7

Empresas desarrollando cumplimiento de requisitos

Guía 7. Datos y gobernanza del dato

Reglamento Europeo de Inteligencia Artificial

Financiado por la Unión Europea
NextGenerationEU

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

Plan de Recuperación, Transformación y Resiliencia

España | digital 2026

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

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

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

Entre las referencias técnicas relevantes actualmente en desarrollo y aplicables, destacan las normas **ISO/IEC 5259-1 a 5259-5 "Artificial intelligence – Data quality for analytics and machine learning (ML)" , prEN 18229 "AI Trustworthiness Framework", prEN XXX Quality and governance of datasets in AI y prEN XXX Concepts, measures and requirements for managing bias in AI Systems,** que servirán de base para el establecimiento de un marco integral de gobernanza, calidad del dato, gestión del ciclo de vida y confianza en los sistemas de inteligencia artificial.

Fecha de versión: 10 de diciembre de 2025

Contenido general

1. Preámbulo 5
2. Introducción 7
3. Reglamento de Inteligencia Artificial 8
4. Gestión de los datos 13
5. Otros elementos a considerar 36
6. Documentación técnica 42
7. Cuestionario de autoevaluación 44
8. Anexos 45
9. Referencias, estándares y normas 76

Índice detallado

1. Preámbulo 5
1.1 Objetivo del documento 5
1.2 ¿Cómo leer esta guía? 5
1.3 ¿A quién está dirigido? 5
1.4 Casos de uso y ejemplos dispuestos a lo largo de la guía 6
2. Introducción 7
2.1 ¿Qué es el gobierno de los datos? 7
2.2 ¿Qué elementos debo implantar y cómo debo hacerlo para desarrollar un adecuado gobierno del dato? 7
3. Reglamento de Inteligencia Artificial 8
3.1 Análisis previo y relación de los artículos 8
3.2 Contenido de los artículos en el Reglamento de IA 9
3.3 Correspondencia del articulado con los apartados de la guía 11
4. Gestión de los datos 13
4.1 Requisitos de información 13
4.2 Recopilación de los datos 13
4.3 Preparación de los datos 15
4.3.1 Medición y mejora de la calidad de los datos 15
4.3.2 Transformación de los datos 23
4.3.3 Agregación de los datos 23
4.3.4 Muestreo de los datos 24
4.3.5 Creación y selección de características 26
4.3.6 Enriquecimiento de los datos 27
4.3.7 Etiquetado de los datos 27
4.3.8 Análisis de sesgos en los datos 29
4.4 Disposición de los datos 31
4.5 Eliminación de los datos 33
5. Otros elementos a considerar 36
5.1 Tratamiento de las categorías especiales de datos personales 36
5.1.1 Las categorías especiales de datos personales en el Reglamento Europeo de la IA y su tratamiento 36
5.1.2 La detección y/o corrección de sesgos de categorías especiales de datos en el marco del sandbox 40
6. Documentación técnica 42

7. Cuestionario de autoevaluación 44
8. Anexos 45
8.1 ANEXO A – Métodos de recopilación de datos 45
8.2 ANEXO B – Calidad del dato 46
8.2.1 ANEXO B.1 – Dimensiones de la calidad de los datos 46
8.2.2 ANEXO B.2 – Controles de calidad de los datos 51
8.3 ANEXO C – Sesgos 57
8.3.1 ANEXO C.1 – Fuentes de sesgo 57
8.3.2 ANEXO C.2 – Técnicas de evaluación del sesgo 60
8.3.3 ANEXO C.3 – Medidas de tratamiento del sesgo 63
9. Referencias, estándares y normas 76

1. Preámbulo

1.1 Objetivo del documento

En esta guía se presentan las medidas que servirán a proveedores y responsables de despliegue para dar cumplimiento a los requisitos del artículo 10 "Datos y gobernanza de datos" del Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo (Reglamento Europeo de IA). Este artículo establece los requisitos de gobernanza de datos que deberá incorporar todo sistema de IA de alto riesgo (HRAIS) y ciertos sistemas de IA de propósito general (Capítulo V, Modelos de IA de propósito general En este sentido, a lo largo de la guía nos referiremos, generalmente, a estos sistemas como "sistema de IA" con el objetivo de simplificar el discurso.

1.2 ¿Cómo leer esta guía?

Esta sección se ha elaborado para **tratar de acompañar al lector** en la lectura de la guía y ayudarle a alcanzar una **comprensión** de los contenidos más **ágil y eficaz**.

