CERMI

COMITÉ ESPAÑOL
DE REPRESENTANTES
DE PERSONAS
CON DISCAPACIDAD

Metodología CERMI para hacer una Auditoría Algorítmica

GOBIERNO DE ESPAÑA · MINISTERIO DE DERECHOS SOCIALES Y AGENDA 2030
Plan de Recuperación, Transformación y Resiliencia
Financiado por la Unión Europea NextGenerationEU
SECRETARÍA DE ESTADO DE DERECHOS SOCIALES

metaDIverso
CERMI DISCAPACIDAD

Índice

1 Introducción

En el escenario actual de desarrollo exponencial de aplicaciones de Inteligencia Artificial (IA), desde el CERMI se ha considerado crítico y necesario elaborar una metodología propia que pueda servir de orientación para la auditoría de muchas de estas aplicaciones algorítmicas, desde la perspectiva fundamental de los derechos humanos y de la protección de las personas con discapacidad y sus familias.

Efectivamente, en una era en la que las tecnologías basadas en algoritmos están profundamente integradas en la infraestructura social, económica y política de nuestras sociedades, es imperativo garantizar que estas tecnologías operen de manera transparente, justa y responsable. La creciente dependencia de los algoritmos, en especial aquellos con un impacto social significativo, requiere una supervisión y revisión sistemáticas para asegurar que sus operaciones y resultados sean comprensibles, equitativos y libres de prejuicios no intencionados.

El CERMI es por encima de todo una organización animada y dirigida por valores, sustentados por la Convención Internacional sobre los Derechos de las Personas con Discapacidad, y en su propósito figura el servir de orientación para otras organizaciones y velar por la defensa de los derechos e intereses de las personas con discapacidad y de sus familias. Motivada por dicha visión de convertirse en el referente de la discapacidad globalmente considerada en España, y motor de innovación social en materia de discapacidad, esta entidad ha abordado este proyecto con rigor, objetividad e independencia.

Este documento ofrece una metodología robusta para abordar estas preocupaciones, diseñada específicamente para algoritmos con un claro impacto social, independientemente de las tecnologías en las que se implanten. La metodología propuesta permite abordar y superar posibles deficiencias en los procedimientos actuales, fortaleciendo las medidas de responsabilidad y asegurando una adecuada rendición de cuentas por parte de todas aquellas personas involucradas en el ciclo de vida del algoritmo. Al llevar a cabo un proceso de auditoría, todas las partes involucradas no sólo estarán contribuyendo a la integridad y confiabilidad de sus soluciones basadas en algoritmos, sino que también estarán reafirmando su compromiso con una tecnología que es respetuosa, justa y beneficiosa para toda la sociedad.

La estructura de este documento es como sigue: en primer lugar, se contextualiza la necesidad de contar con unos principios guía para los sistemas de inteligencia artificial, se describe brevemente el marco normativo actual y se particularizan estos principios para el caso del CERMI. A continuación, se explica el concepto de auditoría algorítmica y se justifica su necesidad, enumerando los diferentes tipos de auditoría en función de su alcance. En la siguiente sección se describen las fases de la auditoría y se especifica el análisis cuantitativo y cualitativo. Por último, se incluyen unas breves conclusiones, un listado de referencias y algunos apéndices con recursos valiosos para entender el panorama regulatorio y proporcionar herramientas prácticas, como la lista de comprobación de criterios de auditabilidad y el modelo CRISP-DM, que pueden ser útiles en el proceso de auditoría.

2 Principios éticos

2.1 Necesidad y evolución de unos principios éticos para la IA

Aunque la inteligencia artificial (IA) surgió alrededor de los años 50, no ha sido hasta el año 2010 cuando se ha empezado a utilizar de manera masiva. Su uso ha venido favorecido por la existencia de más datos que nunca, el incremento de la capacidad de los procesadores y el surgimiento de la tecnología *Cloud* (servicios de computación en la nube). Además, existen cada vez más algoritmos *open source* (de código abierto) que permiten unos tiempos de desarrollo mucho más cortos para diseñar aplicaciones de IA. Esto ha supuesto un crecimiento exponencial en la aplicación de la IA en todos los ámbitos de la sociedad.

En la actualidad usamos IA cada día sin darnos cuenta, cuando nos desplazamos usando los navegadores, cuando interactuamos con los asistentes de voz que ya son un dispositivo más en nuestras casas, cuando nos recomiendan una película o incluso nos realizan un diagnóstico médico. Existen multitud de aplicaciones beneficiosas, como por ejemplo el proyecto Alphafold[1] de Deepmind, que permite a los médicos acortar los tiempos de diseño de proteínas, gracias al uso de la IA y que puede suponer una revolución en la cura de ciertas enfermedades. Pero al mismo tiempo también hay casos que suponen riesgos, como por ejemplo la invasión de la privacidad, la manipulación de la opinión de las personas[2], sesgos en los procesos de contratación[3], o *deepfakes*[4].

Las personas con discapacidad representan el mayor grupo minoritario del mundo. Se calcula que en el mundo hay más de mil millones de personas con discapacidad, la mayoría en países de renta baja y media. En todo el mundo, casi 240 millones de niñas y niños viven con algún tipo de discapacidad. Considerando las cifras es muy relevante tener en cuenta a este grupo social para entender cómo la

[1] https://alphafold.ebi.ac.uk
[2] https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html
[3] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
[4] https://this-person-does-not-exist.com/es

aplicación de la IA puede favorecer su día a día, y asimismo entender si es necesario revisar determinados patrones erróneos o decidir tomar alguna acción al respecto.

Los sistemas basados en IA puede servir de ayuda en muchos casos para las personas con discapacidad, por ejemplo mediante el desarrollo de aplicaciones, que permitan la descripción de imágenes de manera automática para personas con problemas de visión, o el reconocimiento de la voz, o aplicar la lengua de signos o el subtitulado a un video de manera automática para personas sordas[5], o la capacidad de generar textos para personas con dislexia o el diseño de robots para el cuidado de personas con necesidades de apoyo. También, puede ayudar a las personas con discapacidad a encontrar una ruta que proporcione una mejor accesibilidad[6]. Además, otro conjunto de herramientas apoyadas en IA, son aquellas para el seguimiento ocular y el reconocimiento de voz, que pueden permitir a las personas con discapacidad acceder a la información, así como comunicarse mejor.

Adicionalmente, los sistemas de IA pueden asistir a personas con cualquier tipo de problemas en la comunicación, habla o escritura, con aplicaciones como sistemas de recomendación, que según las acciones de la persona usuaria comparadas con las de otras pueden recomendar símbolos o posibilidades lingüísticas adecuados en función de la persona, personalización y mejora de predicciones basado en el comportamiento de las usuarias/os.

Es decir, las aplicaciones de la IA pueden ayudar a las personas en su día a día y en concreto a las personas con discapacidad, pero también hay que ser conscientes de los riesgos, asegurando que en el diseño e implantación de estos sistemas se tienen en cuenta a este grupo social para evitar circunstancias de desamparo, discriminación y vulneración de derechos.

Por ejemplo, un área clave en los últimos años de la implantación de la transformación digital son las redes sociales, donde se pueden encontrar algunos retos para la IA y la discapacidad. Las evaluaciones algorítmicas basadas en

[5] https://g3ict.org/publication/plug-and-pray-a-disability-perspective-on-ai-automated-decision-making-and-emerging-tech.
[6] https://smartcities4all.org/ai-for-inclusive-sidewalks/.

características personales extraídas de estas redes sociales[7] se usan habitualmente para conceder un préstamo o contratar a una persona. Este tipo de evaluaciones algorítmicas, para toda la tipología de personas usuarias, pueden estar siendo sesgadas y llegar a ser discriminatorias. "*La discriminación adopta muchas formas: racismo, sexismo o incluso por edad, pero un tipo de discriminación de la que se habla muy poco es la de personas con discapacidad*"[8]. Muchas veces este tipo de discriminación no es del todo consciente, sino inconsciente, porque no se ha pensado en los distintos tipos de personas y en sus capacidades. Igualmente ocurre en el "mundo físico" cuando no se dispone de los accesos adecuados para todo tipo de personas, o cuando se espera que las personas caminen largas distancias con frecuencia, y no se diseña espacios laborales y sociales a tal efecto. Con esto se está discriminando a quien no puede desplazarse de manera sencilla y se está dando un trato de favor a quienes sí pueden hacerlo. Las tecnologías digitales como la IA tienen la habilidad de poder crear entornos más inclusivos en el ámbito profesional y personal, pero existe asimismo el riesgo de desplegar estas tecnologías de manera que perpetúen los problemas existentes.

La IA puede ser de gran ayuda para el ser humano, pero como toda nueva tecnología tiene sus riesgos. En el caso de los sistemas de IA, es especialmente relevante, porque es una tecnología que puede tomar decisiones automáticas por las personas. Podríamos realizar una clasificación de los sistemas de IA en tres tipos, según la autonomía en la toma de decisiones: **sistemas de soporte a la decisión, sistemas semiautónomos**, y **sistemas completamente autónomos**. En los primeros, la información se utiliza para la toma de las decisiones por parte de una persona, es decir, el sistema de IA sirve de apoyo a la persona que toma la decisión. En el segundo tipo de sistemas, los semiautónomos, el sistema de IA puede tomar la decisión, pero la persona puede intervenir y aprobar o cambiar la decisión. En el tercer tipo, en los llamados sistemas completamente autónomos, el sistema de IA toma la decisión de manera autónoma, y la persona sólo interviene en el diseño del sistema. En la Figura

[7] (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3724556)

[8] https://eu.lansingstatejournal.com/story/money/careers/2020/02/24/employers-resources-adapting-workplace/111340878/

1 se representa esta clasificación de sistemas de IA basada en el nivel autonomía en la toma de decisiones:

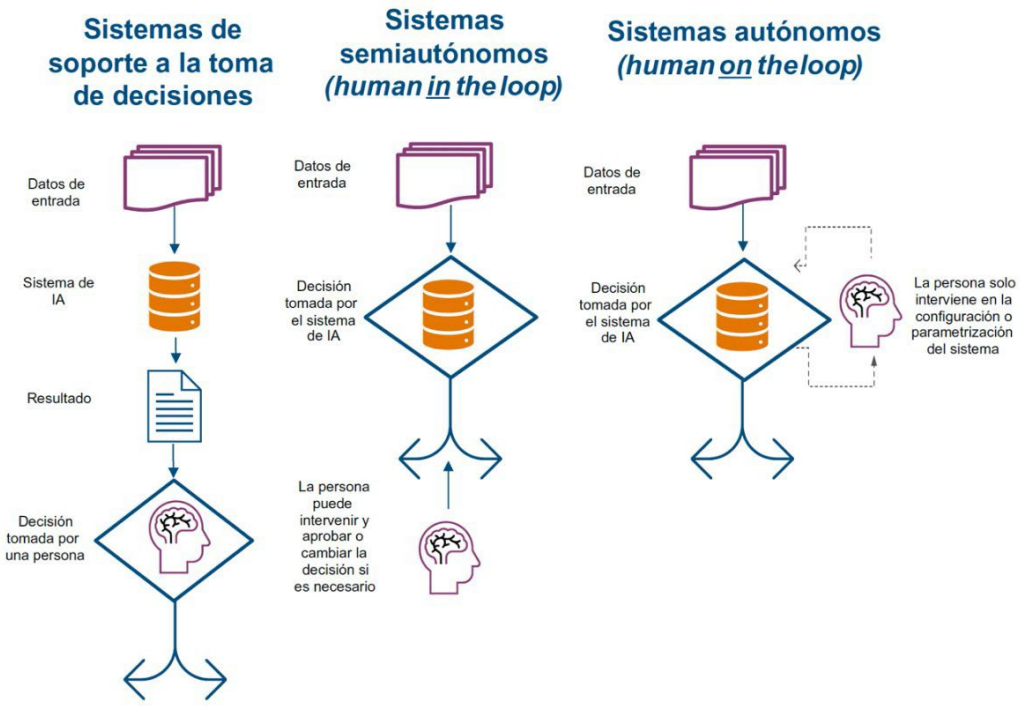


Figura 1. Sistema de toma de decisiones. Fuente: Manual de ética aplicada a la IA.

Descripción figura 1. La imagen representa mediante símbolos los tres tipos de sistemas de IA basados en el nivel de autonomía en la toma de decisiones. La columna de la izquierda representa los sistemas de soporte a la decisión, la columna central los semiautónomos y la columna de la derecha los autónomos.

Desde el momento en que se pueden desarrollar este tipo de sistemas, es clave entender qué principios éticos deberían regir el diseño de los sistemas de IA. Es por ello que desde el año 2016 han surgido multitud de documentos e iniciativas acerca de la IA ética, y más en concreto de cuáles tendrían que ser los principios éticos que deberían regir esta "nueva tecnología" que se está utilizando ahora de manera masiva. En la publicación de estos documentos han participado personas de todos los ámbitos de la sociedad y también organizaciones de todo tipo. A pesar de la multitud de publicaciones de organismos supranacionales, internacionales, entidades públicas y empresas tecnológicas, hoy en día no hay un acuerdo global sobre cuáles deberían ser los principios éticos fundamentales para aplicar a la IA, aunque se suelen encuadrar en cuatro grandes grupos: **autonomía, justicia, beneficencia y no maleficencia**. El

organismo independiente AIethicslab[9], formado por investigadores multidisciplinares, ha creado una herramienta que revisa los documentos publicados e intenta agrupar los principios mencionados en esos documentos en las cuatro categorías mencionadas, la clasificación más actualizada se incluye en la Figura 2:

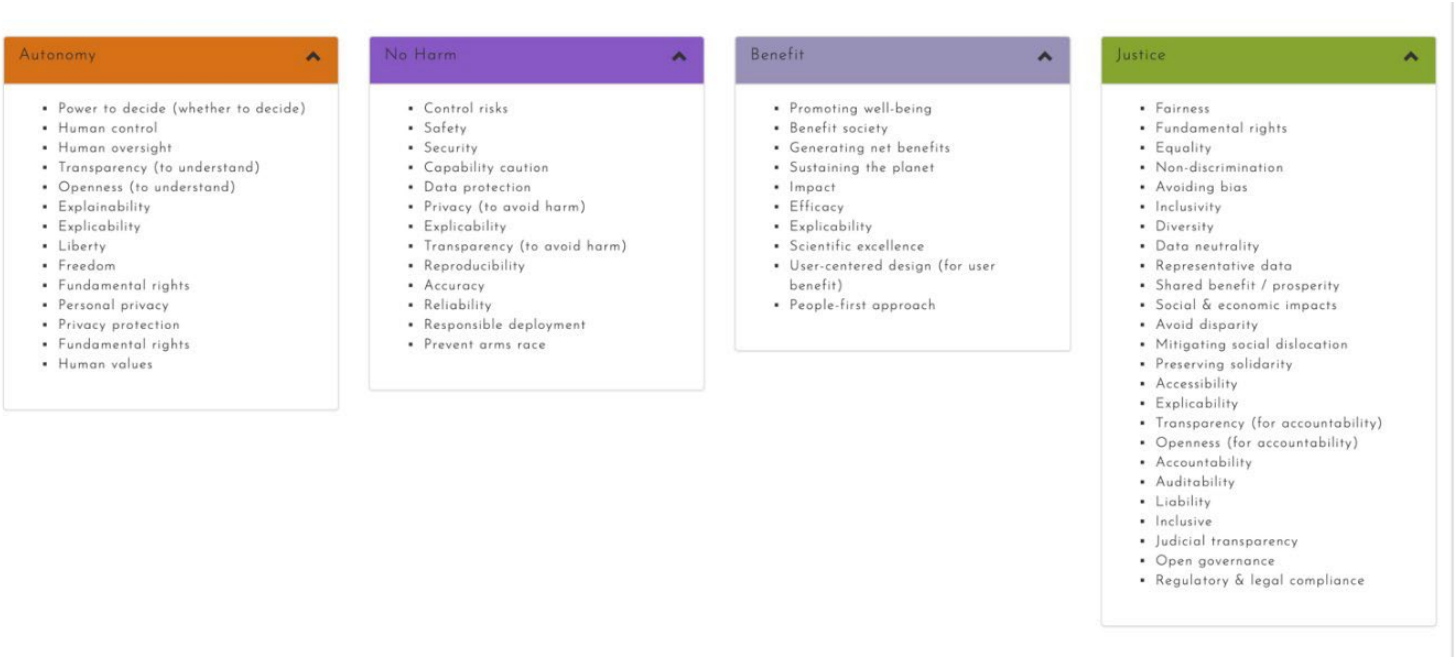


Figura 2. Clasificación de principios éticos de la IA. Fuente: AIethicslab.

Descripción figura 2. La imagen representa cuatro cuadros que recogen las cuatro categorías, por este orden de izquierda a derecha: autonomía, no maleficencia, beneficencia y justifica. Dentro de cada cuadro se desarrolla el contenido de cada categoría.

Si nos centramos en el análisis de los principios éticos que podrían ser más relevantes se describen a continuación en la Figura 3:

Figura 3. Descripción Principios Éticos de la IA.

Descripción figura 3. La imagen representa una tabla con dos columnas y ocho filas donde se recoge el nombre de cada principio (columna de la izquierda) y su descripción (columna de la derecha).

PRINCIPIOS	DESCRIPCION
Privacidad	La privacidad constituye un derecho esencial, un principio ético clave para la IA y además, desde 2018 con la entrada en vigor de la RGPD (Reglamento de Protección de Datos) es una normativa de obligado cumplimiento. En el caso de la IA es clave

[9] https://aiethicslab.com/big-picture/.

	saber la procedencia de los datos, así como no utilizar los atributos protegidos (raza, discapacidad...) para la toma de decisiones.
Equidad	La IA se entrena a menudo con datos de todo tipo que pueden representar ya ciertos sesgos. Hay que velar porque los sistemas de IA no refuercen los prejuicios y las desigualdades humanas. También se pueden encontrar sesgos en el diseño de los algoritmos que se deben revisar. En muchas ocasiones no es sencillo mantener la equidad dado que en algunos casos se puede querer "proteger" a algún grupo concreto frente a otro y esto puede llevar a no poder cumplir fielmente con esa equidad.
Inclusión y no discriminación	Los sistemas de IA deben poder beneficiar a todas las personas. Muy similar a la equidad, la inclusión requiere la consideración de todo tipo de personas independientemente de la raza, el sexo, la edad o la discapacidad, entre otros factores.
Transparencia	Es clave entender qué hay detrás de la decisión de un algoritmo, y cómo han influido las distintas variables en la toma de decisión final. No todos los sistemas de IA presentan esta característica y a efectos éticos es clave comprender cómo se están tomando estas decisiones. La falta de transparencia también podría mermar la posibilidad de impugnar eficazmente las decisiones basadas en resultados producidos por los sistemas de IA y, por lo tanto, podría vulnerar el derecho a un juicio imparcial y a un recurso efectivo, y limita los ámbitos en los que estos sistemas pueden utilizarse legalmente. Transparencia y explicabilidad se usan a veces de manera intercambiable pero no son lo mismo.
Rendición de cuentas (*Accountability* en inglés)	Los sistemas de IA pueden producir resultados impredecibles, dependiendo del tipo de decisiones y de su importancia, no es lo mismo la recomendación de una película que una decisión médica, es necesario que las personas sean conscientes que la responsabilidad final de la decisión no es del algoritmo, sino que detrás siempre hay una persona.
Explicabilidad	Es clave entender qué hay detrás de la decisión de un algoritmo, y cómo han influido las distintas variables en la toma de decisión final. No todos los sistemas de IA presentan esta característica y a efectos éticos es clave entender cómo se están tomando estas decisiones. Aunque las técnicas de explicabilidad están avanzando, aún es complejo para ciertos tipos de algoritmos entender el detalle de cómo se ha tomado la decisión.
Responsabilidad	Muy relacionada con rendición de cuentas, los sistemas de IA pueden tomar decisiones de manera autónoma y es importante mantener a la persona en el ciclo de esta decisión.

