias o mal

funcionamiento del sistema (según la Guía de Supervisión Humana), y monitorizarlas (según la Guía de Solidez).

- La base del razonamiento para seleccionar una métrica u otra para evaluar la precisión del modelo debe documentarse (conforme a la Guía de Documentación Técnica); véase también el apartado 5, donde se abordan los aspectos relativos a la documentación de las acciones indicadas en esta guía.

Ejemplo - Sistema de promoción de empleados

Durante la fase de análisis del sistema de IA, se ha concluido que una de las medidas establecidas para mantener la precisión a lo largo del tiempo, es disponer en la *operación* del sistema de un panel donde se muestre las métricas de precisión y su evolución en el tiempo, de esta manera es posible mantener visible el progreso de la precisión y por tanto poder actuar sobre ella en caso de que abandone los rangos de trabajo definidos. El sistema además de mostrar los datos realiza un registro de la información de precisión presentada acorde a lo establecido en los registros.

Del mismo modo durante el análisis del sistema se encuentra que es importante acompañar, tal y como se recomienda en esta guía, de una información de la incertidumbre, el resultado de la salida del sistema, para ello se plantea la utilización de un conjunto de modelos combinados (*ensemble learning*). La utilización de esta técnica permite establecer un valor de confianza del resultado, basado en las respuestas de los diferentes modelos (*ensembles*) combinadas para poder proporcionar una medida de la incertidumbre del resultado.

4.3.2 Evaluaciones de significancia estadística

Durante el proceso de selección de la métrica de precisión (así como las funciones objetivo), para dar validez estadística a los resultados para el sistema de IA, se deben utilizar métricas específicas que deberán acompañarse de evaluaciones estadísticas. Las evaluaciones de significancia estadística deberán considerar la distribución de los datos (evaluación paramétricos o no), la dependencia entre ellos (independientes e idénticamente distribuidas, i.i.d, o no) y otras suposiciones concretas de cada evaluación, para así elegir adecuadamente la evaluación relevante (ver guía de datos). Entre ellos:

- La evaluación de *ranking* con signo Wilcoxon. No paramétrico, es decir libre de suposiciones sobre la distribución, y datos dependientes.
- Evaluación paramétricos o no.
- Se podrán usar el *Paired student T-Test*, análisis de varianzas, evaluación de Kruskal-Wallis, evaluación Chi-Squared, evaluación de Fisher exacto, test de McNemar, (incluyendo comparación múltiple, la corrección Bonferroni y la tasa de descubrimiento falsa) u otros.

Muchos de éstos requerirán test adicionales de normalidad en los datos, por ejemplo, análisis de varianzas (ANOVA) para determinar si las medias de más de dos grupos son iguales, o sus diferencias en precisión (*accuracy*) entre 3 o más modelos son estadísticamente significativos, para lo cual ANOVA asume distribuciones normales y que la varianza sea homogénea. También tener en cuenta la aplicabilidad de este según el Teorema Central del Límite (*Central Limit Theorem*).

Ejemplo - Sistema de promoción de empleados

En este sistema de IA, como se ha explicado en el apartado anterior, el proveedor ha realizado la selección de *Discounted Cumulative Gain* como métrica para evaluar la precisión del sistema.

Sobre este sistema de IA, se realizan una batería de evaluaciones de significancia estadística utilizando la métrica establecida. Para ello se realizan una serie de conjuntos de datos de prueba y validación agrupados por sectores profesionales, para mantenerlos en conjuntos de datos con relación entre ellos y que permitan consolidar adecuadamente los resultados. Habiendo considerado que los resultados posibles de los diferentes conjuntos de datos no tienen una distribución normal, se ha seleccionado *ranking* con signo Wilcoxon, para realizar dicha evaluación como mejor herramienta para validar el proceso.

La evaluación de la significancia estadística es utilizada durante el proceso de entrenamiento/prueba para verificar que el modelo se ajusta a las especificaciones definidas en la fase de diseño, y acorde a la finalidad prevista (disponer de un indicador para la promoción de empleados). Permite comparar la evolución del modelo entre sí a lo largo de su evolución en el tiempo, o comparar diferentes modelos candidatos entre sí en un momento del proceso (*ensemble learning*), sin dejar de considerar la métrica de precisión ni las especificaciones y guiando en el proceso de entrenamiento y prueba.

Para más información se puede consultar ISO SC42 N1011.ISOIEC TS 4213_2 (Clasif. Perf) [116]

4.3.3 *Benchmarks* de bases de datos y modelos

En la información completa sobre la métrica o métricas de precisión seleccionadas para el sistema de IA, se debe proporcionar información de la realización de *benchmarks (pruebas de rendimiento)* de los datos y del modelo. Estos *benchmarks*, deben tener en consideración a la finalidad prevista del sistema, así la precisión se puede poner en contexto no solo interno dentro del propio sistema de IA, si no el contexto externo en relación con datos y modelos en relación con pruebas repetibles y medibles.

Se usarán y reportarán todas las métricas convenientes para medir la precisión del modelo y conforme a los protocolos establecidos por las evaluaciones *benchmark* de cada tarea de aprendizaje automático. Deberán usarse y documentarse los *benchmarks* más usados por la comunidad científica relevante en aprendizaje automático e IA y sus campos de aplicación.

Los *benchmarks* dependen de la tarea y el estado del arte del campo; por tanto, deberán actualizarse conforme pasa el tiempo y los modelos y bases de datos evolucionan. Ejemplos de bases de datos y modelos correspondientes que actúan bien como el estado del arte o bien como modelos de referencia base (*baseline*) en la tarea de clasificación de imágenes son:

- En modelos de visión por computador o que procesan imágenes:
  - Benchmark / Base de datos ImageNet: Modelos base AlexNet, VGG-19, ResNet-50, EfficientNet-B7.
  - Base de datos MNIST: Modelos base RDML, MCDNN, LeNet.
  - Base de datos CIFAR-10: Modelos base EfficientNet-B7, ColorNet, DenseNet.

    - Base de datos LFW: Modelos base FaceNet, DeepID3, DeepFace, etc…
- En modelos de lenguaje, Los modelos LSTM, BERT, GPT1, 2, 3, [...], etc. se consideran *baselines* con respecto al estado del arte. Considerar, además las métricas específicas a la tarea concreta (traducción automática, comprensión lectora, resumen, etc.). Para ello, ver tabla H.1 [4] para un conjunto de tareas y métricas significativas en modelos de lenguaje.
- En modelos generativos de imagen se miden el fotorrealismo o diversidad de muestra (SSIM, MMD, IS, MS, FID, LPIPS). Por ejemplo, para evaluar la consistencia física: IoU (en modelos de segmentación de imágenes) o FVPS [5].
- Bases de datos para la imparcialidad incluyen COMPAS recid, COMPAS viol. recid, Diabetes, OULAD, *Credit card clients and many others* (ver Tabla 1 en [9]).

Más medidas para asegurar la precisión y documentación del modelo pueden verse en la ISO SC42_N1011_ISOIEC_TS_4213_2 [116].

Responsable del despliegue

Organizativamente, el responsable del despliegue deberá considerar:

- El responsable del despliegue tiene la responsabilidad de consultar el manual de instrucciones del sistema de IA para conocer, aplicar y mantener vigilado el modelo (según guía de datos y guía de supervisión humana).
- Tanto responsable del despliegue como todas las partes implicadas o *stakeholders* (desde el mayor gestor responsable o top mánager, *product owner*, *project manager y data scientist*) deberán tener acceso a la justificación de la precisión del sistema y sus métricas asociadas en cualquier momento del ciclo de vida útil del mismo. Para ello se usarán interfaces adaptadas a diferentes audiencias que puedan requerir una explicación del modelo o su auditoría, para así facilitar la transparencia del proceso de razonamiento.

El responsable del despliegue deberá familiarizarse y comprender al nivel adecuado la interpretación de las nociones y métricas del modelo en cuanto a su precisión y su relación con la finalidad.

5. Documentación técnica

En la guía de documentación técnica que se proporciona dentro del marco del sandbox de IA, se indica de manera detallada y acorde al Reglamento Europeo de IA, la estructura y el contenido de la documentación técnica de sistema de IA de alto riesgo. En este apartado vamos a entrar en detalle en aspectos específicos de como documentar el proceso de selección y aseguramiento de la precisión que se ha detallado en esta guía.

El Reglamento Europeo de IA establece que no solo la definición y selección de métricas de precisión es relevante, si no también elevar su alcance al responsable del despliegue del sistema en las instrucciones pertinentes. Así, refleja en el párrafo 2 y 3 de su artículo 15 sobre precisión, solidez y ciberseguridad:

AI Act

Art.15.2 y 3 – Precisión, solidez y ciberseguridad

2. Para abordar los aspectos técnicos sobre la forma de medir los niveles adecuados de precisión y solidez establecidos en el apartado 1 y cualquier otro parámetro de rendimiento pertinente, la Comisión, en cooperación con las partes interesadas y organizaciones pertinentes, como las autoridades de metrología y de evaluación comparativa, fomentará, según proceda, el desarrollo de parámetros de referencia y metodologías de medición.

3. En las instrucciones de uso que acompañen a los sistemas de IA de alto riesgo se indicarán los niveles de precisión de dichos sistemas, así como los parámetros pertinentes para medirla.

Las instrucciones sobre cómo se propone obtener la precisión, cómo reportarla (guía de transparencia), cómo interpretar y usar sus umbrales y métricas asociadas a ella se documentarán en los apartados específicos detallados en la guía de documentación técnica acorde a la selección y valoración de estas, tal y como se ha detallado en los apartados anteriores y en el anexo 7.1.

En el anexo IV de la documentación técnica mínima requerida para los sistemas de IA de alto riesgo, encontramos algunos puntos referentes a la precisión del sistema:

- Punto 2(g): *"Los procedimientos de validación y prueba utilizados, incluida la información acerca de los datos de validación y prueba empleados y sus características principales; los parámetros utilizados para medir la* ***precisión****, la solidez y el cumplimiento de otros requisitos pertinentes establecidos en el capítulo III, sección 2, así como los efectos potencialmente discriminatorios; los archivos de registro de las pruebas y todos los informes de las pruebas fechados y firmados por las personas responsables…"*

Este punto es el núcleo documental sobre la documentación de la precisión, porque obliga a describir cómo se mide, con qué datos, bajo qué condiciones y con qué métricas. Se debe documentar el diseño de las pruebas, los conjuntos de validación, las fórmulas o indicadores de precisión (referirse al anexo 7.1), los resultados obtenidos y las evidencias que los respalden.

- Punto 3: "*Información detallada acerca de la supervisión, el funcionamiento y el control del sistema de IA, en particular con respecto a sus capacidades y limitaciones de funcionamiento, incluidos* ***los niveles de precisión*** *para las personas o colectivos de personas específicos en relación con los que está previsto que se utilice el sistema y el nivel general de precisión esperado en relación con su finalidad prevista…"*

Este apartado exige declarar el nivel de precisión alcanzado y esperado del sistema en condiciones reales de uso, diferenciando por grupos o contextos cuando sea relevante. Se debe documentar la metodología para medir esa precisión operativa, las limitaciones conocidas, los márgenes de error y la justificación de que el rendimiento es adecuado para la finalidad prevista.