Si el lector no tiene experiencia en materia de gobernanza de datos, se le recomienda **como mínimo una primera lectura completa de la guía**, prestando especial atención a los capítulos clave inherentes al artículo del Reglamento Europeo de la IA (capítulos 4 y 5.1). Si el lector tiene amplia experiencia en materia de gobernanza de datos, deberá focalizarse en los **capítulos imprescindibles inherentes al artículo del** Reglamento Europeo de la IA (capítulos 4 y 5.1), sin perjuicio de que se le recomienda igualmente una lectura completa de la guía.

Adicionalmente, se dispone una serie de Anexos cuyo contenido es introducido y referenciado en las correspondientes secciones de la guía. Cabe destacar la relevancia e importancia de estos Anexos, que son catálogos de elementos indispensables para completar y dar forma al contenido de la guía (métodos de recopilación de datos, dimensiones y controles de calidad de los datos, fuentes de sesgo comunes, técnicas de evaluación y medidas de tratamiento de los sesgos).

1.3 ¿A quién está dirigido?

Los requisitos descritos en el artículo 10 "*Datos y gobernanza de datos*" del Reglamento IA están orientados fundamentalmente al desarrollo del sistema, realizado por el proveedor. En dicho artículo, no se especifican requisitos para aquel que hace uso del sistema, es decir, el responsable del despliegue. En caso de que el responsable del despliegue, en una situación determinada, participase en el desarrollo del sistema, o fuera responsable de los datos de entrenamiento o prueba, éste debería aplicar las medidas desarrolladas para el proveedor. A pesar de todo, se interpela al responsable del despliegue a realizar un uso responsable y ético del sistema en todo momento.

No obstante, atendiendo a lo expuesto en el *artículo 26* "Obligaciones de los responsables del despliegue de sistemas de IA de alto riesgo", *en su punto 4,* en la medida en que el responsable del despliegue ejerza un control sobre los datos de entrada, dicho responsable del despliegue se asegurará de que los datos de entrada sean pertinentes a la vista de la finalidad prevista del sistema de IA de alto riesgo. Por ello, en este caso el responsable del despliegue deberá cumplir con las mismas medidas que el proveedor.

En este contexto, las medidas detalladas a lo largo de esta guía, tanto organizativas como técnicas, están orientadas a servir como guía para el proveedor, salvo en los casos mencionados anteriormente en que afectan también al responsable del despliegue.

1.4 Casos de uso y ejemplos dispuestos a lo largo de la guía

Con el fin de **facilitar** la **comprensión de la guía**, se incorporan en ésta **diferentes ejemplos** que pretenden servir como **referencia** para la adecuación de los HRAIS conforme a los requisitos de datos y gobernanza de datos del Reglamento.

Estos ejemplos se desarrollan en base a los **casos de uso** descritos en la **Guía práctica y ejemplos para entender el Reglamento IA.**

En concreto, los casos de uso utilizados en la elaboración de esta guía son el sistema de IA para la promoción de empleados, la bomba de insulina inteligente y el sistema de IA de reconocimiento biométrico para registrar el tiempo y la asistencia al trabajo.

Finalmente, cabe destacar que siempre que se ponga un ejemplo, se hará de manera ilustrativa. Tanto el proveedor como el responsable del despliegue han de considerar la aplicación de todas las medidas indicadas en esta guía, según corresponda. Además, los ejemplos expuestos son específicos de los casos de uso. Esto implica que las propuestas son específicas para los modelos considerados como ejemplo, y no una solución general para otros tipos de modelo, o incluso modelos de la misma tipología. Cada organización deberá, acorde a esta guía, establecer las medidas oportunas para su tipo de sistema de IA y su finalidad prevista.

2. Introducción

2.1 ¿Qué es el gobierno de los datos?

En el contexto de la IA, la gobernanza de los datos es el conjunto de elementos (políticas, procedimientos, procesos, normas, etc.) que se implementan para garantizar que los datos utilizados en el entrenamiento, validación y prueba de los sistemas de IA son adecuados, pertinentes, suficientemente representativos y cumplen los requisitos de calidad y completitud establecidos.