Supervisión humana	Es importante distinguir entre los sistemas semiautónomos (*Human in the Loop*) donde las personas intervienen en determinadas decisiones y los sistemas autónomos (*Human on the loop*) donde la persona sólo ha intervenido en el diseño del algoritmo, pero las decisiones son tomadas de manera autónoma por éstos.
Seguridad y robustez	El principio de seguridad y robustez aplicado a la IA debe ser entendido por un lado en lo que respecta a las soluciones de IA para no comprometer el entorno y por otro lado la capacidad de resistir amenazas externas
Sostenibilidad	La irrupción de la IA en la sociedad puede contribuir a mejorar la sostenibilidad y proyectos como los de *IA for good* demuestran que la IA puede ayudar por ejemplo al cambio climático o a la mejora de la limpieza en los océanos.

No es sencillo decidir cuáles son los principios adecuados para elegir, dado que todos tienen su importancia. A la hora de elegir estos principios, y priorizarlos adecuadamente, esto va a estar muy relacionado con la industria en la que se van a aplicar estos principios. En el caso que nos ocupa de la discapacidad, los principios van a estar muy relacionados con la equidad o la explicabilidad, pero además va a ser importante incluir en un análisis ético, áreas como la accesibilidad de las aplicaciones, que es un tema crítico para este grupo social.

Hasta ahora se han ido enunciando los principios éticos clave para una IA confiable y responsable. Estos principios éticos son los que deberían inspirar cualquier regulación alrededor de la IA. La regulación supondrá una serie de requisitos a cumplir, y probablemente una serie de penalizaciones si estos no se cumplen, pero la ética está encuadrada, como se detallaba, en qué se debe hacer.

2.2 Aspectos regulatorios[10]

Como se ha detallado en las secciones anteriores, ocurren situaciones en las que el uso de la IA puede ser un perjuicio más que un beneficio. Entre los avances realizados hasta el momento se puede resaltar la Estrategia de la Comisión Europea

[10] En el Anexo I se recoge una evolución y explicación más detallada de la regulación europea sobre inteligencia artificial.

sobre los Derechos de las Personas con Discapacidad 2021-2030 o la propuesta por Unicef sobre las soluciones digitales accesibles e inclusivas para niñas con discapacidad. Estas dos iniciativas clave han puesto de manifiesto la necesidad de proponer mejoras en la próxima regulación sobre IA en Europa (Ley Europea de IA), que están siendo llevadas a cabo por organizaciones como el EDF (Foro Europeo de Discapacidad).

En 2021 la Comisión Europea adoptó la Estrategia sobre los Derechos de las Personas con Discapacidad 2021-2030[11]. Con esta estrategia la Comisión quiere mejorar la vida de este grupo ciudadano. En la actualidad las personas con discapacidad todavía se enfrentan a barreras considerables y a tener un riesgo más alto de exclusión social. El objetivo de esta estrategia es progresar en los derechos de las personas con discapacidad independientemente del sexo, la raza, la religión, la edad o su orientación sexual. Las iniciativas emblemáticas de la estrategia son:

- **Accesibilidad:** Proporcionar información y buenas prácticas en esta área antes de terminar 2022.
- **Tarjeta Europea de Discapacidad:** La Comisión Europea impulsará la regulación de una tarjeta de discapacidad implantada en toda la UE. Esto permitirá una movilidad mejor a las personas con discapacidad entre los países de la Unión Europea.
- **Recomendaciones sobre vida independiente e inclusión en la comunidad:** Esto contribuirá a que las personas con discapacidad puedan vivir en viviendas accesibles y con apoyo de la comunidad en 2023.
- **Marco común de servicios sociales** de excelencia para las personas con discapacidad en 2024.
- **Un conjunto de medidas** respecto a la incorporación de las personas con discapacidad al mercado laboral a finales de 2022

[11] https://ec.europa.eu/social/main.jsp?catId=1484&langId=en.

- **Creación de una plataforma de discapacidad** que reúne a las autoridades nacionales responsables de la aplicación de la CDPD, las Organizaciones de Personas con Discapacidad y la Comisión Europea.

- **Estrategia de RRHH** renovada para la Comisión Europea que incluya acciones para promocionar la diversidad y la inclusión de las personas con discapacidad.

La **Comisión Europea** proporcionará soporte a los países para establecer sus planes y estrategias nacionales en lo que respecta a estos derechos.

En 2022 Unicef publicó una serie de recomendaciones basadas en un estudio que incluía una revisión de la literatura al respecto, sobre accesibilidad en las herramientas digitales. Este estudio explica el concepto de tecnología asistida, accesibilidad e inclusión digital. Además, describe la necesidad de tener en cuenta la inclusividad cuando se aborda el desarrollo de estas herramientas digitales (muchas de ellas basadas en IA) y proporciona recomendaciones al respecto. El estudio se centra en las niñas con discapacidad, pero proporciona una base excelente para entender el impacto.

En 2021 se publica el primer borrador de la Ley Europea de IA, y como se detallaba en el apartado de principios el enfoque utilizado es un enfoque basado en riesgos. Las categorías que definen son las siguientes: **riesgos inaceptables, altos, limitados y mínimos.** En particular, los **sistemas de riesgo inaceptable** deben prohibirse totalmente o con pocas excepciones. Esto incluye los sistemas gubernamentales de puntuación social y de identificación biométrica en tiempo real en espacios públicos. Referente a los **sistemas de alto riesgo** requieren de una evaluación de conformidad y una auditoría de la seguridad, privacidad, solidez e impacto del sistema. Esto incluye también infraestructuras críticas y de transporte, plataformas educativas y de contratación, servicios privados y públicos, y sistemas de aplicación de la ley que puedan interferir con los derechos fundamentales de las personas. Los sistemas de **riesgo limitado** requieren una confirmación de la transparencia del sistema y el consentimiento del usuario. Las personas usuarias deben ser conscientes de que están interactuando con una máquina para tomar la decisión informada de continuar o retroceder. Algunos ejemplos son los chatbots y la IA conversacional, los sistemas de reconocimiento de emociones y los filtros faciales. Por último, los **sistemas de riesgo mínimo** están menos regulados y requieren el

Código de conducta. Se refiere principalmente a sistemas que no implican la recopilación de datos sensibles o que no interfieren con los derechos humanos. Esto incluye videojuegos con IA o filtros de spam.

La Ley Europea de IA aún no es definitiva y desde el punto de vista de la discapacidad se han hecho distintas propuestas desde la OCDE (Organización Internacional para la cooperación y desarrollo económico) y el EDF (Foro Europeo de la Discapacidad).

Un reciente informe de la OCDE[12]propone modificaciones que se pueden agrupar en dos grandes grupos: pasos para facilitar que los sistemas de IA consideren la discapacidad y pasos para regular el mal uso de estos sistemas respecto a la discapacidad.

Respecto al informe de la OCDE en lo que se refiere a los pasos para facilitar que los sistemas de IA consideren la discapacidad se recogen 4 puntos fundamentales:

- **Partes interesadas y familia:** La tecnología diseñada para personas con discapacidad no involucra solamente a las propias personas sino a un "ecosistema" como familias, cuidadores/as, asesores/as o educadores/as. Esto es importante tenerlo en cuenta en el desarrollo de aplicaciones para incluir un interfaz para la persona usuaria y otra para las personas de este ecosistema, madres, padres o comunidad educativa.

- **Espectros, género y grupos de edad:** Es importante tener en cuenta que la discapacidad no es sencilla de representar, no son monolitos, sino que normalmente se trabaja con espectros que presentan múltiples parámetros. Un ejemplo podría ser que las personas con dificultad de aprendizaje pueden tener problemas de salud mental. Representar esta situación en un sistema de IA no es sencillo.

- **Vocabulario y marcos de conocimiento accesibles:** La OMS (Organización Mundial de la Salud) ya ha resaltado la necesidad de establecer un marco de

[12] https://oecd.ai/en/wonk/eu-ai-act-disabilities.

competencias de salud digital, de la misma manera es importante tenerlo en cuenta para la accesibilidad a la tecnología. Esto puede afectar a herramientas desarrolladas con IA para el aprendizaje, que deben incluir las diferentes casuísticas.

- **Necesidad de evaluar el impacto:** A pesar de las categorías de riesgo descritas, no existen evaluaciones de impacto que aborden los riesgos específicos para las personas con discapacidad, un ejemplo podría ser el reconocimiento de emociones, que requeriría un estudio distinto en el caso de la discapacidad.

Referente a los pasos para regular el mal uso de estos sistemas respecto a la discapacidad, se pueden agrupar en cinco puntos:

- **Sistemas de riesgo alto e inaceptable:** Es importante que en esta categoría se revisen e incluyan los riesgos adicionales que afectan a las personas con discapacidad. Entre ellos, como ejemplo, se podrían incluir los sistemas policiales o de seguridad, que podrían reconocer erróneamente dispositivos de ayuda a las personas con discapacidad como objetos peligrosos. Adicionalmente, también se podría incluir aquí las discriminaciones adicionales en lo que a la selección de trabajadores/as se refiere con herramientas de IA, como ya se mencionaba previamente.

- **Sistemas de riesgo limitado y el reconocimiento de emociones:** Los sistemas de reconocimiento de emociones son bastante sensibles a los sesgos para cualquier tipo de población, no sólo las personas con discapacidad. Por eso es importante tener en cuenta en este tipo de sistemas los derechos de las personas con discapacidad, y cómo alinearlos con el desarrollo de estas herramientas basadas en IA.

- **Silos e implicación humana:** Más del 40% de las personas adultas con una discapacidad afirman sentirse socialmente aisladas. Es importante resaltar que el uso de la tecnología como la IA puede facilitar muchas tareas, pero no sustituir la interacción social.

- **Escenarios de abuso y maltrato:** Las personas con discapacidad se convierten casi en 2,2 veces más en víctimas de violencia[13], manipulación o ataques

[13] https://disability.royalcommission.gov.au/news-and-media/media-releases/people-disability-face-much-greater-risk-violence-people-without-disability.

sociales. En muchas ocasiones los algoritmos existentes en redes sociales[14] discriminan a estas personas o las identifican como no humanas.

- **Omisiones y responsabilidad:** No solamente son las acciones las que pueden resultar discriminatorias o perjudiciales para este grupo social, sino también las omisiones. Por ejemplo, cuando un sistema de IA excluye a una determinada población o se basa en fuentes no veraces.

- **Creación y propiedad de datos:** Las herramientas de IA pueden ayudar con el aprendizaje adaptativo y la IA conversacional, pero es importante identificar las partes implicadas en el desarrollo de este proceso. Es clave entender cómo se están generando y almacenando los datos y quién es el propietario de ellos. En muchos casos se llevan a cabo auditorías para poder entender el detalle.

Si revisamos las recomendaciones realizadas por EDF a la Comisión Europea a finales de 2021, estas pueden agruparse en los siguientes puntos, que incluyen algunos de los ya mencionados por la OCDE. Los puntos clave son los siguientes:

- **Accesibilidad:** Resalta la necesidad de incluir la accesibilidad de manera horizontal independientemente del nivel de riesgo que se le atribuya a la aplicación de IA. Recuerda que los sistemas deben ser consistentes con la regulación europea sobre accesibilidad[15].

- **No discriminación e igualdad:** La Ley Europea de IA debe prohibir algunos usos listados en su Anexo III, en lo que se refiere a las prácticas de identificación biométrica. En concreto se refiere a que deben ser prohibidas la identificación y categorización de personas por sistemas de IA, que determinen el acceso a educación, empleo, servicios y beneficios privados o públicos o para la concesión de asilo o control de fronteras.

- **Privacidad y protección de datos:** Esta privacidad y protección de datos debe estar asegurada para las personas con discapacidad. Además, se deben establecer medidas para que se informe a las personas con discapacidad y éstas puedan oponerse a la recopilación de estos datos.

- **Mecanismos sólidos de quejas:** Ley Europea de IA debe garantizar la manera de presentar quejas y/o denuncias en caso de abuso. Estas

[14] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3724556.
[15] https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32019L0882

medidas deben ser accesibles para las personas con discapacidad. El Reglamento también debe garantizar evaluaciones previas del impacto sobre los derechos humanos de los sistemas de IA de alto riesgo antes de ponerlos en funcionamiento, incluida la evaluación de la accesibilidad de estos sistemas para las personas con discapacidad.

- **Involucración de organizaciones de personas con discapacidad y representación en los conjuntos de datos:** Los miembros de la UE están obligados por el artículo 4.3 de la CDPD (Convención sobre los derechos de las personas con discapacidad) [16]y deben consultar e involucrar a las personas con discapacidad, y a sus correspondientes organizaciones representativas en los estados miembros, en el desarrollo, implantación y monitorización del reglamento europeo (Ley Europea de IA).

Como conclusión a las recomendaciones de la OCDE y a las peticiones formales de modificación del reglamento Ley Europea de IA, queda patente que aún son necesarias acciones respecto a la normativa para que todas las reivindicaciones de las personas con discapacidad se contemplen en el texto final. A la fecha de redacción de este documento aún está pendiente el texto definitivo de la Ley Europea de IA.

2.3 Principios éticos para el CERMI

La utilización de la Inteligencia Artificial tiene el potencial de transformar muchos aspectos de la sociedad, y es fundamental considerar desde el inicio cómo esa transformación puede afectar a las personas con discapacidad. Para asegurar que estos cambios beneficien a todas las personas y no perpetúen o exacerben las desigualdades existentes, se proponen seis principios éticos adaptados del Código Ético del CERMI y combinados con los principios éticos mencionados anteriormente de IA.

[16] https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities/article-4-general-obligations.html

A continuación, se incluyen los principios relevantes que el CERMI ha decidido priorizar.

1. Inclusión y no discriminación.

Los sistemas de IA deben respetar, proteger y promover los derechos humanos de las personas con discapacidad atendiendo al mandato contenido en los tratados internacionales de derechos humanos, especialmente en la CDPD. No puede haber sistemas de IA que sirvan para la vulneración de los derechos de las personas con discapacidad.

Los sistemas de IA deben servir para luchar contra todo tipo de discriminación y garantizar la diversidad y la inclusión de todas las personas con discapacidad, con especial atención a las personas sujetas a mayor riesgo de exclusión por la intersección de las características personales de sexo, edad y discapacidad con otros factores de exclusión social (como la raza o la etnia, la religión y la orientación e identidad de género, la ruralidad, entre otros).

La brecha digital en sus diferentes dimensiones es la principal barrera a la inclusión de este grupo ciudadano en los sistemas de IA.

Es fundamental garantizar que los sistemas de IA no aumenten los sesgos discriminatorios contra las personas con discapacidad o contribuyan a crear nuevos sesgos, conscientes o inconscientes.

2. Transparencia y explicabilidad.

Todas las etapas y componentes de los sistemas de IA deben ser transparentes y fáciles de comprender y permitir en todo momento preguntas sobre su funcionamiento. Esto significa poder acceder a la información sobre cómo se implanta cada una de las fases de un sistema de IA, sobre los factores que influyen en la predicción u otras decisiones creadas a partir de los sistemas de IA y sobre los conjuntos de datos utilizados para la creación, entrenamiento y validación de dichos sistemas.

Las personas con discapacidad deben recibir información clara y en formatos universalmente accesibles sobre cada uno de los componentes de los sistemas de IA y sobre cómo estos datos y componentes influyen en los resultados finales de dicho sistema.

Las personas con discapacidad deben saber en todo momento que interactúan con sistemas de IA.

Es imprescindible garantizar la protección y privacidad de los datos.

3. Accesibilidad universal.

Los sistemas de IA deben combatir la brecha digital, garantizando la accesibilidad universal en todos los procesos de interactuación, comunicación y acceso a la información con las personas usuarias.

4. Unidad y paricipación.

El abordaje de la IA por parte del CERMI se realiza desde el principio de unidad y cohesión de todo el movimiento CERMI, situando a las personas con discapacidad y sus familias como eje central de cualquier actuación en materia de IA.

Los sistemas de IA deben estar al servicio de las personas con discapacidad y de su inclusión. Las personas con discapacidad deben participar activamente en los equipos encargados de desarrollar sistemas de IA y aparecer representadas en los conjuntos de datos sobre los que se crean y entrenan estos sistemas.

Es fundamental garantizar la participación activa de las personas con discapacidad en todas las etapas del ciclo de vida de los sistemas de IA, especialmente como validadoras de estos sistemas.

Es fundamental que se garantice la supervisión humana de los sistemas de IA. En esa labor de supervisión deben diseñarse mecanismos de participación de la sociedad civil de la discapacidad.

5. Igualdad entre mujeres y hombres.

La brecha digital en sus diferentes dimensiones afecta de forma especialmente significativa a las mujeres y las niñas con discapacidad, al reconocimiento de sus derechos y a su imagen social. Esta brecha digital contribuye a perpetuar los sesgos que ya existen en el mundo analógico y aumentarlos exponencialmente.

Los sistemas de IA deben respetar, proteger y promover los derechos humanos de las mujeres con discapacidad atendiendo al mandato contenido en los tratados internacionales de derechos humanos, especialmente la CDPD y la CEDAW. Es preciso incluir la perspectiva transversal del enfoque de género en la evaluación del impacto ético de los sistemas de IA.

Se debe impulsar la participación activa de las mujeres con discapacidad en todas las etapas del ciclo de vida de los sistemas de IA.

Es imprescindible corregir la invisibilidad de las mujeres con discapacidad en los conjuntos de datos utilizados como base, entrenamiento y validación de los sistemas de IA, garantizando entre otras medidas la desagregación por sexo de los datos.

6. Sostenibilidad.

Los sistemas de IA deben contribuir a los objetivos de sostenibilidad desde un enfoque inclusivo.

Estos principios deben ser considerados como guías fundamentales en la implantación y uso de sistemas de IA dentro del CERMI y sus entidades. La adhesión a estos principios asegurará que la IA se utilice de manera que beneficie a las personas con discapacidad, en lugar de marginarlas o discriminarlas.