- Punto 4: "Una descripción de la idoneidad de los parámetros de rendimiento para el sistema de IA concreto."

Aquí se requiere justificar que las métricas o indicadores empleados para medir la precisión son adecuados y representativos del uso previsto del sistema. Se debe documentar por qué se eligieron esos parámetros, su relevancia técnica y cómo reflejan correctamente el rendimiento real o esperado del modelo.

- Punto 9: *"Descripción detallada del sistema establecido para evaluar el funcionamiento del sistema de IA en la fase posterior a la comercialización, incluido el plan de vigilancia poscomercialización…"*

Este punto vincula la precisión con su seguimiento continuo tras la comercialización. Exige mantener un procedimiento documentado para monitorizar la evolución del rendimiento (incluida la precisión) a lo largo del tiempo, registrar degradaciones o desviaciones y establecer acciones correctivas.

Si se quiere ir un paso más allá del mínimo requerido por el reglamento y para documentar adecuadamente la precisión, puesto que dependerá de la calidad de los datos, para evaluarla adecuadamente en términos de precisión de forma adecuada, el proveedor deberá facilitar en la documentación dos elementos descriptivos relevantes: La tarjeta del modelo y la tarjeta descriptiva de la base de datos empleada.

5.1 Tarjeta del Modelo

Una tarjeta del modelo(s) usados por el sistema, con sus métricas de precisión, solidez, capacidades operativas, limitaciones y sus relaciones en torno a cambios en la precisión y solidez del sistema. Este apartado está en relación con lo descrito en la guía de Documentación Técnica, pues deberá incorporarse a la dicha documentación.

Las tarjetas de modelo (*Model Cards*) deben incluir secciones especificando, como mínimo:

- Tamaño del modelo, incluyendo el número de parámetros en las configuraciones clave.

- Fecha de publicación.
- Tipo de modelo, indicando si ha sido desarrollado desde cero o si se ha utilizado un modelo de base.
- Finalidad prevista.
- Términos de identidad, con referencia a grupos frecuentemente vulnerables o afectados, como los centrados en la orientación sexual, género y raza.
- Artículos y referencias clave, en caso de que existan.
- Detalles del uso del modelo en entrenamiento, validación y evaluación, junto con información sobre rendimiento, eficiencia y limitaciones.
- Implicaciones de mayor alcance y consideraciones éticas, riesgos y recomendaciones [50].

En particular, las tarjetas de modelo deben incluir, con respecto a las métricas de precisión (performance) o calidad de este, respuesta a las siguientes preguntas en un apartado específico para las métricas, que midan e ilustren los errores de precisión del modelo desproporcionadas entre subgrupos

- Medidas de precisión del modelo: ¿Qué métricas se reportan y por qué fueron seleccionadas? Deben especificarse todas las métricas que reflejen el potencial impacto real del modelo, incluyendo medidas de precisión y otras métricas relacionadas con el rendimiento (*performance*).
- Umbrales de decisión: Si se usan, ¿por qué se seleccionaron esos valores específicos? En las tarjetas de modelo en formato digital, se recomienda incluir una barra deslizante interactiva para mostrar las métricas de precisión en función de distintos umbrales de decisión.
- Enfoques hacia la incertidumbre y variabilidad: ¿Cómo se calculan las medidas de incertidumbre y estimaciones de estas métricas? Esto podría incluir desviación estándar, varianza, intervalos de confianza o divergencia KL.
- Métodos de cálculo y origen de los valores: Detalles sobre cómo se obtienen los valores de estas métricas deben estar incluidos (por ejemplo, media de cinco ejecuciones, validación cruzada de 10 iteraciones o pliegues, etc.).
- Interpretabilidad del modelo: ¿Qué métodos o herramientas se han implementado para facilitar la comprensión del modelo y sus predicciones? (Por ejemplo, SHAP, LIME, o visualizaciones de importancia de características).
- Adaptabilidad del modelo: Información sobre la capacidad del modelo para adaptarse a nuevos datos o condiciones, incluyendo detalles sobre procesos de *fine-tuning* o actualización periódica.
- Uso responsable: Advertencias o limitaciones del modelo en términos de su aplicación para evitar malentendidos o uso indebido. Esto podría incluir advertencias sobre aplicaciones sensibles o contextos inapropiados, y cómo estos podrían impactar negativamente en los resultados.

Las tarjetas de modelos constituyen una herramienta de transparencia, importantes también en relación con la solidez (ver guía de solidez correspondiente) y documentación técnica, favoreciendo la auditoría técnica y no técnica por los analistas, así como mecanismos de retroalimentación por el responsable del despliegue más inclusivos.

Sin embargo, hasta que su estandarización o formalización no se lleve a cabo de forma que evite representaciones de resultados engañosos o confusos, la utilidad y precisión de las tarjetas de modelo se basan en la integridad del creador de la misma tarjeta [50].

Ejemplo - Detección de denuncias falsas

Para cumplimentar la tarjeta del modelo, siguiendo las directrices establecidas en esta guía, el proveedor del sistema de IA proporciona en dicha tarjeta la entre otras cuestiones, la información siguiente:

- Se indica que se ha seleccionado área bajo la curva ROC como métrica de precisión y a ésta se le ha complementado con la información de la matriz de confusión del sistema, como hemos indicado, entendido el sistema de IA como una clasificación binaria.
- Dado que el sistema de IA realiza un procesamiento del texto de las denuncias, también se ha considerado la perplejidad como métrica de precisión. En la ficha del modelo se detalla cómo durante el entrenamiento y prueba se ha analizado los resultados de perplejidad.
- Para gestionar la incertidumbre y la variabilidad, añade información a la tarjeta del modelo, del procedimiento seguido. Se han realizado divisiones de los datos de validación aleatorias, y por conjuntos específicos de origen, para estudiar cómo se comportan las métricas de precisión en los diferentes conjuntos.

Nota: El proveedor, ha completado la información de la tarjeta del modelo añadiendo los campos que se han indicado en la guía, y en este ejemplo hemos indicado aquellos de mayor relevancia y en consonancia con el tratamiento de este caso de uso desarrollado a lo largo de la guía. Los participantes del sandbox, para el caso de las tarjetas del modelo, deberán completar todos los datos indicados en este apartado.

5.2 Tarjeta de base de datos

Las tarjetas de bases de datos (*Datasheets for datasets*) [10] son aconsejables para dar visión más holística de las métricas de precisión provistas y de la proveniencia de los datos usados para el entrenamiento del modelo. Si ésta cambia con el uso del modelo, las tarjetas del modelo deberán también actualizarse.

Deberán estudiarse metodologías de documentación asociadas a la transparencia como las mencionadas, entre otras adecuadas a venir potencialmente en el futuro, para dar una visión más amplia a las métricas de precisión y visualizar, si las hubiese, posibles disparidades interseccionales, por ejemplo, en la precisión en la tarea realizada por el modelo.

Se puede consultar un ejemplo de disparidad interseccional para la precisión (*accuracy*) en clasificación comercial de género está disponible en [17], y un ejemplo de tarjeta de modelo se puede ver en https://github.com/openai/whisper/blob/main/model-card.md.

El proveedor atenderá a taxonomías de imparcialidad y sesgo para elegir aquellas métricas que complementen y enriquezcan la noción de precisión provista por el sistema en la finalidad prevista del sistema.

El proveedor deberá documentar las pruebas hechas para posibles tipos de sesgos que pueda sufrir el sistema afectando a la precisión para categorías de datos que puedan resultar ser discriminadas. Del mismo modo, se deberá documentar cómo las mismas métricas provistas, a nivel global de modelos, no presentan impacto desigual o disparatado (*disparate impact*) para diferentes grupos pertenecientes a variables sensibles. Para éstas, la salida del modelo debe ser imparcial y no discriminar (ver guía de datos).

6. Cuestionario de autoevaluación

Para realizar una autoevaluación del cumplimiento de los requisitos del Reglamento de Inteligencia Artificial referidos en esta guía, se ha generado un cuestionario de autoevaluación global con una serie de preguntas con los puntos clave a tener en cuenta respecto a las obligaciones que dictaminan los artículos del Reglamento de IA mencionados en esta guía.

Será necesario referirse a ese documento para realizar el apartado del cuestionario de autoevaluación correspondiente a esta guía.

7. Anexos

7.1 Métricas de precisión

En este anexo presentamos una serie de métricas de precisión que pueden ser utilizadas y los tipos de modelos con los que se relacionan. Este Anexo no pretende ser una enumeración exhaustiva de métricas existentes, si no una presentación de aquellas que se puedan aplicar a los tipos de modelos presentados. El proveedor deberá tener en cuenta que una tipología específica de modelo o finalidad prevista (o la combinación de ambas) podrá requerir un análisis más específico de la métrica a seleccionar y que podría esta no encontrase en el listado presentado.

En los siguientes puntos se describen las métricas de precisión propuestas para medir la precisión, agrupadas por la tipología del modelo asociado.

1. Métricas de error para modelos regresión: MSE, RMSE, MAE, MAPE. R^2 y R^2 ajustada.

2. En modelos de clasificación, el cómputo de métricas incluirá el uso de elementos básicos de precisión, exactitud, tasas de aciertos y otras métricas de rendimiento del modelo, de acuerdo con la relevancia de la finalidad prevista del sistema del sistema de IA. En este sentido, se incluirán:

- Matriz de confusión.
- Exactitud o tasa de aciertos (*accuracy*).
- Precisión, *recall* y especificidad y sensibilidad, el valor F1, F beta, y/o la divergencia Kullback-Lieber.

2.i. Clasificación Binaria. En particular, los modelos implementados para realizar clasificación binaria deberían reportar:

- Matriz de confusión.
- Tasa de aciertos.
- Valor F (F1 o F beta cuando se quiere dar importancia mayor a precisión (que a exhaustividad o, al contrario), y/o la divergencia Kullback-Lieber.
- Curva de Precisión y *recall*, área bajo la curva ROC (Característica Operativa del Receptor) y área bajo la curva AUROC.
- Curva de respuesta cumulativa, y curva de elevación (*lift curve*).

2.ii Métricas de evaluación en clasificación multiclase: Error de cobertura, el valor de precisión media del ranking de etiqueta, y *Ranking loss*, curva AUROC (útil cuando hay datos no balanceados y el ranking entre predicciones es importante). Además, los modelos de clasificación multiclase deberán reportar:

- Precisión (*accuracy*), precisión macro-media, micro-media y media-ponderada.
- Métricas de distancia o diferencias en distribuciones. Por ejemplo, en destilación, se puede medir la fidelidad entre distribuciones (la métrica de acuerdo medio o la divergencia KL media [33] entre distribuciones).