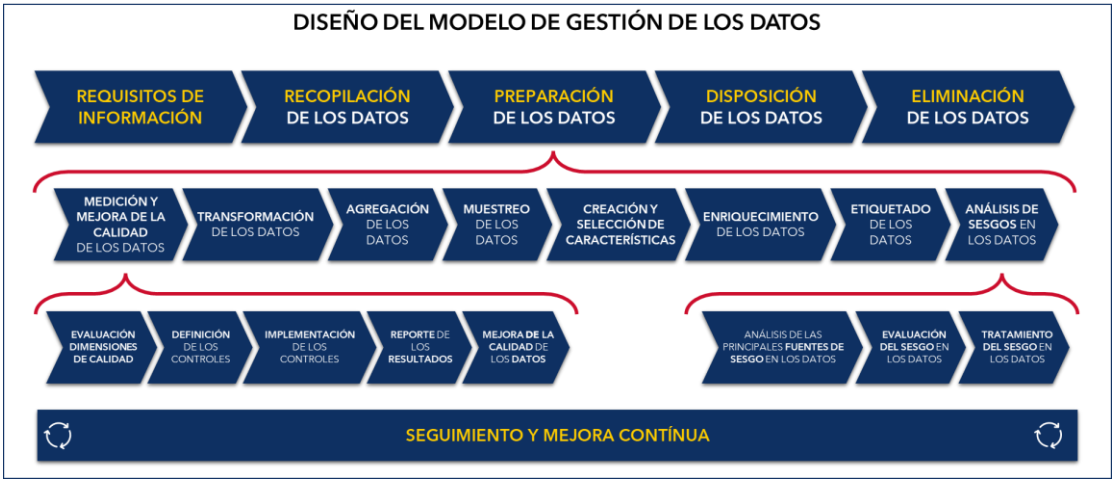
El adecuado gobierno de los datos es esencial para que los sistemas de IA funcionen adecuadamente y conforme a su finalidad prevista. La falta de una gobernanza de los datos puede conducir a resultados sesgados o inexactos, lo que podría llegar a materializarse como un riesgo para la salud, seguridad o derechos fundamentales de los usuarios.

2.2 ¿Qué elementos debo implantar y cómo debo hacerlo para desarrollar un adecuado gobierno del dato?

En esta sección se presentan las fases del modelo propuesto para poder implantar un adecuado gobierno del dato. Este modelo pretende cubrir todos los elementos exigidos en el artículo 10 del Reglamento IA.

No obstante, se ha incorporado algún elemento adicional a lo explícitamente dispuesto en el artículo por la propia completitud y coherencia de los elementos esenciales de un adecuado gobierno del dato (por ejemplo, en el artículo, dentro de los procesos de tratamiento, no se menciona la transformación o el muestreo de los datos).

De una forma general se puede ver el proceso como:



Organizado por fases que se desarrollarán en detalle en el apartado 4 de la presente guía.

3. Reglamento de Inteligencia Artificial

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

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

3.1 Análisis previo y relación de los artículos

Con el objetivo de facilitar la implantación de las medidas propuestas para dar cumplimiento a los requisitos que establece el artículo, se ha reestructurado el contenido de este de manera que algunos puntos que parecen más genéricos (como el 2.h y el 3) se encuentren imbricados en otros más específicos (por ejemplo, el punto 2.e) que es donde tendrían una aplicación más concreta:

1. Los **conjuntos de datos de entrenamiento, validación y prueba** se someterán a las siguientes **prácticas** de **gobernanza y gestión de datos**:
    a) Elección de un **diseño adecuado**.
    b) **Procesos de recopilación** de datos.
    c) **Operaciones** de **tratamiento** para la **preparación** de los datos (*la anotación, el etiquetado, la depuración, el enriquecimiento y la agregación*).
    d) **Formulación de los supuestos** pertinentes con respecto a la **información los datos miden y representan.**
    e) **Evaluación previa de la disponibilidad, la cantidad y la adecuación** de los conjuntos de datos necesarios:
        i. Serán pertinentes y representativos.
        ii. Estarán completos.
        iii. Tendrán las propiedades estadísticas adecuadas.
        iv. Tendrán en cuenta, en función de su finalidad prevista, las características o elementos particulares del contexto geográfico, conductual o funcional en el que se pretende utilizar el HRAIS.
    f) **Análisis de sesgos** (*atendiendo aquellos que puedan afectar a la salud y la seguridad de las personas o dar lugar a algún tipo de discriminación prohibida por el Derecho de la Unión).*
    g) **Detección y remediación** de **lagunas o deficiencias** en los datos:

i. Carecerán de errores.

2. En el caso de que los **datos de salida** vayan a ser datos de **entrada de otro modelo**, deberían ser sometidos a las **prácticas de gobernanza y gestión de datos** mencionadas.
3. En la medida en que sea **estrictamente necesario** para garantizar la **vigilancia, la detección y la corrección de los sesgos**, se podrán tratar las **categorías especiales de datos personales, ofreciendo siempre las salvaguardias adecuadas** para los **derechos y las libertades** fundamentales de las personas.

A partir de este análisis, en la sección 4 plantearemos un enfoque para poder abordar cada una de estas tareas de forma adecuada.