3 Objetivos y justificación de la auditoria algorítmica

3.1 La necesidad de revisar el uso de la IA

Sin duda la IA puede ayudar a mejorar la autonomía de las personas con discapacidad, pero también es verdad que los algoritmos, si no están entrenados convenientemente con datos que incluyan una muestra representativa de la población, y en este caso teniendo una muestra suficiente de personas con una determinada discapacidad, pueden incurrir también en una discriminación. Puede resultar difícil identificar los sesgos, dado que los tipos de discapacidad son muy variados. Además, al realizar esta toma de decisiones de manera automática puede ocurrir que se contribuya a aumentar esas discriminaciones de manera más rápida. Por ejemplo, las administraciones públicas recurren cada vez más a algoritmos que deciden quién debe recibir una ayuda médica o una prestación social por incapacidad[17]. Por ejemplo, en Estados Unidos, el SSDI (Seguro de discapacidad de la seguridad social) empezó a utilizar las redes sociales para investigar y clasificar a los ciudadanos solicitantes, dado que se detectaba un nivel de fraude elevado, como a veces ocurre con otros sistemas de ayuda al ciudadano. Aunque esta vigilancia en redes no parece la más adecuada, considerando la disponibilidad de los datos en las redes sociales, se utilizaron para la toma de decisiones mediante algoritmos, aunque los datos podían ser no veraces. Este tipo de acciones ha supuesto, en algunos casos, reducciones importantes en las ayudas recibidas por las personas con discapacidad simplemente por una toma de decisiones no correctamente informada. Existen casos también en el sector privado, donde los algoritmos de selección de personal pueden incorporar una discriminación por discapacidad. Esto es así porque los algoritmos están enfocados en buscar determinadas características y a partir de éstas encuentran correlaciones y patrones. Estos algoritmos, a la hora de la toma de decisión, suelen ser revisados en lo que se refiere a la raza y el género, pero el

[17] https://cdt.org/insights/report-challenging-the-use-of-algorithm-driven-decision-making-in-benefits-determinations-affecting-people-with-disabilities/

impacto en las personas con discapacidad no se suele considerar. El factor discapacidad no es sencillo de tener en cuenta por la gran diversidad de esta realidad, y si además la discapacidad coincide con otro atributo usado como raza o género que ya está discriminado (por ejemplo, mujeres con discapacidad), es aún más difícil tener en cuenta la discapacidad. Normalmente, los conjuntos de datos con los que se entrenan los algoritmos de empleo se basan en candidatos tradicionales sin discapacidad. Además, relacionado con la selección de personal están también las pruebas de personalidad y distintas pruebas de gamificación (orientadas a entender la personalidad mediante un juego). El algoritmo busca muchas veces características como la estabilidad emocional, la extroversión, la impulsividad, midiendo los datos del juego y también expresiones faciales. Las herramientas de IA suelen ser incapaces de leer estas emociones en las personas con discapacidad. En el sector seguros, la IA puede tomar decisiones de incrementar la prima sin tener en cuenta las características específicas de personas con discapacidad, y resulta difícil saber si esto está ocurriendo, por la falta de transparencia en la toma de decisiones, de ahí la publicación de recomendaciones al respecto[18] en Europa que tratan de regular y proteger a la persona consumidora final.

Como se puede observar en estos ejemplos, las personas con discapacidad pueden estar sometidas a desventajas y discriminaciones en los sistemas de IA, dado que normalmente sus cualidades se manifiestan físicamente de una forma que un algoritmo no suele haber visto en un conjunto de datos de entrenamiento previo.

Como conclusión se trataría de buscar una solución para poder combatir y minimizar los riesgos mencionados del uso de la IA para las personas con discapacidad. Según el estudio ya mencionado[19], se proponen acciones en tres áreas: innovar de manera inclusiva, aplicar la normativa y aumentar la transparencia y explicabilidad de las herramientas basadas en IA.

[18]https://www.beuc.eu/sites/default/files/publications/beuc-x-2021088_regulating_ai_to_protect_the_consumer.pdf
[19] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3724556

Con respecto a la **manera de innovar inclusiva**, se ha conseguido avanzar mucho en el desarrollo de aplicaciones que incluyan la accesibilidad gracias a considerar a las personas con discapacidad desde el inicio del diseño, aunque aún queda camino por recorrer. En primer lugar, las empresas, instituciones y organismos públicos reconocen no recoger suficientes datos de las personas con discapacidad. Esto va a impactar en el diseño adecuado de los algoritmos, que puede incluir sesgos en lo que respecta a la discapacidad. En segundo lugar, la calidad de los datos es clave siempre, y de cara a este grupo social hay que tener especial cuidado dado que los sistemas de IA dependen de datos históricos[20] que podrían no tener representado a este sector. En tercer lugar, es clave la brecha tecnológica en general respecto a la IA, y no sólo para este colectivo, que puede producir situaciones injustas. Por último, al igual que se han empezado a incluir desarrolladores/as con discapacidad o personas usuarias con discapacidad, para servir de usuarios/as de prueba en el desarrollo de aplicaciones, también es clave incluir en las fases iniciales de diseño de los sistemas de IA a personas con discapacidad. En definitiva, se trata de intentar tener una representación adecuada de la diversidad, al igual que se trata de hacer con la raza, el género o la procedencia geográfica.

En lo que se refiere a **trabajar en la explicabilidad** de los sistemas de IA, es una de las claves para el desarrollo de una IA más ética, dado que ayuda a entender lo que está ocurriendo. Cuando estos sistemas toman decisiones importantes para la persona, es necesario que se incluya un detalle del porqué se han tomado esas decisiones, cuáles son las variables que se han utilizado para la toma de decisiones y cuáles de ellas tienen más peso a la hora de influir en la decisión final. En definitiva, se trata de continuar trabajando en técnicas que permitan entender mejor esa toma de decisiones algorítmicas. En multitud de situaciones, esto no va a ser sencillo, dado que los sistemas más complejos no son transparentes *per se*, sino que hay que diseñarlos de manera específica para que la incluyan. Ya sea seleccionando algoritmos más sencillos o bien usando técnicas de explicabilidad adecuadas.

Por último, pero no menos importante, en lo que se refiere a aplicar la **normativa** sin duda es necesario regular el uso de los sistemas de IA. En Estados

[20] <https://doi.org/10.5569/2340-5104.11.01.03>

Unidos, por ejemplo, el Estado de Nueva York publicó una directiva que prohíbe a las agencias de información del consumidor y a los prestamistas utilizar cierta información para determinar la solvencia de una persona[21], esto no es del todo nuevo, ya existen también limitaciones en cuanto a la información médica. En Europa sentó precedente respecto a la privacidad la RGPD, que entró en vigor en 2018. Además, como se evidenciaba anteriormente, ya se está generando una regulación sobre IA al respecto, que se prevé se publique en 2024.

3.2 Necesidad de la auditoría de los algoritmos

A pesar de los avances de las iniciativas regulatorias descritas en los apartados anteriores, así como de la consideración de un conjunto de principios a la hora de desarrollar y usar aplicaciones de IA, e incluso las recomendaciones para considerar los derechos humanos como base para estos sistemas, en muchas ocasiones puede resultar necesaria la auditoría de estos sistemas. Con el rápido avance y adopción de la inteligencia artificial (IA) en diversos ámbitos de nuestra vida cotidiana, es imperativo considerar el impacto social, ético y legal de los algoritmos.

El concepto de "auditoría" proviene de la práctica financiera, que además es el tipo de auditoría más desarrollado y practicado, y supone típicamente el proceso de examinar los registros financieros de una empresa para asegurar que son precisos y cumplen con las leyes y regulaciones pertinentes[22]. Esta forma de auditoría busca verificar la exactitud e integridad de la información financiera proporcionada por la organización, ofreciendo confianza y fiabilidad en tales declaraciones.

Por otro lado, las auditorías éticas en las organizaciones son otra actividad común de auditoría y se suelen referir a un enfoque sistemático que realiza una descripción, análisis y evaluación de los aspectos relevantes de la ética de una organización[23]. También han sido definidas como comprobaciones regulares, completas y documentadas del cumplimiento con las políticas y procedimientos

[21] https://www.nysenate.gov/legislation/bills/2019/S2302
[22] Porter, B., Simon, J., & Hatherly, D. (2014) Principles of External Auditing (4th ed.). Wiley.
[23] Kaptein, M. (1998) Ethics Management: Auditing and Developing the Ethical Content of Organizations. Springer

éticos elaborados por una organización. Estas auditorías buscan evaluar la adhesión de una empresa a sus propios estándares éticos y normativas, y pueden considerar la conducta y cultura organizacional, las políticas internas y las relaciones con los grupos de interés[24].

Existen otro tipo de evaluaciones, como las evaluaciones de riesgo, también referidas como análisis de impacto, que pueden incluir técnicas tanto cuantitativas como cualitativas, y que se utilizan para evaluar el riesgo de un sistema e identificar qué aspectos necesitan atención, desarrollar estrategias de mitigación de riesgos, así como la adopción de medidas preventivas[25].

De este modo, las auditorías éticas en sistemas de Inteligencia Artificial (IA) pueden ayudar a determinar si un sistema de IA se adhiere a un conjunto preestablecido de expectativas, ya sean regulaciones, estándares, principios o métricas específicas del sector o de la organización. Este tipo de auditorías se enfoca en asegurar que los sistemas de IA operan de una manera que es ética y alineada con las expectativas tanto legales como sociales y organizacionales, protegiendo así los intereses y derechos de todas las partes involucradas[26]. Este tipo de auditoría, comúnmente conocida también como auditoría algorítmica, surge como una disciplina esencial para evaluar, validar y asegurar que los sistemas de IA operen de manera justa, transparente y sin perjudicar a grupos vulnerables. En particular, las personas con discapacidad, que a menudo enfrentan barreras adicionales en su interacción con la tecnología, van a requerir garantías especiales para asegurarse de que los algoritmos no perpetúen o aumenten posibles problemas de discriminación, no inclusión o dificultades de accesibilidad.

[24] Rosthorn, J. (2000) Business Ethics Auditing – More Than a Stakeholder's Toy. Journal of Business Ethics, 27, 9-19.

[25] Raab, C.D. (2020) Information privacy, impact assessment, and the place of ethics. Computer Law and Security Review. 37, 105404.

[26] Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021) Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. Science and Engineering Ethics, 27(44), 1-30.

Alineados con los principios descritos anteriormente, podemos enumerar cuatro motivos fundamentales que justifican la necesidad de una auditoria algorítmica:

- **Transparencia y Confianza**: La IA, por su naturaleza, puede operar como una "caja negra", en donde sus decisiones no siempre son transparentes o comprensibles para los humanos. Las auditorías algorítmicas ayudan a desentrañar estos mecanismos y a construir confianza entre las personas usuarias y las tecnologías.

- **Diseño Inclusivo**: la auditoría algorítmica puede revelar áreas en las que la IA no está adecuadamente adaptada para las personas con discapacidad, y ayudar a un diseño y desarrollo más inclusivo en iteraciones futuras.

- **Prevención de Discriminación**: los algoritmos mal diseñados o entrenados con datos sesgados pueden discriminar, ya sea intencionadamente o por omisión, a grupos sociales en situación de vulnerabilidad, como las personas con discapacidad. Las auditorías permiten detectar y corregir estos sesgos.

- **Cumplimiento normativo:** A medida que los países y las organizaciones actualizan sus regulaciones y normativas, se hace necesario garantizar y verificar el cumplimiento y adecuación de los sistemas de IA para todas las personas, incluyendo las personas con discapacidad.

En conclusión, la auditoría algorítmica para IA no solo es una herramienta de supervisión técnica, sino también un imperativo ético y social. Además, en el caso concreto de las personas con discapacidad se prioriza una de las poblaciones en mayor riesgo a ser pasadas por alto o mal atendidas por la tecnología emergente. Garantizar que los algoritmos sean inclusivos y justos es fundamental para construir un futuro digital más equitativo.

A modo de resumen, se presentan en la Figura 4 un compendio de los objetivos que pueden cubrirse mediante una adecuada auditoria de sistemas de IA:

Figura 4. Objetivos de una Auditoría de sistemas de IA.

Descripción figura 4. La imagen representa una tabla con dos columnas y trece filas donde se recoge el nombre de cada objetivo (columna de la izquierda) y su descripción (columna de la derecha).

Objetivo	Descripción
Garantía de Calidad	Asegurar que los algoritmos, especialmente aquellos con alto impacto social, cumplan con los estándares de calidad definidos, independientemente de si son desarrollados por instituciones públicas o privadas, investigadores o emprendedores.
Identificación y Corrección de Deficiencias	Detectar y remediar las carencias en los procesos y las medidas de responsabilidad y rendición de cuentas de las acciones algorítmicas.
Promoción de la Reflexión Crítica	Incentivar un análisis introspectivo sobre el posible impacto social y ético de los algoritmos, fomentando un diseño y desarrollo consciente.
Transparencia	Establecer mecanismos que ilustren claramente los pasos y decisiones tomadas en el diseño y desarrollo de los sistemas algorítmicos.
Identificación de Riesgos	Detectar amenazas, errores y riesgos, ya sean actuales o potenciales, en todas las fases del ciclo de vida del algoritmo.
Estrategia de Mejora Continua	Formular planes de acción para perfeccionar procesos y sistemas basados en algoritmos en el futuro, y remediar problemas una vez implantados.
Promoción de la Proactividad	Resaltar la necesidad de aplicar auditorías antes de la implantación y despliegue de sistemas, garantizando una acción preventiva.
Adaptabilidad	Reconocer que los objetivos específicos de la auditoría pueden variar según el auditor, ya sea para generar conocimiento, investigar posibles impactos en ciertos grupos, recomendar mejoras o actuar como autoevaluación.
Evaluación Integral	Adoptar un enfoque holístico, considerando tanto análisis técnicos como cualitativos, evaluando no solo la eficacia del sistema, sino también su integración en el contexto social y las dinámicas que introduce.
Promoción del Conocimiento	Adquirir un profundo entendimiento sobre el algoritmo y su entorno operativo, evaluando su pertinencia, eficacia,

	transparencia, utilidad y deseabilidad desde perspectivas éticas, sociales y culturales.
Responsabilidad Social	Identificar y corregir posibles sesgos o comportamientos perjudiciales del algoritmo, buscando hacer sus resultados más predecibles, menos inciertos y más controlables para la sociedad.
Metodología Estructurada	Establecer un proceso claro y riguroso que guíe la auditoría desde su inicio hasta su conclusión, considerando todas las fases y aspectos relevantes del algoritmo y su impacto.

3.3 Tipos de auditoría algorítmica

En función del objetivo, del tipo de caso de uso, del riesgo percibido y del impacto social previsto, así como de las posibles limitaciones prácticas, existen diversos tipos de auditoría. La clasificación más habitual se realiza en base al aspecto o nivel del sistema que se vaya a inspeccionar, de este modo pueden distinguirse los siguientes tipos de auditoria:

- **Auditoría de Datos**: con el objetivo de evaluar los conjuntos de datos utilizados para entrenar o alimentar al algoritmo, esta auditoría se centra en la calidad, la representatividad, la diversidad y la integridad de los datos. Busca identificar sesgos, errores o inexactitudes en los datos que podrían afectar el rendimiento o la justicia del algoritmo.

- **Auditoría de Modelo**: en este caso el objetivo es examinar el modelo matemático subyacente utilizado por un algoritmo para realizar predicciones o clasificaciones, así como analizar la lógica, las reglas y las operaciones que guían cómo un algoritmo procesa la información. El foco son aspectos tales como la precisión del modelo, la robustez, la transparencia del algoritmo, su complejidad. Suele contemplar asimismo cómo se entrenó (en el caso de modelos de aprendizaje automático), y qué tipo de datos se utilizaron. Busca identificar posibles fallas lógicas o puntos de decisión que podrían llevar a resultados no deseados o injustos y evaluar si el modelo es propenso a sesgos o discriminaciones y cómo estos pueden afectar sus predicciones.

- **Auditoría de Sistemas**: en este caso el objetivo es evaluar la infraestructura completa (física, lógica, organizacional y extendida) en la que se encuentra inmerso un algoritmo, incluidos hardware, software, interfaces, procesos, departamentos y partes afectadas. Esta auditoría

considera cómo el algoritmo se integra con otros sistemas, cómo procesa y almacena datos, y cómo interactúa con los usuarios, con terceras partes o con otros sistemas. Busca identificar posibles vulnerabilidades e impactos a nivel de sistema que podrían comprometer el algoritmo, los datos que procesa y las partes afectadas.

Además del nivel de análisis considerado, hay que tener en cuenta que la auditoria puede focalizarse en diferentes aspectos, por ejemplo, privacidad, transparencia, impacto ético, accesibilidad, gobernanza, etc. En la Figura 5 se presentan de manera resumida los tipos de auditoría y aspectos a analizar:

Figura 5. Ejemplo de metodología vs principios

Descripción figura 5. La imagen representa una tabla con cuatro columnas, la primera describe en cada fila los aspectos a analizar, y las tres columnas siguientes los tres tipos de auditorías, por este orden de izquierda a derecha: auditoría de datos, de modelos y de sistemas.

Aspectos a analizar	Auditoría de Datos	Auditoría de Modelo	Auditoría de Sistemas
Partes afectadas	X	X	X
Privacidad	X	X	
Transparencia		X	X
Explicabilidad		X	X
Sesgos	X	X	X
Impacto Ético	X	X	X
Cumplimiento normativo	X		X
Fiabilidad		X	
Autonomía	X		X
Usabilidad		X	
Accesibilidad		X	X
Gobernanza	X		

Es importante remarcar, como se explicará en el siguiente capítulo, que actualmente no existe un estándar tipo de auditoria algorítmica, por lo que en cada caso hay que determinar el alcance, el plan de auditoría, así como determinar el conjunto de técnicas, herramientas y mejores prácticas a utilizar. Dependiendo del contexto y del sistema en cuestión, podría ser necesario realizar varios tipos de auditoría para obtener una comprensión completa de los riesgos y beneficios asociados.

4 Fases de la auditoría

4.1 Fases de la auditoría

A continuación, se presentan las diferentes fases de una auditoría ética de sistemas de IA. En función del tipo de auditoría, tipo de sistema, caso de uso y contexto específico, la extensión y desarrollo de las diferentes fases puede variar.

Las fases principales, tal y como se presentan en la Figura 3 serían las siguientes:

- Definición de alcance.
- Plan de Auditoría.
- Recopilación de información.
- Análisis y evaluación.
- Informe de Auditoría.
- Plan de acción.