Por ejemplo, la fidelidad (divergencia KL media) se puede usar para medir el grado de alineación entre el modelo aprendido por el modelo estudiante y el modelo profesor, cuando el estudiante simplifica el del profesor, aplicando aprendizaje con destilación.

2.iii Métricas de evaluación en clasificación multi-etiqueta: Éstos deberán reportar al menos una de las siguientes métricas:

- Función objetivo Hamming Loss.
- Ratio del match exacto.
- Índice de Jaccard, o IoU (intersección sobre unión).
- Métricas de distancia o diferencias de distribución.

3. Métricas de evaluación de *clustering*: éstas incluyen el valor de información mutua ajustada, índice de Rand, el valor de Calinski y Harabaz, de Davies-Bouldin, matriz de contingencia, la métrica de completitud, el índice Fowlkes-Mallows, el coeficiente Silueta medio, etc.

4. En otras tareas de aprendizaje automático más concretos, estudiar la métrica más conveniente para cuantificar la precisión del modelo. Por ejemplo:

4.i Métricas de evaluación en problemas de aprendizaje automático con datos no balanceados: Se usan métricas de clasificadores de una clase (*unary* o *one-class classifiers*) y de detección de anomalías. En particular, una curva precisión-*recall* (PRC) y el área bajo ella (AUPRC) son métricas más adecuadas que una curva ROC y la métrica AUROC para mostrar precisión con datos no balanceados. Valor F1 con media ponderada (*Weighted-average F1*) también es una métrica representativa en estos casos.

4.ii Evaluación de modelos del lenguaje: En reconocimiento del habla se usa la perplejidad de los datos de evaluación, tasa de errores por palabra (*Word Error Rate*, WER). Otras métricas más genéricas son Language Model Probability y Word Accuracy, BLEU, METEOR, NIST LRE y otras.

4.iii Para evaluar modelos de imágenes, usar métricas específicas a la tarea; Por ejemplo, en segmentación de objetos en imágenes: intersección sobre la unión (IoU). En modelos generativos, se podrá reportar la Frechet Information Distance (FID) o Inception Score (IS). En modelos de autoencoders como VAE se pueden usar *sharpness*, *Inception score*. FID *score*, entre otras métricas [3].

7.2 Funciones Objetivo

En este anexo presentamos una serie de funciones objetivo, agrupadas por las tipologías de modelo o tareas con los que se relacionan.

7.2.1 Funciones objetivo regresión, clasificación o *ranking*

A continuación, listamos funciones objetivo de regresión, de clasificación o de ranking:

1. En regresión se podrán usar: MAE (*Mean Average Error*), MSE (*Mean Squared Error* o L2 *loss*), MAPE (*Mean Absolute percentage Error*), *Mean Squared Logarithmic Error* (robusta a valores anómalos), similaridad coseno, logaritmo del coseno hiperbólico (LogCosh), Huber *loss* (menos sensible a valores anómalos).

2. En clasificación dependiendo del mecanismo del modelo utilizado.

2.i Clasificación binaria se recomienda la utilización de las funciones objetivo: Hinge *embedding loss* (también para aprender representaciones o *embeddings* no lineares o tareas semi supervisadas), cross-entropía binaria (BCE), divergencia KL, etc.

2.ii Clasificación multi clase se podrán usar funciones objetivo como la probabilidad logarítmica negativa (*negative log likelihood* NLL), Entropía cruzada (*Crossentropy*), Entropía cruzada categórica, Entropía cruzada categórica *sparse*, de Poisson, Divergencia KL (para asegurarse que la distribución de las predicciones es similar a la de los datos de entrenamiento) y para aproximar funciones complejas).

2.iii Clasificación multi-etiqueta: La función objetivo Hamming *Loss*.

7.2.2 Funciones objetivo en otros tipos de modelo

Para otros tipos de modelo del sistema de IA de alto riesgo, se enumeran en este apartado un compendio de otras posibles funciones de objetivo aplicables, con el objetivo de establecer la precisión de los modelos.

- En clasificadores con datos no balanceados, o detección de objetos: Focal *loss*.
- En problemas de ranking, se podrán usar funciones objetivo como la función objetivo de ranking de margen (*Margin Ranking Loss* para predecir la distancia relativa entre entradas), y métricas como AUC entre otras.
- En problemas de ranking también se puede utilizar *Pointwise Methods, Pairwise Methods, Listwise Methods.*
- En problemas de aprendizaje de representaciones o *embeddings* y de aprender similaritud relativa entre entradas y problemas basados en la recuperación basada en contenido: objetivo del margen tripleta (*Triplet margin loss*).
- En modelos generativos como las GANs, se pueden usar funciones objetivo de discriminador y generador como funciones básicas objetivo clásicas de min-max loss de GAN, la *non saturating GAN loss*, y alternativas como la Wasserstein GAN *loss*, función objetivo de GAN condicional (CGAN), *Sharpness Loss*, Distancia Cook, u otras https://neptune.ai/blog/gan-loss-functions. Además, en modelos VAE se puede usar *Log Hyperbolic Cosine Loss* (LogCosh).
- En aprendizaje por refuerzo, las funciones objetivo vendrán dadas por funciones de recompensa adaptadas al problema y entorno a resolver. La elección específica en cada caso deberá evidenciarse en la documentación técnica. Se ha de describir en dicha documentación la motivación de la elección de la función objetivo seleccionada, y cómo esta se relaciona con la precisión establecida para la finalidad prevista del sistema.

7.3 Precisión, sesgos e imparcialidad

En este apartado vamos a abordar la relación de la precisión con dos aspectos del sistema de IA de alto riesgo en particular: sesgos e imparcialidad. El objetivo es, por tanto, establecer las fronteras entre la precisión y el resto de los aspectos aquí tratados para que el proveedor del sistema de IA tenga una visión completa.

7.3.1 Sesgos y precisión

La presencia de sesgos puede afectar fuertemente a la precisión del sistema, incluso sin un correcto análisis de sesgos, falsear una métrica de precisión, invalidando su utilidad, por ello es muy importante, en el proceso de establecimiento de las métricas adecuadas para la precisión del sistema de IA, en su desarrollo y validación. Realizar una verificación de sesgos y mitigación de éstos en los datos, y en el funcionamiento y salida del sistema de IA.

Se deben verificar los posibles sesgos a evitar, en el diseño, y a verificar y mitigar durante el ciclo de vida de uso del modelo de IA según guía de datos; éstos se detallan en la ISO en Sesgo ISO/IEC TR 24027:2021, *Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision making*, el catálogo de sesgos (https://catalogofbias.org/) y el catálogo de sesgos ejemplificado [7].

Especialmente en la etapa de diseño, la taxonomía de sesgos cognitivos BIASeD, o equivalente, ha de ser tenida en cuenta metódicamente para considerar y evitar cada sesgo y cada efecto listados en la misma [112].

- Verificación de nociones de imparcialidad. Las métricas de precisión serán fiables si son justas o imparciales (*fair*), y transmiten la información necesaria para facilitar la transparencia del modelo para llegar a ellas (según la guía transparencia).
- Verificación de la explicabilidad del modelo, para así, facilitar los dos puntos anteriores (ver guía de datos).

Se deberá evitar una serie de sesgos en los datos, considerando las perspectivas de uso del modelo en el contexto de la IA ética:

- Sesgo de representación.
- Sesgo histórico.
- Sesgo de población, de muestreo, de variables omitidas.
- Sesgo de máscara (*masking bias*), interseccional y el sesgo algorítmico [18] entre otros aplicables de relación directa con la finalidad prevista del sistema de IA (ver guía de datos).

Otros sesgos más genéricos podrían también estar presentes según aplicabilidad: el sesgo de selección, de colisionador (*collider bias*), de información, de exhaustividad (*recall*), de comprobación (*ascertainment*), de desgaste o atrición (*attrition*), de tiempo inmortal, el sesgo de variables de confusión (*confounding*), de clasificación errónea, o el efecto Hawthorne, entre muchos [8].

Todos estos sesgos deben monitorizarse desde la recogida de los datos hasta el fin del ciclo de vida de funcionamiento, especialmente tras la puesta en marcha [62]. Para completar la lista de herramientas para facilitar la aplicación de monitorización de sesgos e imparcialidad, de nuevo aparece una relación clara entre las guías de datos y supervisión humana y la noción y métrica de precisión.

Existen en el ecosistema de imparcialidad para sistemas de AI herramientas que pueden ser utilizadas por el proveedor tanto comerciales como, en gran medida de código abierto. Las nociones de imparcialidad deben estar aseguradas, idealmente y cuando tenga sentido, de acuerdo con las escalas de riesgo especificadas en la Guía de Riesgos.

Por ejemplo, los modelos de predicción de reincidencia de crímenes: deben exhibir equidad de exactitud (*accuracy equity*) y paridad predictiva (*predictive parity*) [16]. Por ejemplo, dada la proximidad con el concepto indicado, y la finalidad prevista del sistema de detección de denuncias falsas, el proveedor añade como criterio de precisión añadido al ya seleccionado, cumplir con ambos conceptos.

7.3.2 La imparcialidad para mitigar sesgos

Asumiendo que el preprocesado de los datos se ha hecho de forma adecuada a la finalidad prevista (según guía de datos) y de forma justa para todas las variables (protegidas y no), cuando se haya detectado un sesgo de la guía de datos o noción de imparcialidad que no se satisface, se podrán aplicar técnicas de imparcialidad.

Las métricas de imparcialidad relevantes tienen como objetivo descubrir el posible sesgo en los datos, modelo o el propio diseñador / desarrollador del modelo, que pueda requerir acciones de mitigación. Generalmente, éstas se basan en diferencias, ratios, y evaluaciones de imparcialidad estadística. La ontología de posibles nociones y métricas de imparcialidad para medir tipos de sesgos se divide en las siguientes dos categorías (Fig.2 de [18]):

El cómputo de la precisión del modelo tendrá en cuenta adecuadamente, las métricas de imparcialidad para minimizar la posible discriminación del modelo hacia grupos de valores pertenecientes a variables sensibles (como minorías por orientación sexual, raza, género o sexo). Según aplicabilidad, las diferentes nociones de imparcialidad se implementarán a través de métricas de ratios, diferencias o test estadísticos según la aplicabilidad de cada métrica. Por ejemplo:

Nociones de imparcialidad de grupo:

- Imparcialidad de independencia de clase (paridad estadística, paridad estadística condicional).
- Imparcialidad de separación de clase: igualdad de probabilidades (*Equalized Odds*), igual oportunidad, igualdad predictiva, imparcialidad total.
- Imparcialidad de suficiencia de clase (igualdad de exactitud de uso condicional, paridad predictiva, calibración Well).
- Imparcialidad total.
- Igualdad de tratamiento.
- Igualdad de exactitud (*accuracy*) general.

Nociones de imparcialidad relajada:

- Nociones de imparcialidad con umbral y basada en evaluaciones estadísticas [18].