3.2 Contenido de los artículos en el Reglamento de IA

AI Act

Art.10 – Datos y gobernanza de datos

1. Los **sistemas de IA de alto riesgo** que utilizan técnicas que implican el **entrenamiento** de **modelos** de IA con **datos** se desarrollarán a partir de conjuntos de **datos de entrenamiento, validación y prueba** que cumplan los **criterios de calidad** a que se refieren los **apartados 2 a 5** siempre que se utilicen dichos conjuntos de datos.

2. Los **conjuntos de datos de entrenamiento, validación y prueba** se someterán a **prácticas** de **gobernanza y gestión de datos** adecuadas para la finalidad prevista del sistema de IA de alto riesgo. Dichas prácticas se centrarán, en particular, en lo siguiente:

a) las decisiones pertinentes relativas al **diseño**;
b) los **procesos de recogida de datos** y el **origen de los datos** y, en el caso de los datos personales, la **finalidad** original de la recogida de datos;
c) las **operaciones** de **tratamiento** oportunas para la **preparación** de los datos, como la **anotación,** el **etiquetado**, la **depuración**, la **actualización**, el **enriquecimiento** y la **agregación**;
d) la **formulación de supuestos**, en particular en lo que respecta a la **información** que se supone que **miden y representan los datos**;
e) una **evaluación de la disponibilidad, la cantidad y la adecuación** de los conjuntos de datos necesarios;
f) el **examen** atendiendo a **posibles sesgos** que puedan afectar a la **salud y la seguridad de las personas,** afectar negativamente a los **derechos fundamentale**s o dar lugar a algún tipo de **discriminación prohibida** por el Derecho de la Unión, especialmente cuando las

salidas de datos influyan en las informaciones de entrada de futuras operaciones;

g) **medidas** adecuadas para **detectar, prevenir y mitigar** posibles **sesgos** detectados con arreglo a la letra f);

h) la detección de **lagunas o deficiencias** pertinentes en los datos que impidan el cumplimiento del presente Reglamento, y la **forma de subsanarlas.**

3. Los conjuntos de **datos de entrenamiento, validación y prueba** serán **pertinentes**, suficientemente **representativos** y, en la mayor medida posible, **carecerán de errores** y estarán **completos** en vista de su finalidad prevista. Asimismo, tendrán las **propiedades estadísticas adecuadas**, por ejemplo, cuando proceda, en lo que respecta a las personas o los colectivos de personas en relación con los cuales está previsto que se utilice el sistema de IA de alto riesgo. Los conjuntos de datos podrán reunir esas características para cada conjunto de datos individualmente o para una combinación de estos.

4. Los **conjuntos de datos** tendrán en cuenta, en la medida necesaria para la **finalidad prevista,** las **características** o elementos particulares del entorno **geográfico, contextual, conductual o funcional** específico en el que está previsto que se utilice el sistema de IA de alto riesgo.

5. En la medida en que sea estrictamente necesario para garantizar la **detección y corrección** de los **sesgos** asociados a los sistemas de IA de alto riesgo de conformidad con lo dispuesto en el apartado 2, letras f) y g), del presente artículo, los proveedores de dichos sistemas podrán tratar excepcionalmente las categorías **especiales de datos personales** siempre que ofrezcan las **garantías adecuadas** en relación con los **derechos y las libertades** fundamentales de las personas físicas. Además de las disposiciones establecidas en los Reglamentos (UE) 2016/679 y (UE) 2018/1725 y la Directiva (UE) 2016/680, para que se produzca dicho tratamiento deben cumplirse todas las condiciones siguientes:

a) que el tratamiento de otros datos, como los **sintéticos o los anonimizados**, **no permita** efectuar de forma efectiva la **detección y corrección de sesgos;**

b) que las **categorías especiales de datos personales** estén sujetas a **limitaciones técnicas** relativas a la **reutilización** de los datos personales y a medidas punteras en materia de **seguridad y protección** de la **intimidad**, incluida la **seudonimización**;

c) que las **categorías especiales de datos personales** estén sujetas a medidas para garantizar que los datos personales tratados estén **asegurados**, **protegidos** y **sujetos a garantías adecuadas**, incluidos **controles estrictos** y **documentación del acceso**, a fin de **evitar** el **uso indebido** y garantizar que **solo las personas autorizadas** tengan

acceso a dichos datos personales con **obligaciones** de **confidencialidad** adecuadas;
d) que las **categorías especiales de datos personales no** se **transmitan** ni transfieran a terceros y que estos **no puedan acceder** de ningún otro modo a ellos;
e) que las **categorías especiales de datos personales** se **eliminen** una vez que se haya **corregido el sesgo** o los datos personales hayan llegado al **final** de su **período de conservación**, si esta fecha es anterior;
f) que los **registros de las actividades** de tratamiento con arreglo a los Reglamentos (UE) 2016/679 y (UE) 2018/1725 y la Directiva (UE) 2016/680 incluyan las **razones** por las que el tratamiento de categorías especiales de datos personales **era estrictamente necesario** para **detectar** y **corregir sesgos,** y por las que ese objetivo **no podía alcanzarse** mediante el tratamiento de **otros datos**.