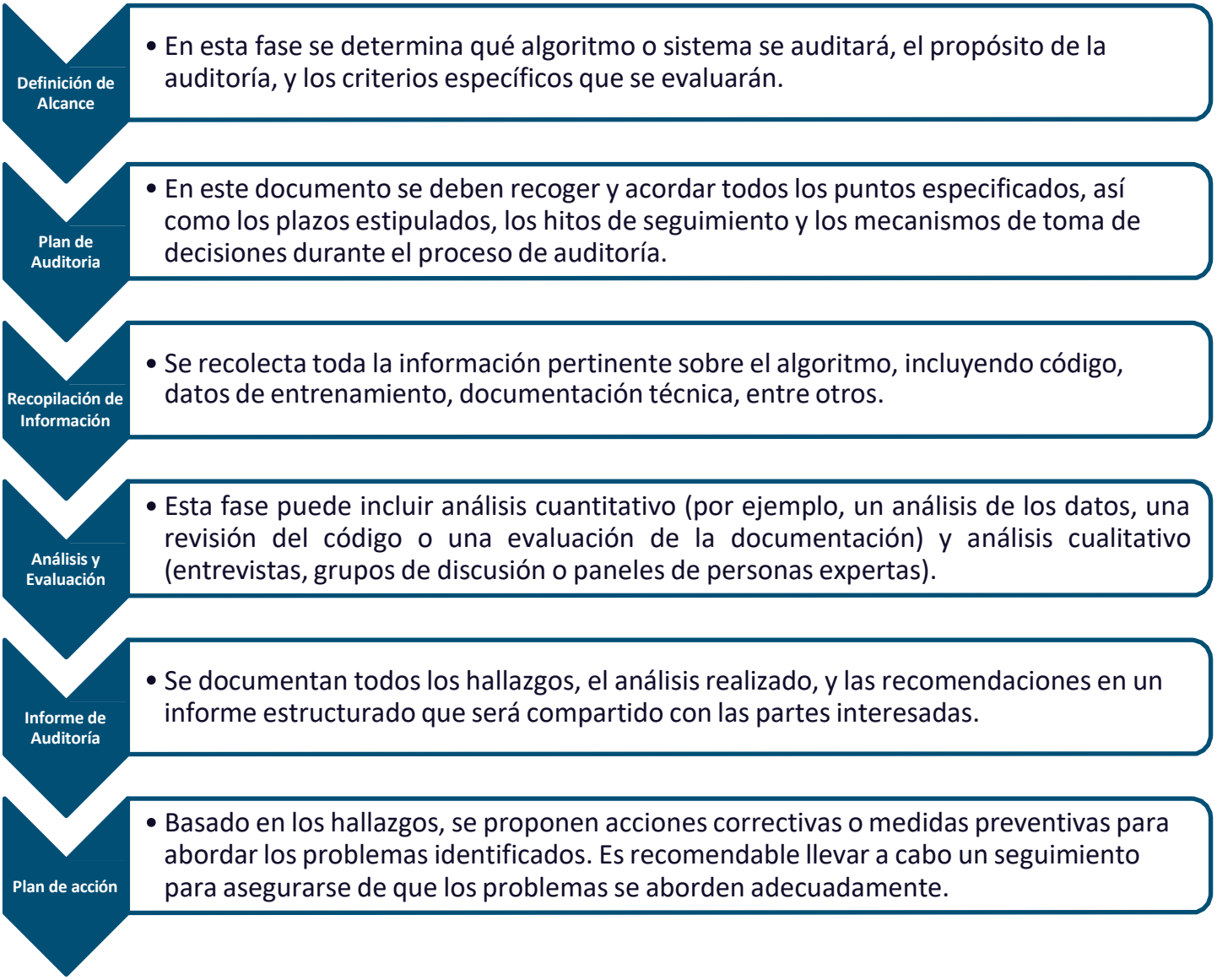


Figura 6. Fases de la Auditoría.

Descripción figura 6. La imagen representa seis rectángulos horizontales en cuya esquina izquierda aparece el nombre de cada fase, acompañado de la descripción en el interior del rectángulo.

Una correcta **definición de los objetivos y del alcance** de la auditoría es el primer paso imprescindible en el proceso. Esta fase establece claramente qué algoritmo o sistema se va a auditar, cuál es el propósito de la auditoría y qué criterios específicos se evaluarán. La importancia de esta fase es clara, ya que permite delinear el perímetro de la auditoría, identificar los recursos, técnicos y humanos, necesarios para la misma, y disponer los criterios específicos o estándares frente a los que realizar la auditoría. Durante esta fase suele ser de utilidad el uso de una lista de

comprobación de la *auditabilidad* del sistema o algoritmo en cuestión, es decir, utilizar una lista de comprobación previa de los requisitos necesarios para llevar a cabo la auditoría. En el Apéndice 2 se incluye un ejemplo de esta lista de criterios de auditabilidad.

> Ejemplo: Una empresa quiere auditar un algoritmo de recomendación utilizado en su plataforma de streaming para asegurarse de que no haya sesgos no deseados.
>
> a. Se identifica el algoritmo de recomendación.
> b. El objetivo es identificar y corregir sesgos en las recomendaciones.
> c. Criterios: Diversidad de géneros recomendados, equilibrio en recomendaciones según demografía del usuario/a, etc.

El resultado de la primera fase será la elaboración del **Plan de Auditoría**, en el que se deben recoger y acordar todos los puntos especificados, así como los plazos estipulados para el desarrollo de esta, los hitos de seguimiento y los mecanismos de toma de decisiones durante el proceso de auditoría.

A continuación, se procedería con la **Recopilación de Información**, que puede consistir en la revisión de la documentación técnica disponible, guías de usuario/a, notas de desarrollo, procedimientos, acceso a los datos de entrenamiento del modelo, e incluso el código fuente del algoritmo. En función del sistema o modelo auditado puede resultar más sencillo o complicado acceder a toda esta información, lo que condicionará la metodología a utilizar durante el proceso de auditoría.

> Ejemplo: En la plataforma de streaming, se recopila (si se dispone de ello):
>
> a. Código del algoritmo de recomendación.
> b. Historial de visualizaciones y preferencias de las personas usuarias.
> c. Documentación sobre cómo funciona el algoritmo, cómo se espera que haga recomendaciones y cualquier cambio hecho en versiones anteriores.

En este punto daría comienzo la fase, habitualmente más extensa de la auditoria, el **Análisis y Evaluación**. Esta fase puede suponer diferentes tipos de análisis, tanto cuantitativos como cualitativos. Se pueden identificar áreas de interés o problemas evidentes, buscar errores comunes o indicios de mala práctica, evaluar los datos en busca de irregularidades, revisar la documentación y asegurarse de que sea clara y refleje lo que realmente hace el algoritmo, usar métricas apropiadas para

determinar cómo se comporta el algoritmo en diferentes condiciones, aplicar técnicas estadísticas para identificar cualquier sesgo en las decisiones del algoritmo, examinar el algoritmo para vulnerabilidades o potenciales brechas de seguridad, llevar a cabo entrevistas con los diferentes grupos de interés o partes afectadas, etc.

Ejemplo: Para el algoritmo de recomendación:

a. Se verifica que el código no tenga errores evidentes y se sigue una estructura lógica. Se observa cómo el algoritmo se comporta con diferentes perfiles de usuarios/as y se mide su precisión y diversidad.

b. Se analizan los datos y se identifica una sobrerrepresentación de ciertos géneros de películas. Se realiza un test para ver si ciertos géneros son sistemáticamente favorecidos o desfavorecidos.

c. Se encuentra que la documentación no menciona cómo se ponderan las preferencias de las personas usuarias.

d. Se verifica si los datos de las personas usuarias están protegidos y si el algoritmo puede ser manipulado externamente.

Una vez completada la fase de análisis y evaluación se elabora el **Informe de Auditoría**, fase en la que se documenta el proceso, los hallazgos más relevantes, los aspectos positivos y las posibles áreas de mejora. En esta fase es importante registrar detalladamente los resultados de cada prueba y análisis realizado, para proporcionar sugerencias y acciones correctivas, que permitan una trazabilidad posterior. El informe puede consistir en una parte de carácter más interno dirigida a la organización y de una parte más externa, que pueda ser compartida con partes interesadas externas.

Ejemplo:

- a. Se documenta que el algoritmo tiene una tendencia a favorecer ciertos géneros y no protege adecuadamente los datos de preferencias de las personas usuarias.
- b. Se sugiere ajustar el peso de ciertos parámetros y fortalecer la seguridad del algoritmo.
- c. Se presenta el informe al equipo de desarrollo y a la dirección de la empresa.

El objetivo de la auditoria debe ser la evaluación y mejora continua del sistema, por lo que es importante desarrollar y acordar un **Plan de Acción** y recomendaciones concretas para ser llevadas a cabo. Estas acciones deben venir acompañadas de unos plazos de ejecución, los recursos necesarios y la persona responsable de su

seguimiento. Una vez implantadas las mejoras, se revisa todo el proceso de auditoría para asegurarse de que se han abordado todas las preocupaciones y se podría dar por concluido el proceso de auditoría. Habría adicionalmente que actualizar cualquier documentación relevante basada en los cambios realizados, y establecer una posible fecha para la siguiente revisión o proceso de auditoría.

Ejemplo:

- a. Se establece un plan para revisar y ajustar los parámetros del algoritmo y para mejorar la seguridad de los datos de usuarios/as.
- b. El equipo de desarrollo realiza los cambios necesarios.
- c. Se vuelve a probar el algoritmo para verificar que ya no tiene los sesgos detectados y que la seguridad es robusta. Se verifica que el algoritmo corregido ya no favorece ciertos géneros y que la seguridad de los datos es adecuada.
- d. Se actualiza la documentación del algoritmo y se documenta todo el proceso de auditoría.

A continuación, se detallan las diferentes metodologías, técnicas y herramientas con relación al análisis cuantitativo y al análisis cualitativo, que como se ha indicado, suelen ser las fases de mayor extensión del proceso de auditoría.

4.2 Análisis cuantitativo

Generalmente, los análisis cuantitativos de las auditorías de impacto ético se centran en dos áreas fundamentales en lo que se refiere a principios éticos: equidad y explicabilidad. Si se tiene en cuenta el ciclo de *machine learning*, desde la recopilación de datos, limpieza, modelado del algoritmo y despliegue, en cada parte de este proceso será relevante entender qué está ocurriendo en términos de esos principios éticos. La privacidad, en lo que a análisis cuantitativo se refiere, no suele incluirse en este tipo de análisis, asumiendo que hay un proceso en marcha separado dentro de la organización que se encarga ya de aplicar la normativa RGPD.

Figura 7. CRISP-DM vs tipo de análisis

Descripción figura 7. La imagen representa una tabla con dos columnas y siete filas donde se recoge el nombre de cada fase del modelo CRISP-DM (columna de la izquierda) y el tipo de análisis (columna de la derecha).

CRISP-DM	Tipo de análisis
Entendimiento del negocio	N/A
Comprensión de los datos	Exploratorio Frecuencia de los datos
Preparación de los datos	NA
Modelado	Detección de sesgos Explicabilidad
Evaluación	Revisión de sesgos y explicabilidad
Implantación	Monitorización continua

Es clave entender el ciclo de *machine learning* para poder realizar este análisis cuantitativo y para eso el modelo CRISP-DM puede resultar de utilidad. Desde hace años CRISP-DM, representado en la Figura 4, es el modelo más usado para el diseño de proyectos de minería de datos y actualmente según la Data Science Process

Alliance [27]es el más utilizado para los proyectos de *machine learning*. El modelo CRISP-DM, explicado con más detalle en el apéndice, permite además realizar una primera aproximación previa al análisis cuantitativo teniendo en cuenta los pasos necesarios y el impacto ético que puede haber en cada uno (por ejemplo, el acceso a los datos, conjunto de datos desbalanceados, sesgo en el diseño de los algoritmos, etc.) sin necesidad de ser una persona experta en las técnicas matemáticas.

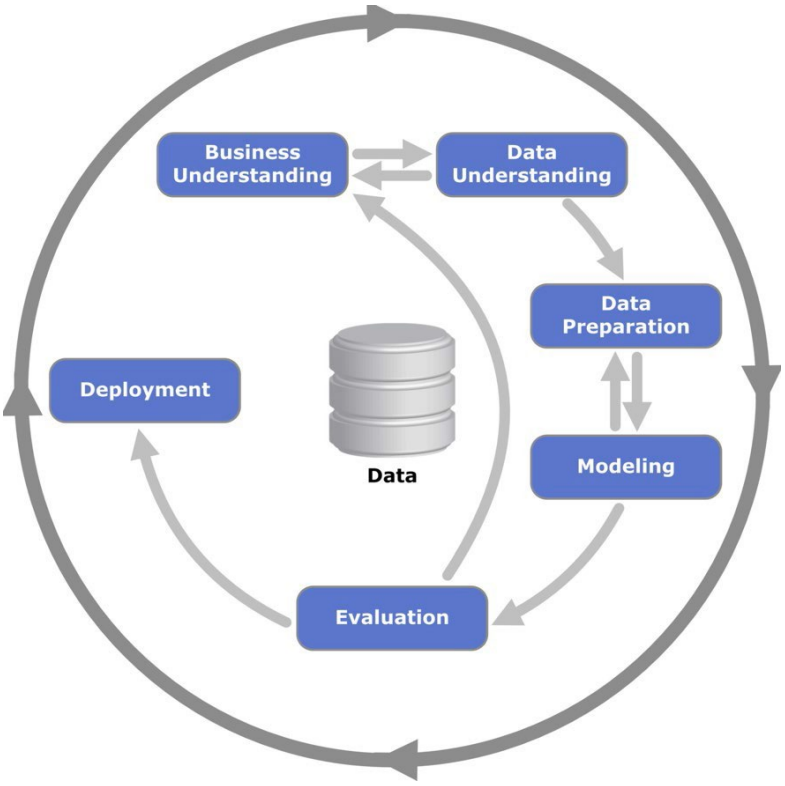


Figura 8. Metodología CRISP-DM.

Descripción figura 8. La imagen describe en un círculo las fases del proceso en el que consiste la metodología CRISP-DM).

La idea fundamental de usar CRISP-DM es que es una metodología que puede ser entendida de manera sencilla por cualquier profesional, aunque luego se requiera en cada uno de los pasos distintas personas expertas para la realización de este

[27] https://www.datascience-pm.com/crisp-dm-still-most-popular/

análisis cuantitativo. No obstante, en la literatura académica se puede encontrar alguna propuesta del modelo de CRISP-DM adaptada a los tipos de sesgos[28], si se quiere profundizar en ello.

En las primeras etapas de preparación de los datos, evaluación y modelado se requiere para llevar a cabo este análisis conocimientos estadísticos y matemáticos, pero en algunos casos es posible tener una idea a más alto nivel, utilizando técnicas como las de toma de muestras y ejecución del algoritmo para ver cuál es el resultado. Respecto a la parte de modelado es cada vez más habitual que los diseñadores de modelos de *machine learning*, consideren en elegir modelos más interpretables. En general, a los modelos de *machine learning* que permiten mayor interpretabilidad se les denomina de caja blanca (por ejemplo un modelo de regresión lineal) y a los modelos que no tienen esa interpretabilidad se les denomina de caja negra (por ejemplo ChatGPT). Esto significa que es más sencillo explicar la decisión tomada por el algoritmo. En muchas ocasiones esto no es posible por el tipo de algoritmo utilizado en el desarrollo del modelo, pero sí pueden buscarse técnicas de explicabilidad como LIME, SHAP o metodologías similares para explicar estos algoritmos.

En el primer paso de la metodología de CRISP-DM, **Entendimiento del Negocio** *o* del problema que se va a resolver es donde se decide junto con la persona experta de negocio qué tipo de atributos van a ser usados (ej. salarios, edad, antigüedad del cliente, saldo mensual). Es aquí donde se valida qué tipo de algoritmos han sido usados para la resolución de problemas similares y qué otros atributos adicionales podrían seleccionarse, para complementar el análisis. En lo que respecta al análisis cuantitativo no aplica en esta fase, pero sí es importante que estén documentados la selección de los atributos, así como la definición clara del problema de negocio que se quiere resolver. En esta fase es complejo de manera cuantitativa detectar sesgos, porque no existe una definición de equidad genérica para los distintos sectores, pero sí debiera acordarse una definición para las partes interesadas que intervienen en esta etapa.

[28] https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1210&context=hicss-55

En el segundo paso **Comprensión de los Datos**, entendimiento de los datos, es uno de los pasos clave para el diseño de un modelo de *machine learning*. Es en este punto donde se tienen que realizar los siguientes análisis:

- Procedencia de los datos.
- Almacenamiento de los datos.
- Posible introducción de sesgos en los datos.

Los dos primeros análisis tienen que ver con el cumplimiento del RGPD, y el último tiene que ver con la selección de los atributos y si se ha podido producir algún tipo de sesgo en esta selección. Aquí es importante utilizar análisis de covarianza, así como ver la posible correlación de las variables seleccionadas.

Es en este punto donde es clave identificar los tipos de datos a incluir para no cometer errores posteriores en la predicción. Por ejemplo, la aplicación de un sistema de navegación de coches autónomos que no incluía la posibilidad de que las personas fueran en bicicleta y el coche autónomo consideró que no era un obstáculo[29]. Este mismo caso podría ocurrir con una silla de ruedas o cualquier otro vehículo usado por personas con discapacidad. Es importante considerar siempre la muestra adecuada para el entrenamiento de los datos para evitar posibles sesgos que puedan conllevar una generalización no adecuada del algoritmo. Otro caso por ejemplo es el de la compañía HIREVUE[30] americana, que dispone de una plataforma de análisis de video para acelerar el proceso de contratación. Esta plataforma no tenía en cuenta a las personas con discapacidad en la muestra de los datos, esto generaba un sesgo de cara a la contratación.

En este paso es donde se hace un análisis exploratorio de los datos y es clave ver la frecuencia de los mismos, para entender posibles faltas de representatividad de estos, así como la relación entre las variables. Es aquí donde pueden aparecer las variables proxy. Las variables proxy son aquellas que no tienen especial relevancia

[29] Troy Griggs and Daisuke Wakabayashi, "How a Self-Driving Uber Killed a Pedestrian in Arizona," *New York Times*, March 21, 2018,
[30] HireVue website, *HireVue,* acceso en Noviembre de 2023, htps://www.hirevue.com/.

aisladas pero que tratadas en conjunto pueden revelar información sensible. Por ejemplo, si se está validando un derecho de la persona en base a una serie de características a, b y c la variable proxy sería una variable combinada usando a, b y c que normalmente se acaba creando como una variable combinada.

En lo referente a la frecuencia de los datos en primer lugar se hace un análisis de los datos en base a las variables utilizadas y se ve la frecuencia que tienen cada una de ellas en el conjunto de datos (*dataset*). Veamos un ejemplo cogiendo el típico *dataset* del Titanic que trata de predecir la supervivencia de estos pasajeros, con la estructura que se muestra a continuación.

Básicamente incluye el número de pasajeros, la clase donde viajó, el nombre, la edad, el género, la tarifa, la cabina y si sobrevivió o no.