7.3.2.1 Taxonomías y métricas de imparcialidad.

Una guía comprensiva debe evaluarse previo uso. Pueden verse en [9], la *Fairness Ontology* [18], y en taxonomía de sesgos cognitivos BIASeD [112]. Habrá casos en que no se puedan obtener todos los principios [63], y el compromiso de tal conflicto debe documentarse según la guía de documentación técnica y en la tarjeta del modelo.

Técnicas para evaluar y establecer imparcialidad, incluyen, por ejemplo, entre otras técnicas:

- Ponderación (*reweighting*) para mitigar el sesgo en la fase de colección de datos de entrenamiento.
- Técnicas para evaluar los nuevos datos transformados, como el *balanced accuracy* o el *average odds difference*, para evitar impacto dispar (*disparate impact*) o resultados que no preservan la igualdad (*unequal outcomes*) entre variables protegidas [62].
- Calcular el Área Bajo las Curvas ROC absoluta para grupos protegidos y no (*Absolute Between-ROC* Área, ABROCA) [113]).
- Cuando se dude qué métricas de imparcialidad deben evaluarse, se puede aplicar la ontología de imparcialidad [18]. El grupo privilegiado será considerado como aquel que históricamente se observó tener ventaja sistemática, mientras que el grupo no privilegiado representaría aquel con desventaja sistemática en la historia.
- Las variaciones de acuerdo con estos grupos y la noción de imparcialidad elegida pueden representarse de forma gráfica para describir visualmente los sesgos en combinaciones estratificadas de atributos protegidos con gráficos de ráfagas de sol (*Plotly sunburst plot*). Ver Fig. 6 en [50].
- Para la verificación de nociones de imparcialidad, se podrán establecer equivalencias: por ejemplo la igualdad entre diferentes métricas de la matriz de confusión equivale a la igualdad de oportunidad; tasas iguales de falsos negativos y falsos positivos entre grupos equivale a satisfacer *Equality of Odds* [50].

7.4 Glosario

Término	Definición
Absolute Between-ROC Área, ABROCA	ABROCA mide el valor absoluto del área entre la curva ROC del grupo de referencia y las curvas ROC de uno o más grupos de comparación. De esta forma, ABROCA cuantifica la divergencia entre las curvas ROC de distintos grupos a través de todos los umbrales posibles, agregando esta divergencia sin importar qué subgrupo tiene mejor rendimiento en algún umbral específico. Esto permite evaluar imparcialidad en el rendimiento del modelo para diferentes subgrupos. Ver desarrollo de la técnica en https://homes.cs.washington.edu/~jpgard/papers/lak19_slicing.pdf
Accuracy	La accuracy, o exactitud, en el contexto de modelos de clasificación es la proporción de predicciones correctas sobre el total de predicciones realizadas. Es una medida de rendimiento que indica qué tan bien el modelo clasifica correctamente las instancias en el conjunto de datos.
Accuracy equity	Un sistema de IA muestra equidad en la precisión (*accuracy equity*) si puede discriminar de igual manera entre la clasificación posible de su espacio de salida para grupos diferentes.

Término	Definición
Accuracy macro-media, micro-media y media-ponderada	Un macro-media calculará la métrica de forma independiente para cada clase y, a continuación, sacará la media (por tanto, tratará a todas las clases por igual), mientras que una micro-media agregará las contribuciones de todas las clases para calcular la métrica media. En una clasificación multiclase, es preferible el micro-media si se sospecha que puede haber desequilibrio entre las clases (es decir, si tiene muchos más ejemplos de una clase que de otras). Por otro lado, para calcular una media ponderada, cada número del conjunto de datos se multiplica por un peso predeterminado antes de realizar el cálculo final.
ANOVA	*Analysis of Variance* (ANOVA). Fórmula estadística utilizada para comparar las varianzas entre las medias (o promedios) de distintos grupos. Se utiliza para determinar si existe alguna diferencia entre las medias de distintos grupos. Este cociente muestra la diferencia entre la varianza dentro del grupo y la varianza entre grupos, lo que en última instancia produce una cifra que permite concluir que se apoya o rechaza la hipótesis nula. Si hay una diferencia significativa entre los grupos, no se apoya la hipótesis nula.
Average odds difference	Es la media de la diferencia en las tasas de falsos positivos y verdaderos positivos entre los grupos no privilegiados y privilegiados. Un valor de 0 implica que ambos grupos se benefician por igual.
Balanced accuracy	Está métrica se puede utilizar para analizar la precisión (performance) de un modelo de clasificación. Se calcula sumando la ratio de verdaderos positivos con la ratio de verdaderos negativos y se divide entre dos. Es especialmente útil cuando las clases no están balanceadas. Cuanto más cerca está de 1 este valor, mejor es el modelo para realizar clasificaciones correctas.
Base Line model	Son modelos sencillos que funcionan como referencia para el sistema de IA. Su principal función es contextualizar el resultado de los entrenamientos del modelo de IA. Generalmente estos modelos carecen de complejidad y tiene poco poder predictivo. Sirven de referencia y para establecer comparaciones con el sistema de IA y permiten conocer mejor los datos de trabajo.
BIASeD	Se trata de una categorización o taxonomía de sesgos cognitivos conocidos, desde la perspectiva de los sistemas de IA. Este catálogo y el trabajo de investigación realizado, pretende alinear los sesgos en IA con los propios sesgos del género humano. El estudio completo se puede encontrar en https://arxiv.org/abs/2210.01122
BLEU	*BLEU es una métrica utilizada en traducción automática para evaluar la calidad de una traducción generada en comparación con una traducción de referencia. Su cálculo se basa en la coincidencia de n-gramas entre el texto traducido y el texto de referencia, proporcionando una puntuación que refleja la similitud y precisión de la traducción generada.*

Término	Definición
Central Limit Theorem	El teorema del límite central (CLT) afirma que la distribución de una variable muestral se aproxima a una distribución normal (es decir, a una "curva de campana") a medida que aumenta el tamaño de la muestra, suponiendo que todas las muestras son idénticas en tamaño e independientemente de la forma real de la distribución de la población. Dicho de otro modo, la CLT es una premisa estadística según la cual, dado un tamaño de muestra suficientemente grande de una población con un nivel finito de varianza, la media de todas las variables muestreadas de la misma población será aproximadamente igual a la media de toda la población.
Clasificación (clasificación binaria, multiclase, multietiqueta	La clasificación binaria implica predecir entre dos posibles categorías o etiquetas. La clasificación multiclase incluye más de dos etiquetas posibles, asignando una sola etiqueta a cada instancia. La clasificación multi-etiqueta permite asignar múltiples etiquetas a cada instancia, lo cual es útil en situaciones donde un mismo ejemplo puede pertenecer a más de una categoría simultáneamente.
Clustering	Como tipo de aprendizaje no supervisado, el *clustering* es el proceso de ordenar un grupo de objetos de tal manera que los objetos del mismo grupo (que se denomina cluster) sean más similares entre sí que a los objetos de cualquier otro grupo. Existen diferentes tipos de algoritmos de *clustering* (K-means, MeanShift, DBSCAN etc.).
Coeficiente Silueta medio	Este es un coeficiente que mide la calidad del agrupamiento, en el que valores más altos indican *clusters* mejor definidos, tomando valores entre -1 y +1. La interpretación es la siguiente: - Valores cercanos a -1 indican un agrupamiento incorrecto. - Valores cercanos a cero indican *clusters* traslapados. - Valores cercanos a +1 indican *clusters* altamente densos. Se calcula considerando: - a: Promedio de la distancia entre una muestra y todos los demás puntos de la misma clase (*cluster*) - b: distancia entre la muestra y todos los puntos del siguiente *cluster* más cercano. Así s = (b-a)/max(a,b)
Cross-entropía binaria	Se trata de una función objetivo, que mide el rendimiento de un modelo de clasificación cuya salida es un valor de probabilidad entre 0 y 1. La pérdida de entropía cruzada aumenta a medida que la probabilidad predicha diverge de la etiqueta real. Así, predecir una probabilidad de 0,012 cuando la etiqueta de observación real es 1 sería malo y daría lugar a un valor de pérdida alto. Un modelo perfecto tendría una pérdida logarítmica de 0.

Término	Definición
Curva de elevación (lift curve)	Mide el rendimiento de un clasificador elegido frente a un clasificador aleatorio. La curva muestra la relación entre el número de casos que se predijeron positivos y los que son efectivamente positivos y, por tanto, mide el rendimiento de un clasificador elegido frente a un clasificador aleatorio. El gráfico se construye con el número acumulado de casos (en orden descendente de probabilidad) en el eje de abscisas y el número acumulado de verdaderos positivos en el eje de ordenadas.
Curva de respuesta cumulativa	Esta curva muestra la proporción de ganancias en el total de positivos con respecto a la proporción de registros del conjunto de prueba. Así, podemos analizar qué proporción del conjunto de pruebas es necesaria para obtener un determinado porcentaje de ganancia en el total de positivos.
DCG (Discounted Cumulative Gain)	Se utiliza para sistemas de IA donde se establece un ranking u ordenación, es decir, cuando la puntuación verdadera de una muestra d es un valor discreto en una escala que mide la relevancia con respecto a una consulta q. Para una consulta q dada y las muestras D = $\{d_1, ..., d_n\}$, se considera la k-ésima muestra. La ganancia G_k mide la utilidad de esta muestra, mientras que el descuento D_k = 1/log(k+1) penaliza los documentos recuperados con un rango inferior. La suma de los términos de ganancia descontados G_kD_k para k = 1...n es la ganancia acumulada descontada (DCG)
Disparate impact	Es la relación de probabilidad de resultados favorables entre los grupos desfavorecidos y privilegiados, permitiendo detectar sesgos o errores de precisión entre ambos conjuntos, privilegiados y desfavorecidos.
Distancia Cook	La distancia de Cook es una medida utilizada para identificar valores atípicos en modelos de regresión. Calcula el cambio en las predicciones del modelo al eliminar una observación específica. Un valor alto en la distancia de Cook indica que esa observación tiene una gran influencia sobre el modelo, lo que puede indicar que es un valor atípico
Divergencia KL o Divergencia Kullback-Lieber	La característica principal de la divergencia KL, es medir como una distribución de probabilidad se diferencia de otra. Por su naturaleza, se trata de una métrica no simétrica, por lo que en sí misma no es una métrica o medida de distancia. Considerada como un enfoque de optimización se puede considerar que tiene dos formas: Forward y Reverse KL.
Especifididad y sensibilidad	La sensibilidad es la métrica que evalúa la capacidad de un modelo para predecir verdaderos positivos de cada categoría a predecir por nuestro modelo. La especificidad es la métrica que evalúa la capacidad de un modelo para predecir los verdaderos negativos de cada categoría disponible. Estas métricas se aplican a cualquier modelo categórico.