6. Para el desarrollo de sistemas de IA de alto riesgo que **no empleen técnicas** que impliquen el **entrenamiento** de modelos de **IA**, los apartados **2 a 5** se aplicarán únicamente a los conjuntos de **datos de prueba**.

3.3 Correspondencia del articulado con los apartados de la guía

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

Artículo Reglamento	Requerimiento Reglamento	Sección guía
10.2.a	¿Qué elementos debo implantar y cómo debo hacerlo para desarrollar un adecuado gobierno del dato?	Apartado 2.2
10.2.b	Recopilación de los datos	Apartado 4.2
10.2.c	Preparación de los datos	Apartado 4.3
10.2.d	Requisitos de información	Apartado 4.1
10.2.e	Medición y mejora de la calidad de los datos	Apartado 4.3.1
10.2.f	Análisis de sesgos en los datos	Apartado 4.3.8
10.2.g	Medición y mejora de la calidad de los datos	Apartado 4.3.1

10.2.h 10.4	Medición y mejora de la calidad de los datos	Apartado 4.3.1
10.5	Tratamiento de las categorías especiales de datos personales	Apartado 5.1

4. Gestión de los datos

En este apartado de la guía se van a desarrollar en detalle las etapas del modelo de gestión de datos presentados en la introducción.

4.1 Requisitos de información

¿Qué es?

Es el proceso consistente en **establecer** con qué **información** necesitamos alimentar nuestro sistema de IA para lograr el objetivo para el que éste ha sido diseñado.

¿Cómo debo abordarlo?

Todo sistema de IA se crea con un objetivo o finalidad, por ejemplo, dar solución a un problema existente o cubrir una necesidad identificada. Esta será la primera fase del ciclo de vida de un sistema de IA. En lo que respecta a los datos, lo primero que debemos hacer es identificar qué información necesitamos para cubrir esa necesidad o resolver ese problema.

Ejemplo

Tomando como ejemplo el **sistema de IA de la bomba de insulina inteligente**, el primer paso será analizar la solución que queremos implementar y el objetivo que buscamos alcanzar, que en este caso es determinar la cantidad de insulina que un paciente diabético necesita en un momento determinado. Así, los datos que necesitaremos recabar, por ejemplo, serán el nivel de azúcar en sangre, el ritmo cardíaco o volumen de oxígeno en sangre.

4.2 Recopilación de los datos

¿Qué es?

Es el proceso consistente en **obtener** los **datos** que contengan la información necesaria para el desarrollo de nuestro sistema de IA, que serán los datos que alimentarán nuestro sistema de IA para lograr el objetivo para el que éste ha sido diseñado.

¿Cómo debo abordarlo?

El proceso de recopilación de datos, generalmente, se lleva a cabo mediante la extracción de datos de diferentes fuentes y este es un elemento especialmente relevante. ¿Por qué? Porque la precisión y calidad de la solución que desarrollemos mediante nuestro sistema de IA dependerá, entre otros elementos, de lo representativos y suficientes que sean estos datos (la cantidad, suficiencia, completitud y adecuación de los datos, junto con otros elementos relacionados, los trataremos más en detalle en la sección 4.3).

Por ello, si recabamos datos de una única fuente de información corremos el riesgo de que la precisión de nuestro sistema de IA genere ciertas dependencias con dicha fuente de datos. Es decir, que obtengamos buenos resultados cuando utilicemos datos de dicha fuente para probar nuestro sistema, pero nos encontremos con un escenario de precisión totalmente diferente si probamos nuestro sistema con datos de una fuente nueva. En otras palabras, si entrenamos nuestro sistema con datos sesgados, corremos el riesgo de obtener un sistema que toma decisiones sesgadas (para más detalle de los sesgos en los datos, ver sección 4.3.8).

Por otro lado, uno de los problemas habituales es tener múltiples repositorios de la misma información. Para ello, es recomendable seleccionar la base más fiable (fuente de la verdad) con Data Owner y un modelo de control de la calidad en el origen del dato (no solo en el repositorio que aprovisiona el modelo).