Figura 9. Listado de muestra de datos del dataset Titanic

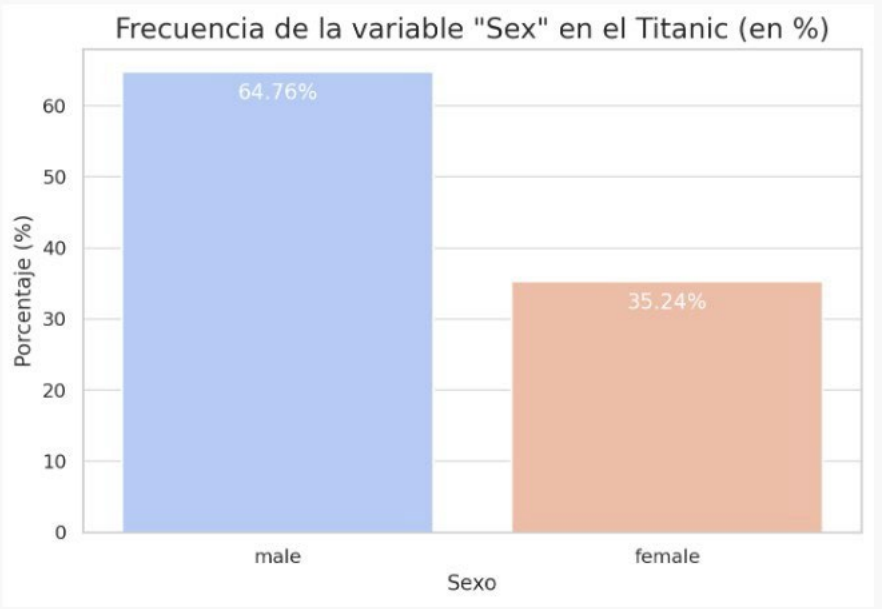
Descripción figura 9. La imagen representa una tabla con doce columnas y doce filas donde se recogen varios datos de las/os pasajeros del TITANIC.

Pas	Survi	Pcla	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7,25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71,2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7,925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53,1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8,05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8,4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51,8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21,075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11,1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30,0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16,7	G6	S

Si analizamos la frecuencia de los datos fijándonos en las características de sexo, se puede ver a continuación la distribución correspondiente.

Figura 10. Frecuencia de la variable Sexo

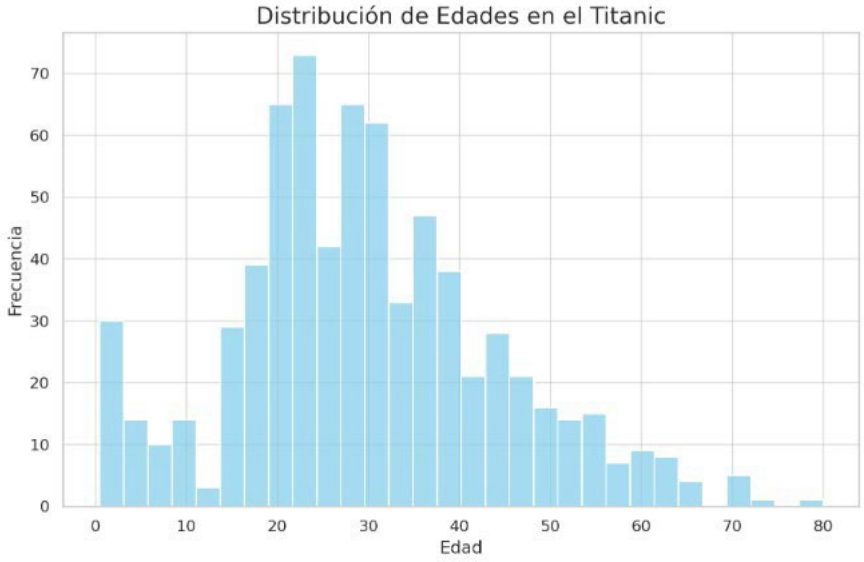
Descripción figura 10. La imagen representa un gráfico de columnas que describe la distribución en porcentaje de la variable sexo en el Titanic (64,76% hombres,35,24% mujeres).



Con esta simple proporción se puede observar que había más hombres que mujeres en el *dataset* analizado y esto puede tener una influencia a la hora de obtener una predicción del modelo. Si hacemos un análisis similar de la edad, esto nos dará también una idea de qué rango de edad se encuentra en el *dataset.*

Figura 11. Frecuencia de edades en el dataset del Titanic

Descripción figura 11. La imagen representa un histograma que describe la distribución de la variable edad en el Titanic.

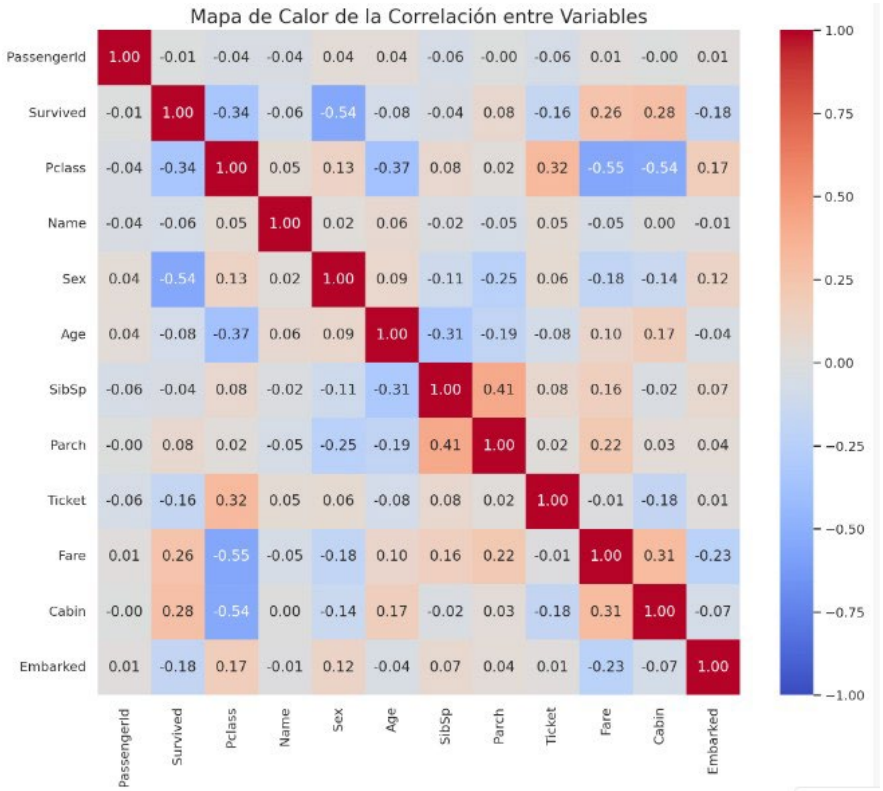


En este punto del proceso, una vez validada la frecuencia de distintas variables se puede establecer una correlación entre las variables para entender cómo se

relacionan entre ellas. Para ello las variables categóricas como el sexo, por ejemplo, se pasan a un valor numérico.

Figura 12. Mapa de correlación de las variables

Descripción figura 12. La imagen representa una matriz de correlación entre las diferentes variables analizadas utilizando un mapa de calor.



Si se tiene en cuenta la variable **sexo**, y la variable **supervivientes**, se ve que hay una correlación bastante alta entre estas variables, igual que ocurre con la tarifa de compra del billete y con la cabina en la que viajaba el pasajero. Si nos estuviéramos refiriendo a variables de asignación de recursos en una subvención, por ejemplo, sería importante detectar estas correlaciones desde un principio para poder corregirlas si fuera necesario.

Es en este paso donde se determina si hay suficientes muestras en el modelo para poder analizar su eficacia teniendo en cuenta la frecuencia de los datos. Además, es también aquí donde se observa si hay grupos que tengan una representatividad muy baja respecto al total de la muestra. Es aquí también donde se estudia si las variables que se están analizando son realmente las adecuadas, o sería necesario analizar variables diferentes teniendo en cuenta lo que el algoritmo trata de predecir.

En el tercer paso **Preparación de los Datos**, se trata de preprocesar los datos antes de utilizarlos para generar el algoritmo. En general, se trata de tener una estandarización de los datos, en un único formato, dado que los datos pueden ser estructurados, semiestructurados y no estructurados. Esta etapa suele ser la más costosa en cuanto a duración se refiere dadas las transformaciones necesarias a implantar. En esta etapa el foco está en las técnicas de preproceso de datos y no suele ser común que incluyan sesgos.

El cuarto paso, **Modelado de Datos** es uno de los puntos clave del diseño, y aunque es el punto donde se concentra todo el esfuerzo del científico/a de datos en sus primeros pasos, suele ser otro de los puntos donde se pueden incluir sesgos y donde el concepto de explicabilidad es más relevante. Con respecto a los sesgos, es importante entender que en esta fase se utiliza la optimización o minimización del error para conseguir el mejor algoritmo, esta estrategia puede entrar en conflicto con la reducción del sesgo. Es decir, la minimización de errores en el proceso de diseño del modelo puede llevar a incrementar el sesgo según se haya definido éste. Es por eso que en algunos casos es extremadamente importante utilizar herramientas que puedan ayudar a detectar y/o mitigar el sesgo en esta fase y que nos ayuden a tomar decisiones. Estas técnicas dependen del modelo de *machine learning* utilizado (clasificación o regresión) y es importante entender que siempre se tendrá que establecer un equilibrio entre la precisión que se quiere obtener y la corrección de los sesgos.

Es en este punto donde se pueden analizar las distintas métricas y el análisis de la matriz de confusión que nos va a permitir obtener métricas relevantes para la detección de sesgos. Estas métricas aparecen detalladas en el anexo 7.4.

Si volvemos al ejemplo del Titanic, lo primero que se hace es entrenar con estos datos una regresión logística o un *random forest* y calcular la precisión de ambos modelos.

En este caso se crean las métricas correspondientes para detectar si hay un sesgo, y se obtiene que los hombres son el grupo desfavorecido, aunque su distribución es mayor en la muestra. Esto, analizando los datos reales, tiene mucho sentido dado que fueron las mujeres y las niñas y niños a quienes se les dio prioridad para abandonar el barco.

Una vez que se conoce cuál es el grupo desfavorecido se pueden aplicar distintos métodos, para hacer un rebalanceo de los datos e intentar que el grupo desfavorecido no lo sea. Si se utiliza la técnica de *reweighing*, esta técnica no cambia los datos del dataset sino que les asigna nuevos "pesos" para cada una de las filas. En este caso, por ejemplo:

- Los ejemplos de hombres que no sobreviven se ponderarán a la baja.
- Los ejemplos de hombres que sobrevivan se ponderarán al alza.
- Los ejemplos de mujeres que sobrevivan se ponderarán a la baja.
- Los ejemplos de mujeres que no sobreviven se ponderarán al alza.

Con esto se ejecuta el modelo de nuevo y se comparan las métricas para ver la evolución representando el modelo a continuación.

Figura 13. Mapa de métricas de sesgos

Descripción figura 13. La imagen representa dos gráficos de líneas que reproducen la evolución de los sesgos: a la izquierda el modelo de algoritmo original, a la derecha el modelo ejecutado de nuevo tras el rebalanceo de los datos.

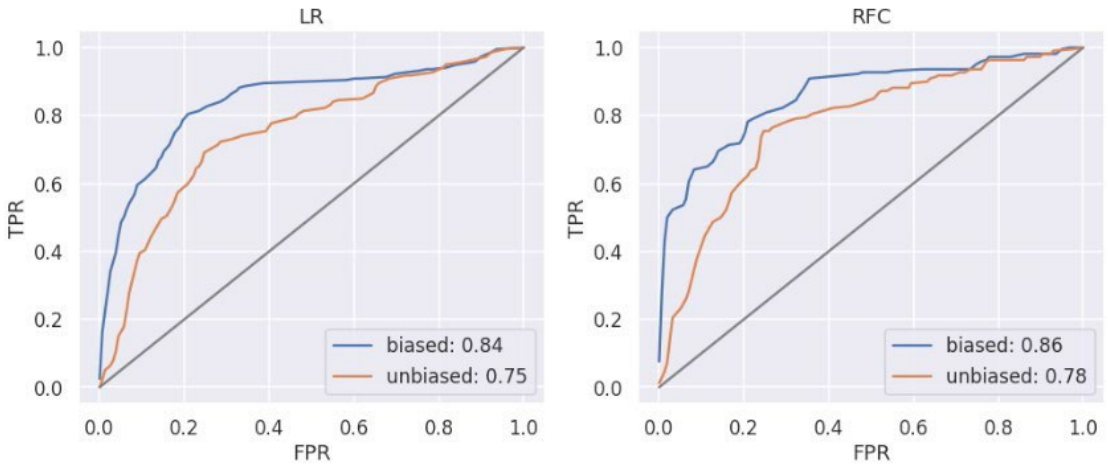


Figura 14. Listado de precisión de los modelos probados

Descripción figura 14. Tabla con tres columnas y tres filas que describe el listado de precisión de los modelos, la primera columna describe el tipo de modelo, la segunda el grado de precisión del modelo con sesgo y la tercera el grado de precisión del modelo sin sesgo.

Modelo	**Precisión *(Accuracy)* (modelo con sesgo)**	**Precisión *(Accuracy) (modelo sin sesgo)***
Random Forest	0,80	0,69
Regresión Logística	0,77	0,71

Como se puede ver la precisión ha variado, es decir, se establece una mejor equidad en el modelo, pero esto afecta a la precisión. Siempre es así y hay que tener en cuenta que cuando se actúa en un modelo para intentar mitigar los sesgos del algoritmo en la mayoría de los casos puede variar su precisión. Es por eso que es importante elegir bien el método a aplicar y tener claro que el algoritmo se va a ver afectado en su toma de decisiones cuando se trata de ajustar el sesgo.

Por esta razón muchas veces se utilizan modelos de *machine learning* más simples pero más explicables.

Adicionalmente, en este punto también es clave contar con la diversidad adecuada en los equipos de desarrollo del modelo, dado que se pueden incluir sesgos también en base al género, raza, procedencia, de las propias personas diseñadoras del modelo.

Es en este punto del proceso donde adquiere mucha relevancia el principio de explicabilidad. Dado que es en este paso donde se eligen los algoritmos a utilizar, es clave tener en cuenta qué tipo de algoritmo se va a seleccionar no sólo teniendo en cuenta la mejor optimización del modelo, sino además la explicabilidad de éste. De manera general, los modelos que son más explicables, en los que se puede entender y explicar la decisión tomada por el algoritmo se denominan modelos de caja blanca. Estos modelos permiten explicar la toma de decisión del algoritmo dada la sencillez de éste. En este tipo se incluyen los algoritmos basados en reglas, la regresión lineal o los árboles de decisiones, entre otros. Los algoritmos de caja negra son aquellos en

los que no es sencillo de entender cómo ha sido la toma de decisiones, y necesitan de técnicas adicionales para obtener esta explicabilidad. Este tipo de algoritmos, en general, suelen obtener mejores resultados que los algoritmos de caja blanca, y a veces, las personas que son científicas de datos pueden tender a su uso basándose solamente en este criterio y olvidando la explicabilidad. En la Figura 15 se puede ver la relación de la explicabilidad con la complejidad del modelo.

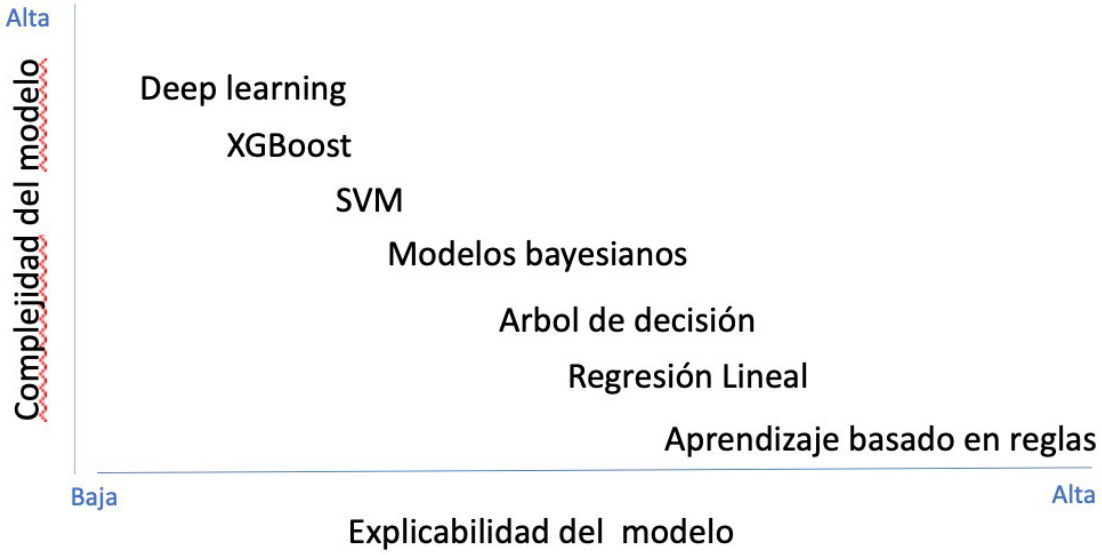


Figura 15. Explicabilidad vs Complejidad del modelo de ML.

Descripción figura 15. La imagen representa un diagrama que describe la relación entre explicabilidad y complejidad del modelo.

Si continuamos con el modelo del Titanic, si se utiliza una Regresión logística en muchos casos la precisión es mejor pero el modelo no puede explicarse de manera sencilla. En cambio, usando Random Forest, sí es más claro entender cómo el modelo está tomando las decisiones y cómo es el impacto de la variable sexo. En el árbol que se muestra a continuación, se ve claramente el patrón de decisión que sigue el modelo de acuerdo a los valores de las distintas variables. El color naranja indica que no sobrevivió y el azul que sí sobrevivió.

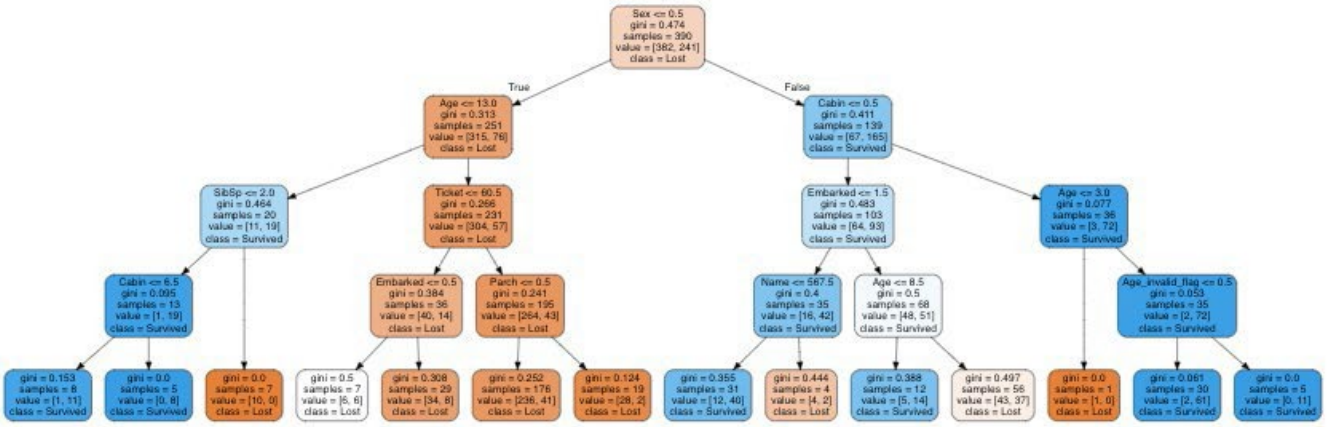


Figura 16. Random Forest para un ejemplo concreto

Descripción figura 16. La imagen representa un diagrama de árbol de decisión que utiliza el algoritmo Random Forest en el ejemplo descrito en el texto.