Término	Definición
F-beta	La puntuación F-beta es una métrica utilizada en el campo de la inteligencia artificial para evaluar el rendimiento de un modelo de aprendizaje automático en un problema de clasificación, teniendo en cuenta tanto la precisión (precisión) como el *recall* (también conocido como sensibilidad o true positive *rate*). Esta métrica es una medida combinada que proporciona un equilibrio entre precisión y *recall*.
F1 con media Ponderada	El promedio ponderado de la puntuación F1 (*Weighted Averages of F1 Score*) es una métrica que se utiliza cuando las clases están desequilibradas y se desea tener en cuenta la importancia relativa de cada clase en la evaluación global del modelo. La puntuación F1 es una métrica que combina tanto la precisión (precisión) como el *recall*. Representa el equilibrio entre la capacidad del modelo para identificar correctamente los casos positivos (*recall*) y la capacidad de clasificar correctamente los casos positivos entre todas las predicciones positivas (precisión).
Prueba exacta de Fisher	La prueba exacta de Fisher es una prueba estadística utilizada para determinar si existe una asociación significativa entre dos variables categóricas en una tabla de contingencia 2x2. Es especialmente útil cuando los recuentos de celdas son bajos, ya que no requiere los supuestos de la prueba Chi-cuadrado.
Focal loss	Una función de pérdida focal aborda el desequilibrio de clase durante el entrenamiento en tareas como la detección de objetos. La pérdida focal aplica un término modulador a la pérdida de entropía cruzada para enfocar el aprendizaje en ejemplos mal clasificados. Es una pérdida de entropía cruzada escalada dinámicamente, donde el factor de escala decae a cero a medida que aumenta la confianza en la clase correcta. Intuitivamente, este factor de escala puede reducir automáticamente la contribución de ejemplos fáciles durante el entrenamiento y enfocar rápidamente el modelo en ejemplos difíciles.
Frechet Information Distance (FID)	La *Frechet Inception Distance*, o FID para abreviar, es una métrica para evaluar la calidad de las imágenes generadas y desarrollada específicamente para evaluar el rendimiento de las redes generativas antagónicas. Fue propuesta como una mejora con respecto a *Inception Score* (IS).
GAN Condicional (CGAN)	Las GAN condicionales (CGAN) son una variante de las Redes Generativas Adversariales (GAN) que incorporan una variable de condición adicional para permitir la generación controlada y condicionada de datos sintéticos. Esto permite la generación específica de datos basada en información adicional suministrada, lo que amplía las capacidades de las GAN en diversas aplicaciones de generación de datos.
Hamming Loss	Es una métrica que mide la precisión de un modelo de clasificación multi-etiqueta al calcular la tasa promedio de etiquetas clasificadas incorrectamente. Proporciona una medida general del rendimiento del

Término	Definición
	modelo en la clasificación multi-etiqueta, independientemente del desequilibrio de las etiquetas.
Hinge embedding loss	Es una función de pérdida utilizada en algoritmos de clasificación binaria, como las máquinas de vectores de soporte (SVM). Su objetivo es maximizar el margen entre las clases y penalizar las clasificaciones incorrectas y las muestras cercanas al margen.
Hubber loss	Es una función de pérdida utilizada en el aprendizaje automático. A diferencia de otras funciones de pérdida, como el Error Cuadrático Medio (MSE), el Huber Loss es menos sensible a los valores atípicos en los datos. El Huber Loss combina las ventajas del Error Cuadrático Medio y del Error Absoluto Medio, adaptando su comportamiento en función de un parámetro de tolerancia.
Hyperband (tool)	Hyperband es un algoritmo de optimización de hiperparámetros que asigna recursos de forma selectiva a diferentes configuraciones de hiperparámetros. Emplea un enfoque de eliminación temprana, descartando rápidamente configuraciones de bajo rendimiento y asignando más recursos a configuraciones prometedoras, lo que permite una búsqueda de hiperparámetros más rápida y eficiente.
Hyperopt	Es una biblioteca de Python que se utiliza para la optimización automática de hiperparámetros. Proporciona herramientas y algoritmos que permiten realizar una búsqueda eficiente de hiperparámetros en el aprendizaje automático.
Inception Score (IS)	Es una métrica utilizada para evaluar la calidad y diversidad de las imágenes generadas por modelos de generación de imágenes, como las GAN. Se basa en la clasificación de imágenes mediante una red Inception pre-entrenada y combina medidas de diversidad y calidad para proporcionar una puntuación que indica la calidad general de las imágenes generadas.
Índice de Fowlkes-Mallows	Es una métrica utilizada para evaluar la calidad de las agrupaciones obtenidas en el análisis de datos. Se basa en la comparación de las etiquetas reales y las etiquetas asignadas por un algoritmo de agrupamiento, y calcula un valor que indica la similitud entre las agrupaciones. Un valor más cercano a 1 indica una mayor similitud, mientras que un valor de 0 indica diferencias completas entre las agrupaciones.
Índice de Jaccard (también IoU)	Es una métrica que mide la similitud entre dos conjuntos o regiones. Se utiliza para calcular la superposición entre los conjuntos dividiendo el tamaño de la intersección entre el tamaño de la unión. Es ampliamente utilizado en el campo de la visión por computadora y otras áreas donde se necesita medir la similitud o la coincidencia entre conjuntos.

Término	Definición
Índice de Rand	Es una métrica utilizada para evaluar la similitud entre dos agrupaciones. Calcula la proporción de pares de datos que están asignados de la misma manera en ambas agrupaciones. Un Índice de Rand más alto indica una mayor similitud entre las agrupaciones, mientras que un Índice de Rand más bajo indica una menor similitud o una asignación aleatoria de los clústeres.
Información mutua Ajustada	La información mutua ajustada es una métrica utilizada en el análisis de agrupamientos para medir la similitud entre dos conjuntos de etiquetas. Esta métrica ajusta la información mutua para corregir por azar, proporcionando una estimación de la dependencia entre las etiquetas sin inflarse cuando el número de agrupamientos es alto.
Kruskal-Wallis	Es una prueba no paramétrica utilizada para determinar si hay diferencias significativas entre las medianas de dos o más grupos independientes. Se basa en la comparación de los rangos de los datos y utiliza una estadística de prueba para evaluar la hipótesis nula de igualdad de medianas.
Language Model Probability	La probabilidad de modelo de lenguaje es una métrica que evalúa qué tan probable es que un modelo de lenguaje genere una secuencia específica de palabras. Esta métrica se basa en las probabilidades condicionales de cada palabra dada su contexto previo y mide tanto la coherencia como la naturalidad de las secuencias generadas.
LogCosh (logaritmo coseno hiperbólico)	Es una función de pérdida suave utilizada en problemas de optimización en aprendizaje automático. Es menos sensible a valores atípicos y busca minimizar la diferencia entre las predicciones y los valores verdaderos de manera estable.
Mean Average Error (MAE)	Es una métrica de evaluación que mide el promedio de las diferencias absolutas entre las predicciones y los valores reales. Es una medida comúnmente utilizada en problemas de regresión para evaluar la precisión de un modelo de aprendizaje automático.
Mean Absolute Percentage Error (MAPE)	Es uno de los KPI más utilizados para medir la precisión del pronóstico. MAPE es la suma de los errores absolutos individuales dividida por la demanda (cada período por separado). Es el promedio de los errores porcentuales.
Margin Ranking Loss	Es una función de pérdida utilizada en problemas de clasificación y aprendizaje por pares. Su objetivo es maximizar la diferencia entre las puntuaciones de elementos similares y no similares para entrenar modelos que puedan clasificar elementos o comparar pares de manera efectiva.
Matriz de confusión	La matriz de confusión es una tabla que muestra las predicciones de un modelo de clasificación en comparación con los valores reales. Tiene cuatro celdas: verdaderos positivos, falsos positivos, falsos negativos y verdaderos negativos. Esta matriz permite evaluar el rendimiento del modelo y calcular métricas como precisión, exhaustividad y puntuación F1. Es una herramienta útil para entender cómo el modelo clasifica diferentes clases y mejorar su precisión.

Término	Definición
Matriz de contingencia	Una matriz de contingencia es una tabla que muestra la distribución conjunta de dos o más variables categóricas. Permite visualizar la relación y las frecuencias de cada combinación de categorías. Es útil para analizar asociaciones entre variables y revelar patrones en los datos.
McNemar	Es una prueba no paramétrica para datos nominales emparejados. Se usa cuando está interesado en encontrar un cambio en la proporción de los datos emparejados. Por ejemplo, podría utilizar esta prueba para analizar estudios retrospectivos de casos y controles, en los que cada tratamiento se empareja con un control. También podría usarse para analizar un experimento en el que se administran dos tratamientos a pares emparejados. Esta prueba a veces se denomina prueba Chi-Cuadrado de McNemar porque la estadística de prueba tiene una distribución de Chi-cuadrado.
Mean Squared Error (MSE o L2 loss)	Es una métrica de evaluación que mide el promedio de las diferencias al cuadrado entre las predicciones y los valores reales en un problema de regresión. Es ampliamente utilizado y penaliza más los errores grandes, pero puede ser sensible a la escala de los datos.
Mean Squared Logarithmic Error	Es una métrica de evaluación utilizada en problemas de regresión. Calcula la diferencia al cuadrado entre los logaritmos de las predicciones y los logaritmos de los valores reales. Es útil para penalizar más los errores en los extremos y cuando se desea enfocarse en la precisión relativa en lugar de la absoluta.
METEOR	*Metric for Evaluation of Translation with Explicit ORdering*, es una métrica de evaluación de traducción que tiene en cuenta tanto la fluidez del texto como la correspondencia semántica entre la traducción y la referencia.
Métrica de Completitud	También conocida como *recall*, mide la capacidad de un modelo para identificar todos los elementos positivos en un conjunto de datos. Se calcula como la proporción de verdaderos positivos sobre la suma de verdaderos positivos y falsos negativos. Una completitud alta indica una buena capacidad para identificar elementos positivos, pero debe considerarse en conjunto con otras métricas para evaluar el rendimiento general del modelo.
Min-max Loss	Es una función de pérdida utilizada en problemas de clasificación. Penaliza las predicciones incorrectas al maximizar la diferencia entre la probabilidad asignada a la clase correcta y la probabilidad máxima entre las clases incorrectas. Su objetivo es mejorar la capacidad del modelo para distinguir entre las clases y reducir los errores de clasificación.
Modelo de espacio vectorial (VSM)	Es un enfoque utilizado en el procesamiento del lenguaje natural para representar y analizar documentos de texto. Los documentos se representan como vectores en un espacio multidimensional, donde cada dimensión corresponde a un término o una característica del texto. Esto permite calcular similitudes o distancias entre los documentos y realizar diversas tareas de análisis de textos.
NIST LRE	NIST *Language Recognition Evaluation*, el objetivo de esta métrica es establecer una medida de base para el rendimiento y la capacidad de reconocimiento de lenguaje hablado a través del teléfono.