Ejemplo

Tomemos ahora como ejemplo el sistema de IA de registro de la **asistencia al trabajo** mediante reconocimiento biométrico. Si, por ejemplo, entrenamos este sistema con imágenes sesgadas por género y raza, existe un alto riesgo de que el sistema también muestre sesgos y discriminación en su comportamiento.

Por ejemplo, si utilizamos principalmente imágenes de hombres blancos para entrenar el sistema de reconocimiento facial, es probable que el sistema tenga dificultades para reconocer y clasificar con precisión a personas de otros géneros y razas. Esto podría llevar a que el sistema cometa errores en la identificación de personas de ciertas razas o géneros, y por lo tanto a una discriminación por parte del sistema.

En este contexto, existen diferentes métodos y vías mediante las que podemos abordar el proceso de recopilación de datos. En el Anexo A se muestra un listado detallado de algunos de estos métodos y vías, con el fin de ofrecer al lector de esta guía algunos ejemplos representativos.

Es importante destacar que el proceso de recopilación de datos está estrechamente relacionado con los procesos de evaluación de la disponibilidad, cantidad y adecuación de los datos detallados en la sección 4.3. Una vez abordamos estos procesos, deberemos analizar si es necesario reevaluar o modificar el proceso de recopilación de datos (por ejemplo, incorporando datos de nuevas fuentes). Este proceso de incorporación de nuevos datos de fuentes adicionales tiene un nombre específico y es el proceso de enriquecimiento de los datos que detallaremos en la sección 4.3.6.

Adicionalmente, hay que señalar que el proceso de recopilación de datos y las fuentes que finalmente se seleccionen dependerá de la necesidad a cubrir y del contexto de la

implementación de cada sistema de IA. También deberá tenerse en especial consideración la normativa aplicable en materia de protección de datos de carácter personal[1].

4.3 Preparación de los datos

¿Qué es?

Es la fase de procesado y acondicionamiento de los datos recopilados en la fase anterior para su disposición y uso en el sistema de IA.

¿Cómo debo abordarlo?

Este procesado y acondicionamiento de los datos está compuesto, generalmente, por un conjunto de operaciones que llevaremos a cabo en función de la naturaleza de los datos que hemos recabado en la fase anterior. Es decir, hemos recabado unos datos y necesitamos acondicionarlos para poder utilizarlos y para ello tendremos que abordar una secuencia de tratamientos en función de cómo estén esos datos.

En las siguientes subsecciones, vamos a describir las principales operaciones o tratamientos que completan esta fase de preparación de los datos.

En cada sección abordaremos, **qué es cada elemento** y **cómo debemos abordarlo**. Adicionalmente, incluiremos la reflexión de **cuándo debemos abordarlo**, esto se debe a que no siempre deberemos enfrentarnos a cada uno de los procesos indicados en esta sección, dependerá de las circunstancias y eso es lo que trataremos de resolver en dicho apartado.

4.3.1 Medición y mejora de la calidad de los datos

¿Qué es?

La calidad de los datos es el grado de adecuación de los datos al propósito para el que han sido definidos o recabados. La medición de la calidad de los datos consiste en analizar cuánto se adecúan los datos a dicho propósito. La mejora es el proceso que nos permite ajustar el grado de adecuación de los datos a su propósito, es decir, nos permite aumentar la calidad de los datos.

Maximizar la calidad de los datos es, sin lugar a duda, uno de los elementos más importantes en el proceso de preparación de los datos.

¿Cuándo debo abordarlo?

La medición de la calidad de los datos y la consecuente mejora de aquellos que no dispongan de una calidad suficiente es un proceso imprescindible en un adecuado gobierno de los datos. Como veremos en las siguientes fases y como hemos comentado

[1] Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos); Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales; Ley Orgánica 7/2021, de 26 de mayo, de protección de datos personales tratados para fines de prevención, detección, investigación y enjuiciamiento de infracciones penales y de ejecución de sanciones penales.

previamente, deberemos abordar cada una de ellas en función de una serie de circunstancias que analizaremos en cada caso. No obstante, la medición y mejora de la calidad de los datos no es un proceso opcional o que dependa de las circunstancias, sino que debe estar siempre presente y adecuadamente implementado en nuestro sistema de gobernanza de datos.

¿Cómo debo abordarlo?