De hecho, tal y como se contempla en la RGPD, las empresas deberían poder explicar las decisiones que toman sus algoritmos, si bien es verdad que no incluye una penalización a las empresas por ello, sí que se incluye como recomendación o buena práctica.

El quinto paso, **Evaluación,** es uno de los más importantes desde el punto de vista de una auditoría. Es en este paso cuando se evalúa si el modelo es suficientemente bueno generalizando los resultados y se pone mucho foco en los resultados de precisión, olvidando si puede haber un posible sesgo no detectado o algún otro tipo de incumplimiento con algún principio clave de la empresa. Por ejemplo, en un modelo de aprobación de préstamo sería bueno generar distintos dataset incluyendo atributos como el género, la raza, la procedencia o la edad y ejecutar el algoritmo para observar que no se está llevando a cabo ninguna discriminación.

El último paso es el **Despliegue** del algoritmo. Esto no afecta a ninguno de los principios como la equidad o la explicabilidad, si bien es verdad que es en el despliegue donde pueden ocurrir los primeros errores y detectar en este punto algún error que pueda afectar a la precisión del algoritmo, pero también a la equidad. Es importante saber detectar este problema en esta fase e intentar identificar el tipo de

error que lo ha podido producir (falta de entrenamiento, falta de datos adecuados, etc.) y revisar en qué fase hay que volver a actuar para corregir el error.

Como se veía anteriormente, existen distintas alternativas a la hora de evaluar el análisis cuantitativo de los sistemas de inteligencia artificial. Para este tipo de análisis se requiere normalmente de conocimientos matemáticos, estadísticos y un conocimiento detallado de las técnicas de *machine learning*. Sin embargo, el surgimiento de herramientas comerciales está contribuyendo a facilitar este tipo de análisis a perfiles profesionales no tan técnicos. Asimismo, para perfiles más técnicos existen cada vez más librerías en Python y R, para el análisis de los sesgos y su mitigación así como su explicabilidad.

En la tabla siguiente (Figura 17) se han tratado de incluir las herramientas más representativas para la detección de sesgos y/o análisis de la explicabilidad. En la tabla se puede ver la fecha de creación de las herramientas, algunas relativamente recientes y además se incluye el área donde más foco pone la herramienta, sesgos o explicabilidad.

Herramientas	Compañía	Año	SaaS/On pre	Sesgos	Detección Sesgos	Mitigación de sesgos	Explicabilid
Aequitas	Aequitas	2018	SaaS	S	S	N	N
Audit-AI	Pymetrics	2018	On premise	S	S	S	N
Datarobot	Datarobot	2020	Ambos	S	S	S	S
Fairlearn	Microsoft	2018	On premise	S	S	S	N
Fairly AI	Fairly	2020	SaaS	S	S	S	S
Fairness 360	IBM	2018	Ambos	S	S	S	S
Fairness Flow	Facebook	2018	On premise	S	S	N	N
Fiddler AI	Fiddler	2018	SaaS	S	S	S	S
H20	H20	2018	SaaS	S	S	S	S
InterpretML	Microsoft	2020	On premise	N	N	N	S
VerifyML	Cylynx	2021	SaaS	S	S	S	N
What if /Mindiff	Google	2018	On premise	S	S	N	N

Figura 17. Tabla con las herramientas más representativas para la detección de sesgos y/o explicabilidad

Descripción figura 17. La imagen representa una tabla donde se indican las herramientas más representativas en la detección de sesgos, la compañía a la que pertenece, el año de creación, el tipo de software que utiliza, siendo las opciones: Software como servicio (SaaS), en instalaciones propias (On Premise), o ambos), si detecta (señalado con la letra S) o no (indicado con la letra N) la existencia de sesgos, y si trata de mitigarlos.

Análisis cualitativo

La metodología cualitativa se presenta como una herramienta indispensable en la auditoría ética de sistemas de inteligencia artificial. Dado que estos sistemas no operan en un vacío, sino dentro de complejos entornos socioculturales, es fundamental entender no solo sus componentes técnicos, sino también sus implicaciones humanas y contextuales. Efectivamente, los sistemas de inteligencia artificial, al integrarse en la sociedad, forman lo que se conoce como sistemas sociotécnicos. Estos sistemas no solo involucran tecnología, sino también las interacciones humanas y las estructuras sociales en las que se insertan. Estos sistemas no son meramente la suma de sus componentes tecnológicos y sociales, sino que su interacción crea dinámicas y propiedades emergentes que no podrían ser anticipadas observando estos componentes de manera aislada. La relevancia de considerar a la IA desde una perspectiva sociotécnica radica en la necesidad de entender cómo la tecnología y la sociedad coevolucionan. Ignorar la dimensión social puede llevar a implantaciones tecnológicas que, aunque sofisticadas, podrían ser inapropiadas, no éticas o incluso perjudiciales para ciertos colectivos dentro de la sociedad.

Desde una perspectiva ética, la consideración de los sistemas de IA como sistemas sociotécnicos implica que una auditoría o revisión de estos sistemas no debe limitarse a analizar la tecnología en sí, sino también cómo esta se inserta y afecta al tejido social en el que opera. Por ejemplo, determinados sesgos y discriminaciones pueden surgir tanto del propio diseño algorítmico como de la interacción con estructuras sociales previamente sesgadas. Por otra parte, en un sistema sociotécnico, la responsabilidad de las acciones y decisiones no recae únicamente en el sistema tecnológico o en sus desarrolladores. Dado que hay una interacción constante con y entre actores humanos, la responsabilidad puede ubicarse en diferentes actores involucrados. Por último, dado el carácter emergente de los sistemas sociotécnicos, es crucial estar atento a consecuencias no previstas de la implantación de la IA en contextos sociales, que pueden manifestarse de maneras impredecibles.

En resumen, un sistema sociotécnico desde el punto de vista ético de la inteligencia artificial reconoce que la tecnología no opera en un vacío. Por el contrario, está inextricablemente ligada a, e influenciada por, estructuras y valores

sociales, y su implantación ética requiere una profunda consideración de estas interacciones y su impacto en el bienestar humano y social. Si aplicamos este enfoque a los sistemas sociotécnicos de IA relacionados con las personas con discapacidad queda patente la necesidad de complementar el análisis cuantitativo y de datos con técnicas de análisis cualitativo.

A través de este enfoque cualitativo, se busca capturar las sutilezas, percepciones y experiencias que los enfoques cuantitativos podrían pasar por alto. En el ámbito de la IA, donde las decisiones algorítmicas afectan directamente a las personas y sociedades, comprender estas dimensiones es esencial para una implantación ética y responsable.

La metodología cualitativa puede permitir identificar: aspectos relativos a sesgos y discriminaciones (¿el sistema reproduce o amplifica desigualdades existentes?), transparencia y explicabilidad (¿las personas usuarias y afectadas entienden cómo funciona el sistema y por qué toma ciertas decisiones?), autonomía y privacidad (¿el sistema respeta la privacidad y autonomía de las personas?), o consecuencias no intencionadas (a veces, los sistemas pueden tener efectos colaterales no previstos que solo un análisis cualitativo puede capturar), entre otros.

La metodología cualitativa ofrece diversas técnicas que se pueden adaptar según el objeto de estudio y el contexto, tales como entrevistas, grupos de discusión, observación participante y no participante, y paneles de personas expertas, entre otras[31].

Las **entrevistas** permiten obtener visiones profundas de los/as participantes sobre un tema, así como información detallada sobre sus experiencias, opiniones y percepciones. Puede ser especialmente útil para entender las perspectivas de quienes diseñan y desarrollan, así como de las personas usuarias finales. Las entrevistas pueden ser semiestructuradas (aunque hay un esquema de preguntas

[31] Aunque no suele ser de aplicación para las auditorías éticas de IA, existe otra técnica de investigación cualitativa, conocida como estudio etnográfico, que involucra una inmersión prolongada en la comunidad o grupo de interés, y puede combinar observaciones, entrevistas y otros métodos para obtener una comprensión profunda del contexto cultural y social en el que se implementa y utiliza un sistema de IA. Puede revelar cuestiones éticas arraigadas en el contexto sociocultural.

predefinidas, el entrevistador/a tiene la libertad de desviarse de ellas para profundizar en ciertos temas según las respuestas del entrevistado/a) o abiertas (no hay un esquema fijo de preguntas, y la conversación puede fluir con base en los intereses del entrevistador/a y del entrevistado/a).

Los **grupos de discusión** ayudan a capturar diversas perspectivas y a identificar puntos de consenso o conflicto. Consisten en reuniones de un grupo pequeño (6-12 personas) para discutir un tema específico, guiados por una persona moderadora, mientras un observador/a toma notas o graba la conversación. Estos grupos pueden resultar particularmente útiles para evaluar la aceptabilidad o preocupaciones éticas de un sistema entre diferentes grupos de interés.

La técnica de **observación** participante y no participante ofrece una visión desde el terreno sobre cómo se utiliza e interactúa con la tecnología. En el primer caso la persona que investiga se involucra en la comunidad o grupo que está estudiando, mientras que, en la observación no participante, la persona que investiga observa sin intervenir activamente en la comunidad o grupo. Esta técnica puede ser útil para entender cómo se utiliza un sistema en un entorno real, cómo los usuarios/as interactúan con él, o para identificar problemas éticos no anticipados en el diseño.

Los paneles de personas expertas consisten en reunir a expertos/as en un área específica para discutir y evaluar un tema, en una sesión que puede ser estructurada con preguntas específicas o más abiertas. Es una técnica muy adecuada para evaluar aspectos técnicos, éticos o legales del sistema. Las personas expertas pueden ofrecer perspectivas informadas y críticas sobre el diseño y la implantación del algoritmo.

Para implantar estos métodos eficazmente en una auditoría ética de IA, es crucial tener claridad en los objetivos de la investigación, seleccionar el método adecuado para el propósito y ser meticuloso en el registro y análisis de los datos recolectados. Es igualmente importante considerar la ética en la propia investigación, garantizando la confidencialidad, el consentimiento informado y el respeto a los derechos de las personas participantes. Estas técnicas, aplicadas de manera rigurosa, proporcionan datos ricos en matices y contextos, fundamentales para una auditoría ética.

5 Conclusiones

A medida que la Inteligencia Artificial adquiere un papel más preponderante en diversos sectores, es vital asegurar que su implantación y uso sean éticos y estén alineados con los principios que defienden los derechos y la inclusión de las personas con discapacidad. El CERMI ha identificado y enfatizado esta necesidad, proporcionando un marco de referencia en el que se pueden basar futuras acciones.

La evolución y necesidad de establecer principios éticos claros para la IA ha sido reconocida a escala mundial. Los aspectos regulatorios, incluida la evolución de la regulación sobre IA en la Unión Europea (Ley Europea de IA), proporcionan un marco legal que aspira a guiar el desarrollo y uso de la IA. Estos marcos son fundamentales para garantizar que la IA no perjudique ni margine a ningún grupo, especialmente a las personas con discapacidad.

La necesidad de revisar el uso de la IA y la importancia de auditar los algoritmos es crucial para garantizar que se tomen decisiones imparciales y justas. Las diferentes modalidades de auditoría algorítmica identificadas permiten a las organizaciones adoptar un enfoque adecuado según sus necesidades y objetivos. Una adecuada auditoría ética de los sistemas de IA, desarrollada con rigor y siguiendo una metodología inspirada en esta guía, puede ayudar a identificar y abordar posibles sesgos, inexactitudes y aspectos no éticos de la IA. Es importante definir claramente el alcance y objetivos de la auditoría, así como el uso de técnicas cuantitativas y cualitativas en el desarrollo de la misma.

Es fundamental que las organizaciones, especialmente aquellas que representan a las personas con discapacidad y sus familias, adopten y apliquen esta guía de auditoría ética. La IA tiene el potencial de ser una herramienta poderosa para la inclusión y equidad, pero solo si se utiliza de manera ética y responsable.

El trabajo que el CERMI ha realizado con este documento pone de manifiesto la necesidad de un enfoque proactivo hacia la ética en la IA, garantizando que los avances tecnológicos sean beneficiosos y no perjudiciales para las personas con discapacidad. Es nuestra responsabilidad colectiva asegurarnos de que las tecnologías emergentes, como la IA, se utilicen de manera que respeten y promuevan los derechos humanos.

6 Referencias

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*, January–June: 1–8. https://doi.org/10.1177/2053951720983865

Kaptein, M. (1998). Ethics Management: Auditing and Developing the Ethical Content of Organizations. Springer.

Lam, M. S., Gordon, M. L., Metaxa, D., Hancock, J. T., Landay, J. A., & Bernstein, M. S. (2022). End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proc. ACM Hum.-Comput. Interact., 6*(CSCW2), Article 512. https://doi.org/10.1145/3555625

Lillywhite, A., & Wolbring, G. (2020). Coverage of Artificial Intelligence and Machine Learning within Academic Literature, Canadian Newspapers, and Twitter Tweets: The Case of Disabled People. *Societies*, 10(23). https://doi.org/10.3390/soc10010023

Lucchi López-Tapia, Y. D. (2023). Justicia digital y discapacidad: aprovechando la oportunidad.

Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. Science and Engineering Ethics, 27(44), 1-30. https://doi.org/10.1007/s11948-020-00282-y

Narayanan, M., & Schoeberl, C. (2023). A Matrix for Selecting Responsible AI Frameworks. *Issue Brief*. Center for Security and Emerging Technology.

Nugent, S., Jackson, P., Scott-Parker, S., Partridge, J., Raper, R., Shepherd, A., Bakalis, C., Mitra, A., Long, J., Maynard, K., and Crook, N. (2020). Recruitment AI has a Disability Problem: questions employers should be asking to ensure fairness in recruitment. Institute for Ethical Artificial Intelligence

Olmeda, M. V., & Ibánez, J. C. (2022). Manual de ética aplicada en Inteligencia Artificial. Anaya Multimedia.

Packin, N. G. (2021). Disability Discrimination Using AI Systems, Social Media and Digital Platforms: Can We Disable Digital Bias? Journal of International and Comparative Law, 8(2), 487. https://doi.org/10.2139/ssrn.3724556

Porter, B., Simon, J., & Hatherly, D. (2014). Principles of External Auditing (4th ed.). Wiley.

Prem, E. (2023). From ethical AI frameworks to tools: a review of approaches. *AI and Ethics, 3*(699-716). https://doi.org/10.1007/s43681-023-00258-9.

Raab, C. D. (2020). Information privacy, impact assessment, and the place of ethics. Computer Law and Security Review, 37, 105404. https://doi.org/10.1016/j.clsr.2020.105404

Rosthorn, J. (2000). Business Ethics Auditing – More Than a Stakeholder's Toy. Journal of Business Ethics, 27, 9-19. https://doi.org/10.1023/A:1006417426632

Salgado-Criado, J., & Fernández-Aller, C. (2021). A wide human-rights approach to artificial intelligence regulation in Europe. IEEE Transactions on Systems, Man, and Cybernetics, DOI: 10.1109/MTS.2021.3056284

Smith, P., & Smith, L. (2021). Artificial intelligence and disability: too much promise, yet too little substance? AI Ethics, 1, 81–86. https://doi.org/10.1007/s43681-020-00004-5

Valle Escolano, R. (2023). Artificial intelligence and rights of people with disabilities: The power of algorithms.

Whittaker, M., Alper, M., Bennett, C. L., Hendren, S., Kaziunas, L., Mills, M., Morris, M. R., Rankin, J., Rogers, E., Salas, M., & West, S. M. (2019). Disability, Bias, and AI. *AI Now Institute at New York University*.

Referencias en línea a las que se ha accedido en octubre de 2023

A European Strategy on Data. (s. f.). European Commission. https://commission.europa.eu/document/d2ec4039-c5be-423a-81ef-b9e44e79825b_es

AI & Disability: EU AI Act. (s. f.). OECD. https://oecd.ai/en/wonk/eu-ai-act-disabilities

AI for Inclusive Sidewalks. (s. f.). Smart Cities for All. https://smartcities4all.org/ai-for-inclusive-sidewalks/

AI HLEG Steering Group of the European AI Alliance. (s. f.). European Commission. https://ec.europa.eu/futurium/en/european-ai-alliance/ai-hleg-steering-group-european-ai-alliance.html

AIS eLibrary. (s. f.). AIS. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1210&context=hicss-55

AlphaFold. (s. f.). EMBL-EBI. https://alphafold.ebi.ac.uk

Article 4: General Obligations. (s. f.). United Nations. https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities/article-4-general-obligations.html

Big Picture. (s. f.). AI Ethics Lab. https://aiethicslab.com/big-picture/

Buchanan, M. (2018, abril 4). *Cambridge Analytica Scandal Fallout*. The New York Times. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

CRISP-DM: Still the Most Popular. (s. f.). Data Science PM. https://www.datascience-pm.com/crisp-dm-still-most-popular/

Dastin, J. (2018, octubre 10). *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Directiva (UE) 2019/882 del Parlamento Europeo y del Consejo. (2019, abril 17). EUR-Lex. https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32019L0882

Employers Resources: Adapting the Workplace. (2020, febrero 24). Lansing State Journal. https://eu.lansingstatejournal.com/story/money/careers/2020/02/24/employers-resources-adapting-workplace/111340878/

Ethics Guidelines for Trustworthy AI. (s. f.). European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Guide to Algorithmic Auditing
https://eticas.tech/guide-to-algorithmic-auditing

GAIA-X. (s. f.). https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html

HireVue. (2023, noviembre). https://www.hirevue.com/

People with Disability Face Much Greater Risk of Violence than People Without Disability. (s. f.). Disability Royal Commission. https://disability.royalcommission.gov.au/news-and-media/media-releases/people-disability-face-much-greater-risk-violence-people-without-disability

Plug and Pray? A Disability Perspective on AI, Automated Decision Making, and Emerging Technologies. (s. f.). G3ict. https://g3ict.org/publication/plug-and-pray-a-disability-perspective-on-ai-automated-decision-making-and-emerging-tech

Policy and Investment Recommendations for Trustworthy Artificial Intelligence. (s. f.). European Commission. https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence

Regulating AI to Protect the Consumer. (s. f.). BEUC. https://www.beuc.eu/sites/default/files/publications/beuc-x-2021-088_regulating_ai_to_protect_the_consumer.pdf

Report: Challenging the Use of Algorithm-Driven Decision Making in Benefits Determinations Affecting People with Disabilities. (s. f.). Center for Democracy & Technology. https://cdt.org/insights/report-challenging-the-use-of-algorithm-driven-decision-making-in-benefits-determinations-affecting-people-with-disabilities/

S2302. (2019). New York State Senate. https://www.nysenate.gov/legislation/bills/2019/S2302

This Person Does Not Exist. (s. f.). https://this-person-does-not-exist.com/es

7 Apéndices

7.1 Evolución de la regulación sobre IA en la Unión Europea (Ley Europea de IA)

Desde 2018 la Unión Europea empezó a diseñar sus primeros documentos al respecto con el objetivo de diseñar una regulación para el uso de IA responsable. En la Figura 6 se muestra una evolución temporal de las acciones clave realizadas, así como los documentos publicados. En **abril de 2018** la UE publica su intención de llevar a cabo un documento de recomendaciones y recoger en este documento la opinión de personas expertas y los distintos actores involucrados en el desarrollo y uso de la IA (empresas, instituciones, gobiernos, sociedad civil, etc.). **En junio de 2018**, se nombra un grupo de personas expertas HLEG (*High Level Expert Group*) [32]formado por 52 personas de todos los países y representando también a las distintas partes involucradas en el desarrollo de la IA. Este grupo de alto nivel tendrá que realizar una serie de recomendaciones, así como definir la hoja de ruta necesaria para poder completar una regulación adecuada. Además, este grupo HLEG contribuye a formar una comunidad más amplia con la creación de la Alianza Europea de IA (*European AI Alliance*) que está formada por más de 4000 miembros de todas las partes de la sociedad, y que se compromete de manera consultiva a proporcionar sugerencias de mejora al Grupo de Alto Nivel.