Término	Definición
Non saturating GAN loss	Se utiliza en GANs para entrenar el generador de imágenes y superar problemas de saturación y desvanecimiento del gradiente. Proporciona una alternativa efectiva a la pérdida clásica de GAN y ha demostrado ser útil en el entrenamiento de generadores más estables y de alta calidad.
Optuna (tool)	Es una biblioteca de optimización de hiperparámetros que se utiliza para automatizar y facilitar la búsqueda de la configuración óptima de los hiperparámetros de un modelo de aprendizaje automático.
Paired student T-Test	Es una prueba estadística utilizada para comparar las medias de dos conjuntos de datos relacionados. Se utiliza para determinar si hay una diferencia significativa entre los dos conjuntos de datos y se basa en el supuesto de distribución normal de los datos emparejados.
Plotly sunburst plot (tool)	Es una representación visual interactiva que muestra la estructura jerárquica de los datos en forma de anillos concéntricos. Cada anillo representa una categoría o subcategoría y el tamaño de las porciones en el gráfico refleja una métrica específica. Es una herramienta útil para visualizar datos jerárquicos y explorar su distribución de manera intuitiva.
Pointwise Methods, Pairwise Methods, Listwise Methods	Todos los modelos de *Learning to Rank* utilizan un modelo base de aprendizaje para calcular s = f (x). La elección de la función de pérdida es el elemento distintivo para los modelos *Learning to Rank*. En general, tenemos 3 enfoques, dependiendo de cómo se calcule la pérdida. *Pointwise Methods*: la pérdida total se calcula como la suma de los términos de pérdida definidos en cada documento di (por lo tanto, puntualmente) como la distancia entre la puntuación predicha si y la verdad fundamental yi, para i = 1 ... Al hacer esto, transformamos nuestra tarea en un problema de regresión, donde entrenamos un modelo para predecir. *Pairwise Methods*: la pérdida total se calcula como la suma de los términos de pérdida definidos en cada par de documentos di, dj (por lo tanto, por pares), para i, j = 1 ... n. El objetivo sobre el que se entrena el modelo es predecir si yi > yj o no, es decir, cuál de los dos documentos es más relevante. Al hacer esto, transformamos nuestra tarea en un problema de clasificación binaria. *Listwise Methods*: la pérdida se calcula directamente en toda la lista de documentos (por lo tanto, *listwise*) con los rangos previstos correspondientes. De esta manera, las métricas de clasificación se pueden incorporar más directamente a la pérdida.
Precisión - Recall (curva) PRC y Área debajo PRC (AUPRC)	La precisión-*recall* (PRC) es una métrica que evalúa el rendimiento de un modelo de clasificación en términos de precisión y *recall* a medida que se varía el umbral de clasificación. La curva PRC traza estos valores, y el área debajo de la curva (AUPRC) resume el rendimiento general del modelo.

Término	Definición
Predictive parity	Métrica de equidad que comprueba si, para un clasificador determinado, los índices de precisión son equivalentes para los subgrupos considerados. Por ejemplo, un modelo que predice la aceptación en la universidad satisfaría la paridad predictiva para la nacionalidad si su índice de precisión es el mismo independiente de la nacionalidad de origen A veces, la paridad predictiva también se denomina paridad de tasa predictiva.
R2	R^2 es una métrica estadística que mide la proporción de variación en la variable dependiente que es explicada por el modelo de regresión. Indica qué tan bien se ajustan los datos al modelo, con valores de 0 a 1, donde un valor más cercano a 1 sugiere un mejor ajuste.
R2 ajustada	Es una medida corregida de la bondad de ajuste de un modelo lineal. Se utiliza para evaluar la precisión del modelo y determinar el porcentaje de varianza en la variable dependiente que se explica mediante las variables independientes. El R2 tiende a sobreestimar el ajuste del modelo de regresión lineal y siempre aumenta a medida que se incluyen más variables en el modelo. El R2 ajustado intenta corregir esta sobreestimación y puede disminuir si la inclusión de una variable no mejora el modelo.
Ranking con signo Wilcoxon	Es una prueba estadística no paramétrica utilizada para comparar dos muestras pareadas y determinar si hay una diferencia significativa entre ellas. Es una alternativa útil cuando los datos no cumplen con los supuestos de normalidad o cuando se trabaja con escalas ordinales.
Regresión	La regresión es una técnica estadística y de aprendizaje supervisado que busca modelar la relación entre una variable dependiente continua y una o más variables independientes. La regresión se utiliza para realizar predicciones y analizar cómo las variables independientes influyen en la variable dependiente.
Reweighting	Es una técnica que minimiza el sesgo, ajusta los pesos o las probabilidades de las observaciones en un conjunto de datos para abordar desequilibrios o mejorar la representatividad de ciertas clases o instancias.

Término	Definición
RMSE	Es una métrica que cuantifica el error promedio entre las predicciones de un modelo de regresión y los valores reales del conjunto de datos. Es utilizado para evaluar y comparar la precisión de diferentes modelos, siendo deseable un valor de RMSE lo más bajo posible.
ROC (curva) Y Área debajo ROC (AUROC o AUC)	La curva ROC es una representación gráfica que muestra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos a medida que se varía el umbral de clasificación. El AUROC es una métrica que resume la calidad de la curva ROC y proporciona una medida cuantitativa del rendimiento del modelo de clasificación binaria. Cuanto mayor sea el valor del AUROC, mejor será el rendimiento del modelo.
Similitud coseno	Es una medida que evalúa la similitud entre dos vectores en función del ángulo entre ellos. Es una medida comúnmente utilizada en diversas aplicaciones para comparar la similitud entre textos, documentos, perfiles de usuario y otros tipos de datos representados en forma vectorial.
Triplet Margin Loss	También conocida como *Triplet Loss* o Pérdida de Margen Triplete, es una función de pérdida utilizada en el aprendizaje automático para aprender representaciones de datos donde las instancias similares están más cercanas entre sí y las instancias diferentes están más separadas. Utiliza triplets de datos y se basa en la comparación de distancias en el espacio de representación.
Unequal outcomes	Hace referencia a situaciones en las que los modelos de *machine learning* producen resultados desiguales o sesgados para diferente grupos o individuos.
Valor Calinsk y Harabaz	Este índice es una medida utilizada en el campo de la estadística y el aprendizaje automático para evaluar la calidad de una agrupación o *clustering* de datos. Se basa en la idea de que un buen *clustering* debería tener una alta cohesión *intra-cluster* (los puntos dentro de cada grupo son similares) y una baja separación *inter-cluster* (los grupos están bien diferenciados entre sí). Cuanto mayor sea el valor del índice, se considera que el *clustering* es de mejor calidad.
Valor F1	El F1 score es una medida estadística que combina la precisión y el recuerdo (*recall*) de un modelo de clasificación. Es útil cuando se tienen clases desequilibradas en los datos. Proporciona una única métrica que representa la precisión y el recuerdo de manera equilibrada, y su valor oscila entre 0 y 1, siendo 1 el mejor resultado posible. Un F1 score alto indica un equilibrio entre la precisión y el recuerdo del modelo en la clasificación de las clases.
Wasserstein GAN loss	Se trata de una de las alternativas más potentes a la pérdida GAN original. Aborda el problema del colapso de modo y la desaparición del gradiente. En esta implementación, la activación de la capa de salida del discriminador se cambia de sigmoidea a lineal. Este simple cambio influye en el discriminador para que emita una puntuación en lugar de una probabilidad asociada a la distribución de los datos, por lo que la salida no tiene que estar en el rango de 0 a 1.

Término	Definición
Word Accuracy	Word Accuracy mide la fracción de palabras que un sistema de reconocimiento o procesamiento de texto clasifica o identifica correctamente, en comparación con una referencia. Es una métrica que evalúa el rendimiento general en la identificación precisa de palabras sin considerar precisión ni recall.
Word Error Rate	Es una medida que cuantifica la precisión del reconocimiento automático de voz al comparar la transcripción generada por el sistema con una transcripción de referencia, brindando una forma de evaluar y comparar la calidad de diferentes sistemas de reconocimiento de voz.

8. Referencias, estándares y normas

8.1 Referencias

8.1.1 Referencias generales

[1] Hullermeier et al. 2019 Aleatoric and Epistemic Uncertainty in Machine Learning: A Tutorial Introduction.

[2] Calibration of Machine Learning Models. Antonio Bella, Cèsar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia 2010

[3] Metrics for Deep Generative Models N Chen · 2018

[4] Language Models are Few-Shot Learners T Brown et al. 2020

[5] Physically Consistent Generative Adversarial Networks for Coastal Flood Visualization Bjorn Lutjens et al. 2021

[6] IBM Uncertainty Quantification 360 Toolkit https://uq360.mybluemix.net

[7] Questioning causality on sex, gender, and COVID-19, and identifying bias in large-scale data-driven analyses: the Bias Priority Recommendations and Bias Catalog for Pandemics. Díaz-Rodríguez et al. 2021 https://arxiv.org/abs/2104.14492

[8] https://catalogofbias.org

[9] A survey on datasets for fairness-aware machine learning Tai Le Quy∗1, Arjun Roy†12, Vasileios Iosifidis‡1, Wenbin Zhang§3, and Eirini Ntoutsi 2022

[10] Datasheets for Datasets. Timnit Gebru et al. 2021

[11] "Searching for a Search Method: Benchmarking Search Algorithms for Generating NLP Adversarial Examples",

[12] EMNLP 2020 Blackbox NLP Workshop track proceedings. https://github.com/QData/TextAttack

[13] TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP, J Morris et al. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020 [In Python]

[14] PLENARY: Explaining black-box models in natural language through fuzzy linguistic summaries K Kaczmarek-Majer, G Casalino, G Castellano… - Information …, 2022 - Elsevier

[15] Companies Committed to Responsible AI: From Principles towards Implementation and Regulation? Paul B. de Laat Philosophy & Technology volume 34, pages 1135–1193 (2021)

Note: SHAP is not from Amazon nor proprietary in Table 3 (SHAP is Scott Lundberg's creation with a MIT open source license, in Github, and was developed with Microsoft Research teams. AzureML also implements XAI algorithms and Amazon has as ML tool, Amazon SageMaker).

[16] COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity 2016

[17] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learn Research), Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New ing

[18] An Ontology for Fairness Metrics. Franklin et al. https://dl.acm.org/doi/pdf/10.1145/3514094.3534137

[19] Human-centred artificial intelligence https://scilog.fwf.ac.at/en/environment-and-technology/15317/human-centred-artificial-intelligence

[20] A Holzinger et al. Digital Transformation in Smart Farm and Forest Operations Needs Human-Centered AI: Challenges and Future Directions

[21] Holzinger, A. 2016. Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? Brain Informatics, 3, (2), 119-131, doi:10.1007/s40708-016-0042-6.

[22] Holzinger, A., Malle, B., Kieseberg, P., Roth, P.M., Müller, H., Reihs, R. & Zatloukal, K. 2017. Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology. arXiv:1712.06657.