La forma idónea de abordar todo proceso de medición y mejora de la calidad de los datos es estableciendo una metodología de calidad de datos adecuada. En esta sección pretendemos acompañar al lector de la guía en el proceso de desarrollo de su metodología de calidad de datos. Para ello, detallaremos a continuación las fases que debemos afrontar en este proceso:

A. Determinar las dimensiones de calidad a evaluar

El primer paso es determinar cuáles son las dimensiones de calidad que vamos a evaluar para nuestros datos. Para ello, debemos conocer y comprender cuales son estas dimensiones y después determinar cuáles de ellas necesitamos para que nuestros datos tengan la calidad adecuada para cumplir su función. Es importante destacar que deberemos abordar este ejercicio para cada uno de los datos cuya calidad queramos garantizar, es decir, las dimensiones de calidad han de analizarse de forma individual para cada dato.

Para facilitar esta tarea, en el Anexo B.1 se disponen y se detallan algunos ejemplos de las dimensiones de calidad de los datos más comunes y relevantes (tales como la completitud, representatividad, auditabilidad, consistencia, relevancia, diversidad o credibilidad).
En este contexto, lo primero que deberemos hacer es seleccionar cuáles de estas dimensiones (u otras que nuestra organización considere y no estén en el inventario facilitado en el Anexo B.1) necesitamos para que cada uno de nuestros datos tengan la calidad adecuada para cumplir su función. Por ejemplo, es bastante probable que la dimensión de completitud sea necesaria para todos nuestros datos. Sin embargo, la dimensión de consistencia entre datos será necesaria únicamente en ciertos casos, por ejemplo, si tenemos dos datos que representan fechas relacionadas entre sí (la fecha de entrada a una infraestructura deberá ser menor que la fecha de salida).

Ejemplo

Seleccionemos para este ejemplo el sistema de IA relacionado con el caso de uso de la **promoción de empleados**, supongamos que somos una organización que decide desarrollar este **sistema de IA** que analiza los perfiles y el rendimiento de los trabajadores y nos ayuda a determinar quién merece en mayor medida un ascenso.

Supongamos también que nos disponemos a determinar las dimensiones de calidad de los datos que nuestro sistema de IA utiliza, entre ellos se encuentran los datos "fecha de incorporación a la empresa", "situación del empleado" y "fecha de salida de la empresa" de los empleados, y decidimos empezar a definir las dimensiones de calidad para éstos.

Ahora, lo que debemos preguntarnos es lo siguiente: ¿Qué dimensiones de calidad necesito que tengan mis datos para que éstos cumplan su objetivo y mi sistema de IA funcione adecuadamente? Comencemos seleccionando para desarrollar este ejemplo, dos de las dimensiones dispuestas en el Anexo B.1, "Completitud" y "Consistencia". Nosotros, que conocemos nuestro objetivo y nuestros datos, determinamos que los datos "situación del empleado" y "fecha de incorporación a la empresa" deben estar siempre informados ya que todo empleado debe disponer de una fecha de incorporación y una situación (por ejemplo, activo, de baja, exempleado). En cambio, el dato "fecha de salida de la empresa" únicamente estará informado para aquellos exempleados que ya no estén en nuestra organización. En cuanto a la "Consistencia", tenemos varias relaciones que han de cumplirse entre estos datos. Por ejemplo, si el dato "fecha de salida de la empresa" está informado, entonces el dato "situación del empleado" únicamente podrá tomar el valor "exempleado". Además, el dato "fecha de salida de la empresa" tiene que ser mayor que el dato "fecha de incorporación a la empresa".

B. Definir los controles de calidad a implementar para cada dimensión

Una vez identificadas las dimensiones de calidad de nuestros datos, lo que debemos hacer es definir los controles de calidad de los datos para cada requisito. Definir un control de calidad consiste en establecer una métrica que nos aporte la información que necesitamos obtener acerca del estado de la calidad de nuestro dato.

La selección de las dimensiones de calidad podía ser una tarea relativamente sencilla puesto que generalmente dependerá de la propia naturaleza de los datos, como hemos visto en la fase anterior. Sin embargo, es probable que llegados a este punto nos preguntemos, ¿qué controles he de definir?, y, ¿cuántos deberé definir? De nuevo la respuesta va ligada al objetivo de este ejercicio que es garantizar que cada uno de nuestros datos tengan la calidad adecuada para cumplir su función. Por ello, deberemos analizar y determinar qué controles son los adecuados en cada caso para poder conseguir este objetivo.

Para facilitar esta tarea, en el Anexo B.2 se disponen y se detallan algunos ejemplos de controles de calidad de los datos más comunes y relevantes para cada una de las dimensiones detallados en el Anexo B.1. Adicionalmente, se incorpora también el detalle de cómo definirlos en cada caso.