[32] https://ec.europa.eu/futurium/en/european-ai-alliance/ai-hleg-steering-group-european-ai-alliance.html

Figura 18. Hitos de la IA (fuente- "Manual de ética aplicada en IA")

Descripción figura 18. La imagen representa una línea de tiempo desde el año 2018 hasta el 2021, indicando las fases que se han dado cada mes, dentro de cada año, dentro de la evolución de la regulación sobre IA en la Unión Europea.

En **diciembre de 2018** se lanza una consulta pública, para obtener sugerencias y aportaciones de las distintas partes involucradas, denominada IA hecha en Europa (*AI made in Europe*). Con ella se trataba de dar respuesta a los retos éticos y sociales, así como a resolver la necesidad de establecer un marco regulatorio. Este plan supuso una serie de acciones a escala de la Unión Europea:

- Necesidad de desarrollar capacidades europeas de IA para experimentación y testeo.
- Diseñar e implantar programas de aprendizaje para preparar a la sociedad europea para el uso de la IA.
- Conseguir que las administraciones públicas sean pioneras en el uso e implantación de la IA.
- Implantar guías éticas basadas en la recomendación de las personas expertas del Grupo de Alto Nivel.
- Revisar los marcos legales europeos para adaptarlos a los retos específicos de IA
- Reforzar el desarrollo y la inversión de la IA para que su uso contribuya al desarrollo económico de Europa.

En **abril de 2019** se publica la primera guía ética para una inteligencia artificial confiable[33] ,donde se detallan los cuatro principios éticos fundamentales así como los siete requisitos para una IA confiable, como se refleja en la Figura 19, incluida en dicho documento.

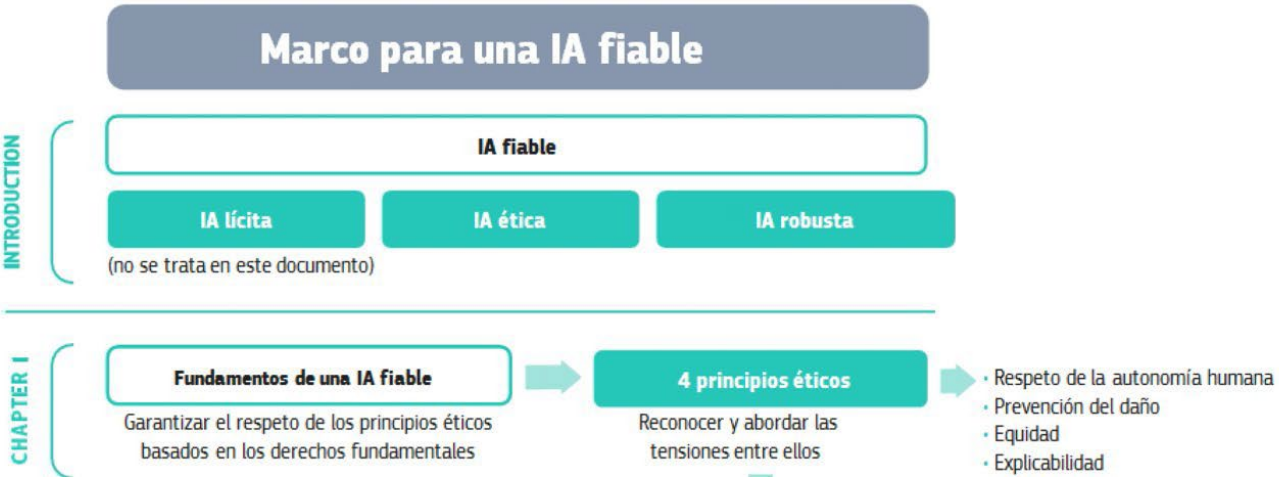


Figura 19. Marco para una IA fiable (fuente: regulación europea)

Descripción figura 19. La imagen representa las directrices establecidas dentro de la primera guía ética para una inteligencia artificial confiable. Dentro del primer capítulo se explican los principios éticos que deben garantizarse: respeto a la autonomía humana, prevención del daño, equidad y explicabilidad.

En **junio de 2019** se publican las recomendaciones para conseguir una IA confiable en Europa[34]. El documento se centra básicamente en 33 recomendaciones que se fundamentan en inclusión, sostenibilidad y competitividad. A continuación, se resumen los puntos fundamentales en los que se centran:

- Asegurar las capacidades de investigación en Europa.
- Transformar el sector privado y usar el sector público como catalizador de un crecimiento innovador.

[33] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[34] https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence.

- Capacitar y proteger a las personas en el entorno en el que viven.
- Impulsar el diseño e implantación de unos datos e infraestructuras comunes para la IA.
- Diseñar la educación para el aprendizaje de la IA como base fundamental y a cualquier nivel.
- Proporcionar la inversión adecuada para conseguir estos objetivos.

Seguidamente a estas acciones se publica en febrero de 2020 el Libro Blanco de la Comisión Europea[35]. Esta publicación se compromete por un lado a mejorar el día a día de la ciudadanía con la IA, pero respetando sus derechos. Además, es un documento consultivo que supone una base para las aportaciones de la sociedad civil, las autoridades públicas, las instituciones académicas, las empresas u organizaciones y las personas físicas. El documento se centra en dos pilares fundamentales como son el establecimiento de **ecosistemas de excelencia y ecosistemas de confianza**. Con el ecosistema de excelencia se pretende establecer esfuerzos conjuntos entre los organismos privados y públicos para que la IA pueda ser implantada de manera ética, y establece puntos clave de colaboración como la investigación, la educación o la creación del talento adecuado. En cuanto a lo que define como ecosistemas de confianza, se refiere a establecer un marco normativo para los miembros de la Unión Europea que genere la confianza suficiente y facilite a las personas encargadas de desarrollar e implantar soluciones de IA la capacidad de diseñar una IA responsable. Un ejemplo de estos ecosistemas es GAIXA-X[36],que ha sido creada para poder disponer de una infraestructura común europea de datos. La plataforma fue fundada por 22 empresas europeas, actualmente cuenta con 850 miembros, de 425 organizaciones diferentes y está trabajando ya en más de 60 casos de uso en la industria.

[35] https://commission.europa.eu/document/d2ec4039-c5be-423a-81ef-b9e44e79825b_es

[36] https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html.

En **julio de 2020** se publica la Lista de evaluación de la IA confiable (ALTAI, *Assessment List of Trustworthy AI*). [37]Este cuestionario trata de proporcionar guías de autoevaluación y de cumplimiento de los 7 requerimientos para una IA confiable. Se desarrolló como resultado de las recomendaciones del Grupo de Alto Nivel, así como con la inclusión de las sugerencias de mejora proporcionadas por las distintas partes interesadas europeas.

En **abril de 2021,** después de tres años de trabajo conjunto, se publica la Propuesta de REGLAMENTO DEL PARLAMENTO EUROPEO Y DEL CONSEJO POR EL QUE SE ESTABLECEN NORMAS ARMONIZADAS EN MATERIA DE INTELIGENCIA ARTIFICIAL (LEY DE INTELIGENCIA ARTIFICIAL) Y SE MODIFICAN DETERMINADOS ACTOS LEGISLATIVOS DE LA UNIÓN [38]. La propuesta incluye un marco regulatorio con cuatro objetivos clave:

- Proporcionar un mercado único para el desarrollo y utilización de la IA
- Mejorar la aplicabilidad de la regulación existente para IA así como su gobernanza
- Garantizar la seguridad jurídica para poder innovar y desarrollar IA
- Poder garantizar que los sistemas de IA usados en Europa son seguros y respetan la legislación vigente.

Se evaluaron distintos enfoques para llevar a cabo este marco regulatorio y se decidió abordar como un enfoque horizontal orientado a riesgos, se puede ver un resumen en la Figura 20. Este enfoque a riesgos tiene la ventaja de permitir una mejor actualización con el paso del tiempo además de proporcionar la capacidad de ajustarse mejor a las normativas actuales de la Unión Europea.

[37] https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

[38] https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex%3A52021PC0206

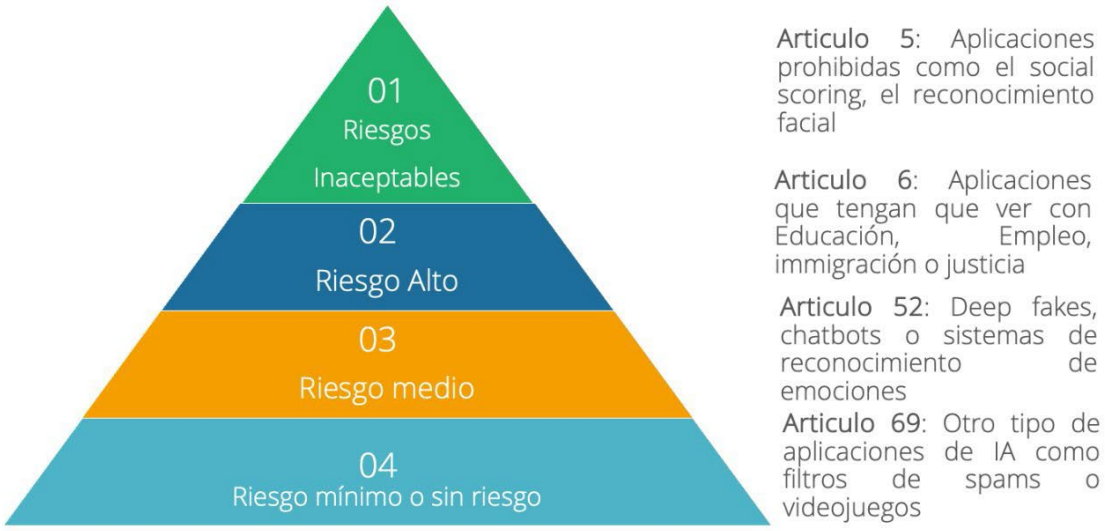


Figura 20. Resumen de la última regulación europea.

Descripción figura 20. La imagen representa una pirámide de riesgos que pueden darse dentro de las aplicaciones de IA, identificando cada nivel con un color. En la parte más elevada de la pirámide se sitúan los riesgos inaceptables, seguido de riesgo alto, riesgo medio y riesgo mínimo o sin riesgo.

Desde la publicación de esta propuesta en 2021 se ha ido evolucionando en la modificación del texto, con las sugerencias de mejora de los Estados y el propio Parlamento Europeo. Actualmente el texto actualizado a junio de 2022[39] por el Parlamento Europeo incluye las últimas modificaciones, junto con las enmiendas realizadas. Asimismo, desde noviembre de 2022, se ha incluido una modificación a la definición de los sistemas de IA, con la incorporación de la IA generativa. La IA generativa se ha incluido dentro de lo que se llaman modelos fundacionales, que abarcan los modelos entrenados con grandes cantidades de datos, diseñados para generar una salida, que se puede usar de muy distintas maneras. La IA generativa es un tipo específico de modelos fundacionales que genera texto, video, audio o imágenes.

[39] https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_ES.html

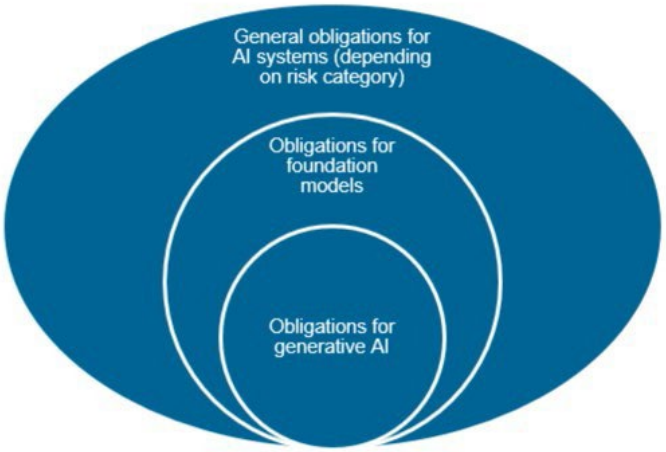


Figura 21. Modelos fundacionales e IA generativa.

Descripción figura 21. La imagen representa un diagrama de Venn apilado, formado por tres círculos, situados uno dentro de otro. En el circulo exterior se sitúan las obligaciones generales para los sistemas de IA (dependiendo de la categoría de riesgo), en su interior se sitúan las obligaciones para modelos fundacionales, y en el circulo interno, las obligaciones para la IA generativa.

En la Figura 21 se puede ver la relación entre los modelos fundacionales y la IA generativa, así como los requerimientos específicos que tienen que cumplir.

En concreto, respecto a los modelos fundacionales se les imponen obligaciones especiales a las empresas proveedoras de la IA generativa:

- Entrenar, desarrollar y diseñar el sistema de forma que existan salvaguardas contra la generación de contenidos que infrinjan la legislación de la UE
- Documentar y ofrecer públicamente un resumen detallado del uso de datos de entrenamiento protegido por derechos de autor.
- Cumplir obligaciones de transparencia más estrictas, la transparencia va muy enfocada a tratar de evitar contenido que pueda manipular o contenido no veraz como “deepfakes”.

En cuanto a los modelos fundacionales en general, las empresas proveedoras estarán obligadas a realizar una serie de acciones antes de poner los modelos en el mercado:

- Poder demostrar cómo ha mitigado los riesgos razonablemente previsibles para la salud, la seguridad, los derechos fundamentales, el medio ambiente, la democracia y el Estado de Derecho.

- Usar únicamente conjuntos de datos que garantice que los conjuntos de datos son adecuados e imparciales.
- Diseñar, desarrollar y probar el modelo base de forma que se garanticen el rendimiento, la previsibilidad, la interpretabilidad, la corrección, la seguridad y la ciberseguridad durante todo su ciclo de vida.
- Desarrollar el modelo de implantación utilizando normas para reducir o recortar el uso de energía, recursos y residuos.
- Tener disponible documentación técnica e instrucciones inteligibles para el modelo de implantación.
- Conservar la documentación técnica a disposición de las autoridades competentes durante un periodo de diez años a partir de la fecha de introducción en el mercado.
- Establecer un sistema de gestión de la calidad para garantizar y documentar el cumplimiento de la Ley de IA.

Actualmente la regulación está siendo aún discutida por los órganos responsables en la Unión Europea y se prevé que entre en vigor a finales de 2024 o principios de 2025.

7.2 Lista de comprobación de criterios de auditabilidad

A continuación, se presenta una lista de comprobación, que puede ser adaptada en cada caso, destinada a guiar a las organizaciones que desean realizar una auditoría ética de un sistema de inteligencia artificial. El objetivo es verificar la disponibilidad de información para valorar las posibilidades reales de llevar a cabo la auditoria con éxito:

1. **Datos de contacto**

   - ❒ Proporcionar una lista de contactos con las áreas de conocimiento, roles y responsabilidades respectivas dentro de la organización.
   - ❒ Disponibilidad de datos de contacto de personas u organizaciones involucradas en el diseño, desarrollo e implantación.

2. **Detalles del algoritmo o sistema**:

   - ❒ Persona física o jurídica propietaria del algoritmo o sistema.
   - ❒ Fecha de lanzamiento y versión actual.
   - ❒ Tipo de licencia del algoritmo o sistema, y condiciones de uso, comerciales y contractuales.

3. **Objeivo y Uso del Algoritmo**:

   - ❒ Descripción de la aplicación y caso de uso.
   - ❒ Evaluación del contexto y de la normativa y regulación aplicable.
   - ❒ Objetivos e indicadores de éxito.
   - ❒ Categorización de personas individuales o colectivos influenciados por el algoritmo.
   - ❒ Identificación de colectivos en riesgo de exclusión.

4. **Descripción Técnica**:

   - ❒ Descripción general del algoritmo o sistema y su arquitectura.
   - ❒ Elección del tipo de algoritmo de ML (incluyendo caja negra vs. caja blanca).

- ❒ Información de entrenamiento y modo de operación.
- ❒ Elección de métricas de rendimiento.
- ❒ Información sobre la API.
- ❒ Referencias adicionales como artículos, sugerencias y citas relacionadas.
- ❒ Medidas adoptadas para asegurar la imparcialidad y eficacia del modelo.
- ❒ Protocolos de seguridad aplicados.
- ❒ Aseguramiento de calidad del modelo.

5. **Acceso al Código**:
   - ❒ Cumplimiento de estándares de calidad.
   - ❒ Detalles sobre el lenguaje de programación.
   - ❒ Herramientas y librerías esenciales para visualización.
   - ❒ Plan de mantenimiento.
   - ❒ Aseguramiento de calidad del código.
6. **Datos:**
   - ❒ Método y herramientas de adquisición de datos.
   - ❒ Calidad de los datos brutos.
   - ❒ Detalles sobre las bases de datos de entrenamiento y evaluación.
   - ❒ Lista de variables utilizadas.
   - ❒ Transformaciones de datos y elección de 'características'.
   - ❒ Representación de grupos y potencial sesgo en los datos brutos.
   - ❒ Datos personales y protección de datos.
7. **Integración en la organización:**
   - ❒ Procesos y actividades vinculadas al sistema.
   - ❒ Referencia al RGPD y otros marcos regulatorios pertinentes.