[23] Holzinger, A., Malle, B., Kieseberg, P., Roth, P. M., Müller, H., Reihs, R. & Zatloukal, K. 2017. Machine Learning and Knowledge Extraction in Digital Pathology needs an integrative approach. Springer Lecture Notes in Artificial Intelligence Volume LNAI 10344. Cham: Springer International, pp. 13-50. doi: 10.1007/978-3-319-69775-8_2

[24] Human-Centered Artificial Intelligence for Designing Accessible Cultural Heritage G Pisoni, N Díaz-Rodríguez, H Gijlers, L Tonolli Applied Sciences 11 (2), 870

[25] Continual Learning for Robotics: Definition, Framework, Learning Strategies, Opportunities and Challenges T Lesort, V Lomonaco, A Stoian, D Maltoni, D Filliat, N Díaz-Rodríguez Information Fusion 220

[26] 2020 A survey on ontologies for human behavior recognition ND Rodríguez, MP Cuéllar, J Lilius, MD Calvo-Flores ACM Computing Surveys (CSUR) 46 (4), 1-33 219 2014 A fuzzy ontology for semantic modelling and recognition of human behaviour ND Rodríguez, MP Cuéllar, J Lilius, MD Calvo-Flores Knowledge-Based Systems 66, 46-60 133 201

[27] Explainability in Deep Reinforcement Learning A Heuillet, F Couthouis, N Díaz-Rodríguez Knowledge-Based Systems 214, 106685 119 2021

[28] Don't forget, there is more than forgetting: new metrics for Continual Learning N Díaz-Rodríguez, V Lomonaco, D Filliat, D Maltoni NeurIPS workshop on Continual Learning 2018

[29] Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence A Holzinger, M Dehmer, F Emmert-Streib, R Cucchiara, I Augenstein, ... Information Fusion 79, 263-278 2022

[30] Personas for Artificial Intelligence (AI) An Open Source Toolbox A Holzinger, M Kargl, B Kipperer, P Regitnig, M Plass, H Müller IEEE Access 10, 23732-23747 2022

[31] Measuring the quality of explanations: the system causability scale (SCS) A Holzinger, A Carrington, H Müller KI-Künstliche Intelligenz 34 (2), 193-19

[32] Interdisciplinary Research in Artificial Intelligence: Challenges and Opportunities R Kusters, D Misevic, H Berry, A Cully, Y Le Cunff, L Dandoy, ... Frontiers in Big Data 3, 45 2020

[33] S. Stanton, P. Izmailov, P. Kirichenko, A. A. Alemi, A. G. Wilson, Does knowledge distillation really work? (2021). doi:10.48550/ARXIV.2106.05945. URL https://arxiv.org/abs/2106.05945

[34] A Neural-Symbolic learning framework to produce interpretable predictions for image classification, PhD Thesis 2022

[35] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. AB Arrieta, N Díaz-Rodríguez, J Del Ser, A Bennetot, S Tabik, A Barbado, ...Information Fusion 58, 82-115

[36] Wilcoxon, Frank. Individual Comparisons by Ranking Methods. Biometrics Bulletin. 1945, 1 (6), 1095 pages 80-83.

[37] Dietterich et al. Approximate Statistical Tests for Comparing Supervised Classification 1092 Learning Algorithms. Neural Computation, Volume 10, Issue 7. 1998, 10 (7), pages 1895-1923. 1093 https://doi.org/10.1162/089976698300017197

[38] Akenine-Möller, Tomas, and Johnsson, Björn. Performance per what? Journal of Computer Graphics 1073 Techniques. 2012, 1, pages 37-41. http://jcgt.org/published/0001/01/03/paper.pdf

[39] Blouw, Peter and Xuan Choo and Hunsberger, Eric and Eliasmith, Chris. Benchmarking keyword 1075 spotting efficiency on neuromorphic hardware. Proceedings of the 7th Annual Neuro-inspired 1076 Computational Elements Workshop. 2019, pages 1-8. https://arxiv.org/pdf/1812.01739.pdf

[40] Suffering-focused AI safety: In favor of "fail-safe" measures Lukas Gloor Center on Long-Term Risk Report

[41] Superintelligence as a Cause or Cure for Risks of Astronomical Suffering

[42] Kaj Sotala and Lukas Gloor Foundational Research Institute, Berlin, Germany Superintelligence as a Cause or Cure ... Informatica 41 (2017) 389–400

[43] Safe Deep RL in 3D environments using human feedback. 2022.

[44] Safeguard By Design Lessons Learned from DOE Experience Integrating Safety in Design

[45] Sustainability at scale: towards bridging the intention-behavior gap with sustainable recommendations S Tomkins, S Isley, B London, L Getoor - Proceedings of the 12th ACM conference on …, 2018

[46] Green AI. Schwartz et al.

[47] Distilling the Knowledge in a Neural Network

[48] EVALUATION METRICS FOR LANGUAGE MODELS Stanley Chen, Douglas Beeferman, Ronald Rosenfeld.

[49] Chip Huyen, "Evaluation Metrics for Language Modeling", The Gradient, 2019. https://thegradient.pub/understandSing-evaluation-metrics-for-language-models/

[50] Model cards for model reporting. M Mitchell, S Wu, A Zaldivar, P Barnes, L Vasserman… - Proceedings of the …, 2019

[51] Frank McSherry Materialize: a platform for building scalable event based systems

[52] Frank McSherry, Kunal Talwar Mechanism design via Differential Privacy, 2008

[53] Nobel Prize Report, Mechanism Design. 2007 Scientific background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2007 Mechanism Design Theory, based in:

[54] L. Hurwicz & S. Reiter (2006) Designing Economic Mechanisms, p. 30

[55] https://divedeep.ai/2022/03/17/data-drift-vs-concept-drift/

[56] University of Oxford researchers have created a tool called capAI, a procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act. CapAI provides organizations with practical guidance on how to translate high-level ethics principles into verifiable criteria that help shape the design, development, deployment and use of ethical AI. This tool can be used to demonstrate that the development and operation of an AI system are trustworthy. The tool is being validated with firms at the moment and the most up-to-date version can be found here

[57] A survey on concept drift adaptation ACM computing surveys (CSUR), 46(4):1–37, 2014, Gama et al.

[58] Analysis of representations for domain adaptation Neurips 2007, Ben-David et al

[59] Dataset Shift in Machine Learning Quiñonero-Candela et al 2022

[60] Generalized out-of-distribution detection: A survey. Yang et al 2022

[61] Understanding Continual Learning Settings with Data Distribution Drift Analysis" Lesort et al. 2022 video: https://www.youtube.com/watch?v=WFhozvAgnsU

[62] Sagemaker Clarify: Amazon AI Fairness and Explainability Whitepaper https://pages.awscloud.com/rs/112-TZM-766/images/Amazon.AI.Fairness.and.Explainability.Whitepaper.pdf https://aws.amazon.com/blogs/machine-learning/learn-how-amazon-sagemaker-clarify-helps-detect-bias/

[63] Data Privacy and Trustworthy Machine Learning. Strobel et al. 2022

[64] Gradual (In)Compatibility of Fairness Criteria Corinna Hertweck, Tim Räz 2022 https://arxiv.org/abs/2109.04399

[65] Decoupling feature extraction from policy learning: assessing benefits of state representation learning in goal based robotics A Raffin, A Hill, R Traoré, T Lesort, N Díaz-Rodríguez, D Filliat ICLR 2019 Workshop on Structure & Priors in Reinforcement Learning (SPIRL) 2019

[66] Continual reinforcement learning deployed in real-life using policy distillation and sim2real transfer RT Kalifou, H Caselles-Dupré, T Lesort, T Sun, N Diaz-Rodriguez, D Filliat ICML Workshop on Multi-Task and Lifelong Learning 2019

[67] S-RL Toolbox: Environments, Datasets and Evaluation Metrics for State Representation Learning

[68] A Raffin, A Hill, R Traoré, T Lesort, N Díaz-Rodríguez, D Filliat NeurIPS workshop on Deep Reinforcement Learning 2018

[69] Deep Unsupervised state representation learning with robotic priors: a robustness analysis

[70] T Lesort, M Seurin, X Li, N Díaz-Rodríguez, D Filliat. 2019 International Joint Conference on Neural Networks (IJCNN) 2017

[71] Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence A Holzinger, M Dehmer, F Emmert-Streib, R Cucchiara, I Augenstein, et al. Information Fusion.

[72] Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study Zech et al 2018 (extendido de Confounding variables can degrade generalization performance of radiological deep learning models. Zech et al. 2018.)

[73] ISO/IEC 25000, Systems and software engineering – Systems and software Quality 937 Requirements and Evaluation (SQuaRE) and ISO/IEC WD 25059:2021, Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) Quality Model for AI systems

[74] STRIDE-AI: An Approach to Identifying Vulnerabilities of Machine Learning Assets Lara Mauri et al. IEEE CSR 2021

[75] Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks Papernot 2016

[76] Steps Toward Robust Artificial Intelligence Thomas G. Dietterich 2017

[77] Improving the Robustness of Deep Neural Networks via Stability Training

[78] SECURING MACHINE LEARNING ALGORITHMS. December 2021 ANNEX D: REFERENCES by input data type and lifecycle stages.

[79] Towards Resilient Artificial Intelligence: Survey and Research Issues

[80] The Robustness of Counterfactual Explanations Over Time, A Ferrario et al.

[81] Research priorities for robust and beneficial artificial intelligence.

[82] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. 2016.

[83] Evaluating Robustness of Counterfactual Explanations Artelt et al 2021

[84] Exploring the Trade-off between Plausibility, Change Intensity and Adversarial Power in Counterfactual Explanations using Multi-objective Optimization. Javier Del Ser et al. 2020

[85] Towards Human-Compatible XAI: Explaining Data Di erentials with Concept Induction over Background Knowledge. Widmer et al 2022

[86] A survey on bias in visual datasets 2022. Fabrizzi et al.

[87] Omitted variable bias: A threat to estimating causal relationships. Wilms et al.

[88] Google. Machine Learning Glossary: Fairness. 2021 [cited 29 November, 2021]; available from: https://developers.google.com/machine-learning/glossary/fairness.

[89] David Lopez-Paz, Krikamol Muandet, Bernhard Scholkopf, and Ilya O. Tolstikhin. Towards a learning theory of cause-e ect inference. In Francis R. Bach and David M. Blei, editors, Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, volume 37 of JMLR Workshop and Conference Proceedings, pages 1452{1461. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/ lopez-paz15.html.

[90] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Sch□olkopf, and Leon Bottou. Discovering causal signals in images. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 58{ 66, 2017. doi: 10.1109/CVPR.2017.14.

[91] ICO, Guidance on the AI auditing framework: draft guidance for consultation. Information
[92] Commissioner's Office, 2020.

[93] PwC, PwC Ethical AI Framework. 2020.

[94] Deloitte, Deloitte introduces trustworthy AI framework to guide organizations in ethical application of technology. August 26, 2020. New York.

[95] Orcaa, It's the age of the algorithm and we have arrived unprepared. 2020.

[96] Epstein, Z., et al., Turingbox: an experimental platform for the evaluation of AI systems. IJCAI International Joint Conference on Artificial Intelligence, 2018. 2018-July: p. 5826-5828. #Discontinued.

[97]. Shi et al. Robustness Verification for Transformers. International Conference on Learning Representations. 2020. arXiv:2002.06622

[98] Incremental Bounded Model Checking of Artificial Neural Networks in CUDA Luiz H. Sena et al.

[99] A Raffin, A Hill, R Traoré, T Lesort, N Díaz-Rodríguez, D Filliat

[100] NeurIPS workshop on Deep Reinforcement Learning 2018

[101] Stable-Baselines3 Reliable Reinforcement Learning Implementations https://stable-baselines3.readthedocs.io/en/master/

[102] T Lesort, M Seurin, X Li, N Díaz-Rodríguez, D Filliat 2019 International Joint Conference on Neural Networks (IJCNN) 2017

[103] Error Analysis tool, part of the Responsible AI Dashboard in Azure: https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai-dashboard

[104] Evolved from "Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure Besmira Nushi Ece Kamar Eric Horvitz HCOM 2018.

[105] Deep Reinforcement Learning that Matters - P Henderson · 2017

[106] L. Hurwicz & S. Reiter (2006) Designing Economic Mechanisms,

[107] Facial Recognition: Analyzing Gender and Intersectionality in Machine Learning. Report. http://genderedinnovations.stanford.edu/case-studies/facial.html#tabs-2

[108] Piggyback: Adapting a Single Network to Multiple Tasks by Learning to Mask Weights. 2018

[109] AS Ross, MC Hughes, F Doshi-Velez. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. IJCAI'17.

[110] K Burns, LA Hendricks, K Saenko, T Darrell, A Rohrbach. Women also Snowboard: Overcoming Bias in Captioning Models. ECCV'18, 771-787.

[111] A Rohrbach, LA Hendricks, K Burns, T Darrell, K Saenko. Object Hallucination in Image Captioning. EMNLP'18.

[112] A Gulati, MA Lozano, B Lepri, N Oliver. BIASeD: Bringing Irrationality into Automated System Design

[113] J Gardner, C Brooks, R Baker. Evaluating the Fairness of Predictive Student Models Through Slicing Analysis

8.1.2 Referencias del Glosario

Para el glosario, se han utilizado las siguientes referencias, que pueden ser consultadas para ampliar los aspectos allí descritos.

- https://homes.cs.washington.edu/~jpgard/papers/lak19_slicing.pdf,
- https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf
- https://www.tibco.com/reference-center/what-is-analysis-of-variance-anova
- https://medium.com/sfu-cspmp/model-transparency-fairness-552a747b444
- https://www.statology.org/balanced-accuracy/
- https://towardsdatascience.com/baseline-models-your-guide-for-model-building-1ec3aa244b8d
- https://arxiv.org/abs/2210.01122
- https://www.investopedia.com/terms/c/central_limit_theorem.asp
- https://machinelearningmastery.com/types-of-classification-in-machine-learning/
- https://www.datacamp.com/blog/clustering-in-machine-learning-5-essential-clustering-algorithms
- https://jdvelasq.github.io/courses/notebooks/sklearn_unsupervised_03_clustering/1-03_metodo_de_la_silueta.html

- https://ml-cheatsheet.readthedocs.io/en/latest/loss_functions.html
- https://orange3.readthedocs.io/en/3.5.0/widgets/evaluation/liftcurve.html
- https://rstudio-pubs-static.s3.amazonaws.com/577248_f94b111668f546e896649e408011969d.html
- https://towardsdatascience.com/learning-to-rank-a-complete-guide-to-ranking-using-machine-learning-4c9688d370d4
- https://medium.com/sfu-cspmp/model-transparency-fairness-552a747b444
- https://keepcoding.io/blog/que-es-la-distancia-de-cook/
- https://github.com/christianversloot/machine-learning-articles/blob/main/how-to-use-kullback-leibler-divergence-kl-divergence-with-keras.md
- https://towardsdatascience.com/evaluating-categorical-models-ii-sensitivity-and-specificity-e181e573cff8#:~:text=Sensitivity%20is%20the%20metric%20that,negatives%20of%20each%20available%20category.
- https://machinelearningmastery.com/fbeta-measure-for-machine-learning/
- https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f
- https://www.statology.org/fishers-exact-test/
- https://paperswithcode.com/method/focal-loss#:~:text=Focal%20loss%20applies%20a%20modulating,in%20the%20correct%20class%20increases.
- https://machinelearningmastery.com/how-to-implement-the-frechet-inception-distance-fid-from-scratch/#:~:text=for%20Real%20Images-,What%20Is%20the%20Frechet%20Inception%20Distance%3F,performance%20of%20generative%20adversarial%20networks.
- https://machinelearningmastery.com/how-to-develop-a-conditional-generative-adversarial-network-from-scratch/
- https://www.linkedin.com/pulse/hamming-score-multi-label-classification-chandra-sharat/
- https://medium.com/udacity-pytorch-challengers/a-brief-overview-of-loss-functions-in-pytorch-c0ddb78068f7
- https://www.cantorsparadise.com/huber-loss-why-is-it-like-how-it-is-dcbe47936473
- https://arxiv.org/abs/1603.06560
- https://github.com/hyperopt/hyperopt
- https://medium.com/octavian-ai/a-simple-explanation-of-the-inception-score-372dff6a8c7a
- http://sedici.unlp.edu.ar/bitstream/handle/10915/76139/Documento_completo.pdf?sequence=1
- https://deepai.org/machine-learning-glossary-and-terms/jaccard-index
- https://towardsdatascience.com/7-evaluation-metrics-for-clustering-algorithms-bdc537ff54d2
- https://cloud.google.com/vertex-ai/docs/tabular-data/tabular-workflows/feature-engineering?hl=es-419#:~:text=combina%20los%20resultados.-,Informaci%C3%B3n%20mutua%20ajustada%20(AMI),si%20se%20comparte%20m%C3%A1s%20informaci%C3%B3n.
- https://deepai.org/machine-learning-glossary-and-terms/kruskal-wallis-test
- https://medium.com/ingeniouslysimple/language-models-15e45dce0805

- https://github.com/christianversloot/machine-learning-articles/blob/main/how-to-use-logcosh-with-keras.md
- https://medium.com/@20__80__/mean-absolute-error-mae-machine-learning-ml-b9b4afc63077
- https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d
- https://pykeen.readthedocs.io/en/stable/api/pykeen.losses.MarginRankingLoss.html
- https://keepcoding.io/blog/medidas-de-calidad-en-matrices-de-confusion/
- https://keepcoding.io/blog/tipos-tests-estadisticos-para-big-data/
- https://statologos.com/prueba-de-mcnemar/
- https://www.britannica.com/science/mean-squared-error
- https://insideaiml.com/blog/MeanSquared-Logarithmic-Error-Loss-1035
- https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall
- https://neptune.ai/blog/gan-loss-functions
- https://en.wikipedia.org/wiki/Vector_space_model
- https://arxiv.org/abs/2010.08029
- https://optuna.org/
- https://www.statstutor.ac.uk/resources/uploaded/paired-t-test.pdf
- https://www.geeksforgeeks.org/sunburst-plot-using-plotly-in-python/
- https://towardsdatascience.com/learning-to-rank-a-complete-guide-to-ranking-using-machine-learning-4c9688d370d4
- Precision-Recall Curves. Sometimes a curve is worth a thousand… | by Doug Steen | Medium
- https://developers.google.com/machine-learning/glossary/fairness#predictive-parity
- https://www.ibm.com/docs/es/cognos-analytics/11.1.0?topic=terms-r2
- https://www.ibm.com/docs/es/cognos-analytics/11.2.0?topic=terms-adjusted-r-squared
- https://www.cienciadedatos.net/documentos/18_prueba_de_los_rangos_con_signo_de_wilcoxon
- https://economipedia.com/definiciones/analisis-de-regresion.html
- https://towardsdatascience.com/fairmodels-lets-fight-with-biased-machine-learning-models-f7d66a2287fc
- https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d
- https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
- https://www.researchgate.net/publication/360499809_SHARP-GAN_SHARPNESS_LOSS_REGULARIZED_GAN_FOR_HISTOPATHOLOGY_IMAGE_SYNTHESIS
- https://ieeexplore.ieee.org/document/9761534
  https://medium.com/beyondminds/advances-in-generative-adversarial-networks-7bad57028032
- https://keepcoding.io/blog/similitud-entre-vectores-o-cosine-similarity/
- https://towardsdatascience.com/triplet-loss-advanced-intro-49a07b7d8905
- https://cristhianhenry57.medium.com/clustering-6adbfae73ded
- https://towardsdatascience.com/the-f1-score-bec2bbc38aa6
- https://neptune.ai/blog/gan-loss-functions

- https://www.wonderflow.ai/blog/what-is-accuracy-in-text-analysis
- https://huggingface.co/spaces/evaluate-metric/wer

8.2 Estándares

Conceptos y términos

- [114] ISO/IEC 22989:2022, Information technology – Artificial intelligence – Artificial intelligence concepts and terminology, https://www.iso.org/standard/74296.html
- [115] ISO/IEC 23053:2022, Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML), https://www.iso.org/standard/74438.html

Precisión/ Accuracy. Algunos de estos estándares han sido ya publicados, mientras que otros se encuentran actualmente en desarrollo o revisión:

- [116] ISO/IEC TS 4213:2022, Information technology – Artificial intelligence – Assessment of machine learning classification performance, https://www.iso.org/standard/79799.html
- prEN 18229-2 AI trustworthiness framework – Part 2: Accuracy and robustness
- prEN Evaluation methods for accurate computer vision systems
- prEN ISO/IEC 23282 Evaluation methods for accurate natural language processing systems

Diseño y Desarrollo

- [117] ISO/IEC DIS 5338, Information technology – Artificial intelligence – AI system life cycle processes, https://www.iso.org/standard/81118.html
- [118] ISO/IEC DIS 5339, Information technology – Artificial intelligence – Guidance for AI applications, https://www.iso.org/standard/81120.html
- [119] ISO/IEC DIS 5392, Information technology – Artificial intelligence – Reference architecture of knowledge engineering, https://www.iso.org/standard/81228.html
- [120] ISO/IEC TR 24372:2021, Information technology – Artificial intelligence (AI) – Overview of computational approaches for AI systems, https://www.iso.org/standard/78508.html
- [121] ISO/IEC CD TS 12791, Information technology – Artificial intelligence – Treatment of unwanted bias in classification and regression machine learning tasks, https://www.iso.org/standard/84110.html

Financiado por
la Unión Europea
NextGenerationEU

GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

Plan de
Recuperación,
Transformación
y Resiliencia

España | digital 2026