- ❒ Actualización y monitoreo de datos.
- ❒ Reentrenamiento del modelo.
- ❒ Aseguramiento de calidad a largo plazo.
- ❒ Control de rendimiento en producción.

7.3 Modelo CRISP-DM

Las tres metodologías más relevantes para proyectos alrededor del análisis de datos son KDD, SEMMA y CRISP-DM. Actualmente esta última, CRISP-DM, es la metodología más usada para los proyectos de *machine learning*. Consta de seis fases:

1) Entendimiento del negocio
2) Comprensión de los datos
3) Preparación de los datos
4) Modelado
5) Evaluación
6) Implantación.

La sucesión de fases no es necesariamente rígida, se puede ir de una fase a otra, que es lo que de hecho ocurre en los proyectos reales. Imaginemos que tenemos que predecir el fraude para los clientes de una determinada empresa, seleccionamos los datos que se consideran relevantes para este análisis, en la fase de entendimiento de los datos (*Data Understanding*). Una vez seleccionados estos datos, se preparan de manera adecuada para poder seleccionar el tipo de algoritmo (limpieza, normalización de datos, etc) en la fase de preparación de datos (*Data Preparation*). Una vez que se tienen los datos preparados comienza la fase de modelado (*Data Modeling*) es decir la búsqueda del método de *machine learning* más adecuado para la resolución de este problema de negocio. Puede ocurrir que con los datos que se han preparado no se encuentren resultados óptimos en ningún modelo y sea necesario volver a cualquiera de las fases anteriores para recopilar esos datos. A esto nos referimos cuando se habla de fases no necesariamente rígidas.

7.3.1 Comprensión del negocio (*Business Understanding*)

La fase inicial se centra en comprender los objetivos y requisitos del proyecto desde una perspectiva empresarial, para lograr transformarlo en un problema de minería de datos con sentido de negocio y crear un plan diseñado para lograr tales objetivos. En esta fase es donde se eligen las tecnologías a utilizar y se definen los criterios de éxito del proyecto.

7.3.2 Entendimiento de los datos (*Data Understanding*)

La fase de comprensión de datos se centra en recopilar los conjuntos de datos que pueden ayudar a lograr los objetivos del proyecto.

Esta fase consta de cuatro etapas básicas:

- **Recopilación de datos**, cuáles son las fuentes necesarias y dónde están.
- **Idenificación de los datos,** cuáles son los datos clave de cada una de esas fuentes.
- **Búsqueda de relaciones entre datos**, cuáles son las relaciones entre esas fuentes de datos
- **Idenificación** de los posibles problemas de calidad de datos

7.3.3 Preparación de los datos (*Data Preparation*)

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final a partir de los datos brutos iniciales. En esta parte se preparan los datos pensando en las técnicas de modelado que se utilizarán posteriormente, dato que en la mayoría de los casos dependiendo de la técnica de modelado se requiere un formateo diferente de los datos.

Esta fase consta de las siguientes etapas:

- Selección de datos, se elige un conjunto de datos basados en los identificados en la fase anterior en base a la completitud, el volumen, formato…etc.
- Limpieza de datos, esta tarea complementa la anterior y requiere mucho tiempo. En casi todos los proyectos de datos suele ser la etapa más larga del proyecto, dado que de la calidad del dato dependerá el resultado. Algunas de las técnicas usadas en esta parte, suele ser normalización, discretización de campos numéricos, tratamiento de nulos o información no disponible, entre otros.
- Estructuración de datos, esta tarea es la que permite crear nuevos atributos o campos a partir de otros de los que ya se disponía, que pueden ayudar a un mejor resultado del modelo.
- Integración de los datos, implica la creación de nuevas tablas, o nuevos datos que puedan ser necesarias como tablas intermedias o tablas de resumen,

- Formateo de los datos, consiste en formatear los valores de los datos que pudieran ser necesarios para la aplicación de modelos posteriores. Puede consistir en cambiar el formato por su tipología (char, integer, date...), o eliminar determinados caracteres especiales.

7.3.4 Modelado de datos (*Modeling*)

En esta fase, se seleccionan y aplican varias técnicas de modelado. Normalmente, existen varias técnicas para el mismo tipo de problema de *machine learning* y generalmente se tiene en cuenta el tipo de datos, el conocimiento de la técnica y el tiempo del que se dispone. Adicionalmente es importante considerar también los costes de ejecución del modelo, generalmente modelos más complejos requieren mayor capacidad de procesamiento que en general suele elevar los costes. El desarrollo del Cloud en los últimos años ha contribuido a reducir estos costes, pero aun así sigue siendo una variable a tener en cuenta para la elección del modelo.

Actualmente, también se tiene en cuenta la explicabilidad de los modelos a seleccionar, y se tiene que decidir en esta etapa qué tipo de equilibrio se quiere obtener entre la complejidad del modelo y la explicabilidad. Normalmente modelos más complejos pueden proporcionar mejores resultados pero su explicabilidad es más reducida. Es decir, la explicabilidad y la complejidad del modelo están relacionadas de manera inversa, como se puede ver en el gráfico, y en muchas ocasiones se decide utilizar un modelo más sencillo por su mayor explicabilidad. No obstante, en los últimos años, han avanzado mucho las técnicas de explicabilidad de los modelos que están permitiendo proporcionar más explicabilidad a modelos más complejos.

A menudo es necesario volver a la fase de preparación de datos, ya que puede surgir una necesidad adicional que no se ha previsto en la fase anterior.

Esta fase tiene cuatro etapas fundamentales:

- Selección de las técnicas de modelado, se suele hacer en base a la experiencia de la persona que diseña, la tipología del problema o la disponibilidad de los tipos de algoritmos. Generalmente es en esta etapa donde se distingue entre modelos supervisados o modelos no supervisados.

- Generación del plan de pruebas, típicamente en los modelos supervisados se separan los conjuntos de datos, en entrenamiento y test. No es el caso de los

modelos no supervisados, en los que lo que se hace es validar el modelo con distintas partes del conjunto de datos.

- Construcción del modelo, una vez obtenidas las métricas adecuadas para la selección del modelo se despliega en producción el más adecuado. En esta etapa también se puede enriquecer el modelo con variables adicionales en base a la experiencia con otros clientes o problemas similares.

- Evaluación de la calidad del modelo, la evaluación se realiza en base a la experiencia, así como la validación del cliente.

7.3.5 Evaluación (*Evaluation*)

La fase de evaluación analiza qué modelo se adapta mejor al negocio y se decide qué hacer a continuación. Siempre hay que tener en cuenta que el modelo se desarrolla con los datos actuales, pero los datos irán cambiando y habrá que evaluar periódicamente si el modelo sigue siendo válido.

Esta fase tiene tres etapas fundamentales:

- Evaluación de los resultados, para comprobar que cumplen los criterios de la empresa.

- Revisión del proceso para establecer mejoras en toda la metodología, así como en las fases iniciales.

- Próximos pasos, probar más iteraciones del modelo, decidir probar otros modelos o la finalización de esta fase.

7.3.6 Implementación (*Deployment*)

La creación del modelo generalmente no es el final del proyecto. Incluso si el propósito del modelo es aumentar el conocimiento de los datos, el conocimiento adquirido deberá organizarse y presentarse de manera que pueda ser utilizado.

Dependiendo de los requisitos, la fase de implantación puede ser tan simple como generar un informe o tan compleja como implantar un proceso de *machine learning* repetible en toda la empresa.

Esta fase consta de cuatro etapas fundamentales:

- Plan de implantación

- Plan de monitorización
- Informe final
- Revisión del proyecto

Es clave entender que esta no es la finalización del proyecto, dado que los datos irán cambiando. Por eso es clave la monitorización continua del modelo, y revisar de manera periódica su comportamiento para asegurarse que sigue siendo válido para el problema de negocio a resolver.

7.4 Métricas para análisis de sesgos

Para explicar el detalle de las métricas que impactan en los sesgos nos hemos basado en este manual. [40]

La matriz de confusión se utiliza para medir cuanto de bueno es el modelo que hemos creado. Sirve para poder tomar decisiones sobre qué quiero optimizar en mi modelo, dependiendo del tipo de aplicación que se esté construyendo. Recordemos que, en el caso de análisis supervisado, lo que tengo son un conjunto de datos etiquetados para cada una de las muestras con los distintos atributos, y la variable objetivo a predecir la tenemos también etiquetada. Como ejemplo podemos usar la variable de contratación, es decir, predecir si una persona candidata va a ser contratada o no. Una vez definida esta variable, se recopilan los datos necesarios para predecir si la persona candidata va a ser contratada o no. En ese conjunto de datos se puede analizar el valor del sesgo en base a cualquier variable incluida en el *dataset,* como podrían ser el género, la edad, etc.

Esto sería lo que llamamos un conjunto de datos etiquetados. Para simplificar se ha incluido una sola variable, y lo que se va hacer para calcular las métricas es comparar cual es la diferencia entre la realidad y la predicción. Es lo que se llama la matriz de confusión, que básicamente es una comparación de los valores reales y predichos.

[40] Olmeda, M. V., & Ibánez, J. C. (2022). *Manual de ética aplicada en Inteligencia Artificial*. Anaya Multimedia.

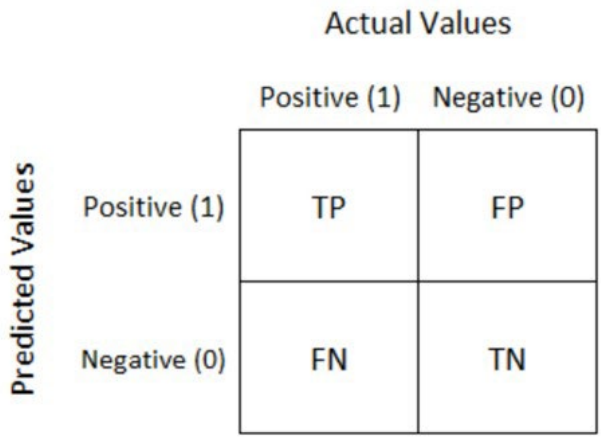


Figura 22. Matriz de confusión del modelo.

Descripción figura 22. La imagen muestra una tabla, con dos columnas y dos filas, correspondiente a una matriz de confusión del modelo creado.

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** $PPV = \frac{TP}{TP+FP}$ $TPR = \frac{TP}{TP+FN}$	**False Positive (FP)** $FDR = \frac{FP}{TP+FP}$ $FPR = \frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** $FOR = \frac{FN}{TN+FN}$ $FNR = \frac{FN}{TP+FN}$	**True Negative (TN)** $NPV = \frac{TN}{TN+FN}$ $TNR = \frac{TN}{TN+FP}$

Figura 23. Tabla de contingencia de las métricas dentro de la Matriz de Confusión.

Descripción figura 23. La imagen muestra una tabla, con dos columnas y dos filas, con las formulas para calcular las métricas dentro de una matriz de confusión.

TP(True positive): Predicción correcta de que la persona ha sido contratada
TN(True negative): Predicción correcta de que la persona no ha sido contratada
FP(False positive): Predicción falsa, se contrató a la persona cuando no se tenía que haber hecho, esto es el error que se denomina Tipo I
FN(False negative): Predicción falsa, no se contrató a la persona cuando sí se tenía que haber hecho, y este es el error que se denomina Tipo II

A continuación, se describen las fórmulas más usadas y se hace un ejemplo para el cálculo de las más usadas.

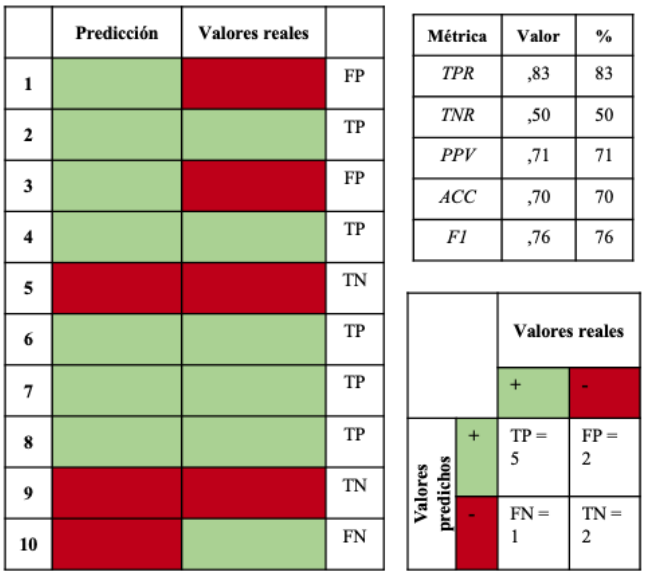
Métrica	Fórmula	Nombre completo	Descripción
TPR	$\frac{TTTT}{TTTT + FFFF}$	**Tasa de verdaderos positivos (*True positive rate, Recall, Sensitivity)***	Este valor junto con el siguiente nos ayuda a medir como de bueno es el algoritmo discriminando entre casos positivos y negativos. En este caso es el **porcentaje de positivos detectado correctamente** por el algoritmo.
TNR	$\frac{TTFF}{TTFF + FFTT}$	**Tasa de verdaderos negativos (*True negative rate, Especificity)***	Se llama especificidad , al contrario que el anterior, es el **porcentaje de negativos detectado correctamente por el algoritmo**
ACC	$\frac{TTTT + TTFF}{TTTT + FFFF + FFTT + TTFF}$	**Exactitud (*Accuracy*)**	Es la **exactitud** y trata de medir lo cerca que ha estado una medición de su valor verdadero, es decir, es la cantidad de predicciones positivas que fueron correctas.
PPV	$\frac{TTTT}{TTTT + FFTT}$	**Porcentaje de casos positivos (*Positive predictive value, Precision)***	Precisión, que lo que indica es el **porcentaje de casos positivos**
F1	$22 * (\frac{\cdots\cdots}{TTTPP + TTTTT})$	**F1 score**	Esta métrica nos resume la precisión y la sensibilidad en una sola métrica y es muy útil cuando las clases son desiguales.

Figura 24. Tabla de las métricas más usadas dentro de la Matriz de Confusión.

Descripción figura 24. La imagen muestra una tabla, con cuatro columnas: Métrica, fórmula, nombre completo y descripción.

Ejemplo de contratación

Suponemos que tenemos una lista de 10 valores sobre la variable que queremos predecir que es contratación de personas. Esta variable es la etiqueta de los datos y tiene el valor de contratada (en verde) y no contratada (en rojo). Para hacer la matriz de confusión necesitamos comparar los valores predichos por el algoritmo, con los valores reales. Con esto vamos calculando los valores de la matriz de confusión, una vez obtenida la matriz de confusión se pueden ir obteniendo el resto de las métricas.



	Predicción	Valores reales	
1			FP
2			TP
3			FP
4			TP
5			TN
6			TP
7			TP
8			TP
9			TN
10			FN

Métrica	Valor	%
TPR	,83	83
TNR	,50	50
PPV	,71	71
ACC	,70	70
F1	,76	76

		Valores reales	
		+	-
Valores predichos	+	TP = 5	FP = 2
	-	FN = 1	TN = 2

Figura 25. Matriz de confusión sobre los valores predichos para la variable contratación de personas

Descripción figura 25. La imagen muestra tres tablas, una de ellas es un listado de los 10 valores sobre la variable que queremos predecir, contiene una columna con el número de orden, predicción (color rojo si toma valor negativo o verde si toma valor positivo) y valores reales (color rojo si toma valor negativo o verde si toma valor positivo). La segunda tabla está dividida en tres columnas, la métrica calculada, el valor calculado y el valor calculado expresado en porcentaje. La tercera tabla muestra los valores obtenidos dentro de la matriz de confusión.

Además de las métricas explicadas con el ejemplo, es importante entender también el funcionamiento de la curva ROC, que se usa además de la exactitud (accuracy) cuando los datos no están balanceados y que va a permitir tomar decisiones sobre la idoneidad del modelo. Además de la curva ROC para representar TPR y FPR, se utiliza mucho la métrica AUC.

CURVA ROC(*Receiver Operating Characteristic curve*) y AREA AUC (Area under the curve)

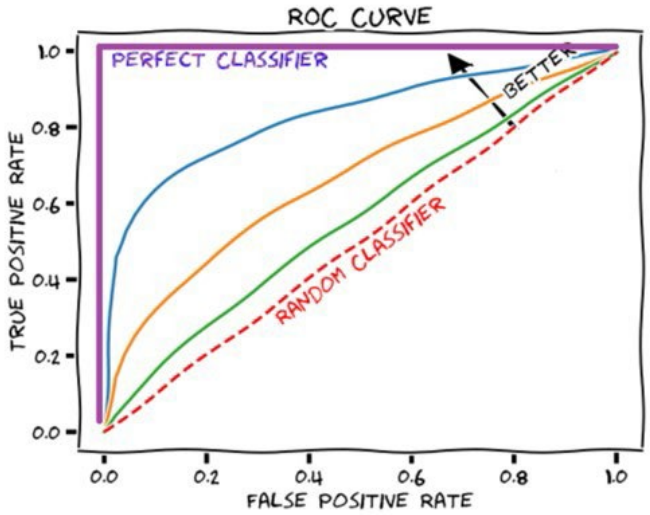


Figura 26. Curva de ROC

Descripción figura 26. La imagen muestra una gráfica de la Curva de ROC. En el eje de abscisas se muestra el ratio de falso positivo, y en el eje de ordenadas se muestra el ratio verdadero positivo.

La curva ROC nos va a medir lo adecuado que puede ser el modelo clasificando los verdaderos positivos y falsos positivos teniendo en cuenta diferentes umbrales y diferentes clasificadores, digamos que es la manera sencilla de representar los valores de distintas matrices de confusión para distintos algoritmos y poder compararlos entre ellos. Si vamos estableciendo distintos umbrales lo que vamos a ir obteniendo son valores FPR y TPR que podemos ir marcando en la curva ROC y que nos va a permitir entender como está clasificando nuestro algoritmo. En la curva ROC las distintas líneas representan distintos algoritmos que puedo ir comparando, y como van cambiando el TPR y FPR dependiendo del umbral que establezcamos. La parte AUC es el área que queda por debajo de la curva que vamos obteniendo, el valor de AUC estará siempre entre 0.5 y 1. Cuanto más cercano a 1 sea el valor de AUC , indica que mejor es el clasificador.

Estas métricas y gráficos ayudan a tomar decisiones sobre la idoneidad del modelo, y para saber que modelo finalmente tendremos que seleccionar habrá que responder a preguntas como ¿el modelo tolera mejor los falsos positivos o los falsos negativos?, porque así podemos decidir si tenemos que seguir entrenando el modelo o nos vale con la tasa de errores que tenemos Habrá que responder a preguntas como ¿es mejor contratar a alguien cuando no parecía ser el candidato adecuado? o ¿es mejor no contratarlo cuando si lo teníamos que haber hecho? Siempre son los objetivos de negocio o el tipo de aplicación los que van a marcar la

pauta. Imaginemos que no es una aplicación de contratación, sino de detección de tumores o de conducción autónoma donde unos errores pueden tener más relevancia que otros.