Ejemplo

Siguiendo con el ejemplo descrito en la fase anterior, estamos analizando ahora qué controles necesitamos para el requisito de completitud de los datos "situación del empleado", "fecha de incorporación a la empresa" y "fecha de salida de la empresa" y para el requisito de consistencia entre los dos últimos. Apoyándonos en el inventario de controles facilitado en el Anexo B.2 decidimos definir los siguientes controles para garantizar la calidad de nuestros datos:

ID	Dato	Tipo de control	Descripción	Definición del control
0001	Situación del empleado (A)	Completitud	Ratio de datos informados (no nulos) en el set de datos	Registros no nulos del dato A / Registros totales del dato A
0002	Fecha de incorporación (B)	Completitud	Ratio de datos informados (no nulos) en el set de datos	Registros no nulos del dato B / Registros totales del dato B
0003	Fecha de salida (C)	Completitud	Ratio de datos informados (no nulos) en el set de datos	Registros no nulos del dato C / Registros totales del dato C
0004	Fecha de incorporación (B) y Fecha de salida (C)	Consistencia	Ratio de datos donde la regla de consistencia "B<C" se cumple correctamente	Registros donde se cumple B<C / Registros totales donde B y C están informados

Adicionalmente, es importante destacar que la definición de los controles de calidad de los datos deberemos abordarla en cada uno de los puntos del ciclo de vida de los datos en los que consideremos oportuno para garantizar que llegan con la calidad adecuada cuando son utilizados por nuestro sistema de IA.

Ejemplo

Supongamos que descargo los datos de una fuente de datos abierta y los vuelco sobre una hoja Excel. Esta hoja Excel la cargo directamente en una base de datos en SQL y de ahí hago una extracción sobre mi plataforma de Python donde tengo implementado mi sistema de IA. En este ciclo de vida tenemos nuestros datos en tres repositorios diferentes y para ello tienen lugar tres procesos diferentes, supongamos los siguientes:

a) Descarga de internet a la hoja Excel.
b) Carga de la hoja Excel a base de datos SQL.
c) Extracción de una selección de los datos de la base de datos SQL a plataforma Python.

La definición de los controles de calidad podrá ser diferente en función del repositorio. Retomando por un momento el ejemplo anterior, si al pasar de “b)” a “c)” ya no seleccionamos la “fecha de incorporación a la empresa”, en el punto “b)” podremos definir los controles mencionados en el ejemplo anterior, pero en “c)” ya no podremos definir los controles que implicaban al dato que no hemos seleccionado.

En este contexto, deberemos decidir en qué puntos de ese ciclo de vida debemos definir y posteriormente implementar los controles de calidad. Esto lo deberemos determinar analizando nuestro ciclo de vida y las necesidades que consideremos oportunas para garantizar la calidad de nuestros datos. En cuanto a medición de la calidad de los datos, idealmente debería realizarse en los repositorios origen. A partir de ahí los controles en las diferentes capas o puntos del ciclo de vida deberían estar más focalizados a la calidad del proceso (es decir, que los datos se hayan copiado/ingestado/transferido correctamente; siendo por tanto controles más de nivel

técnico y no tanto funcionales). Con esto se evita la duplicidad de los mismos controles en diferentes capas y la dificultad que conlleva para su gestión y remediación.

En el ejemplo anterior, es un ciclo de vida muy corto y es posible que tengamos capacidad de definir e implementar controles de calidad en todos los puntos. Pero ¿Y si tenemos un ciclo de vida mucho mayor? Entonces deberemos determinar qué puntos son los más críticos y cuáles son los que nos permiten garantizar mejor la calidad de nuestros datos.

Por último, deberemos determinar para cada control los requisitos de calidad necesarios, esto nos permitirá evaluar si la calidad de nuestros datos es suficiente o no. Por ejemplo, si la completitud de una de las características del conjunto de datos que alimenta nuestro sistema de IA es imprescindible, deberemos establecer un requisito de calidad para el control de completitud de ese dato del 100%.

C. Implementar los controles definidos

La siguiente fase es implementar los controles de calidad definidos en la fase anterior. Consiste en desarrollar técnicamente los controles que hemos definido, es decir, traducir las reglas definidas al lenguaje técnico de cada punto del ciclo de vida del dato. Por ello, este ejercicio dependerá del tipo de plataforma y lenguaje sobre los puntos en los que hayamos definido los controles. A continuación, listamos algunos ejemplos con el fin de ayudar al lector a comprender mejor este proceso:

Ejemplo

Supongamos que hemos definido los controles de calidad indicados en el ejemplo anterior y que los hemos definido para los tres puntos del ciclo de vida del dato especificados en dicho ejemplo. En este contexto, debemos ahora implementar estos controles y, tal y como hemos explicado, el cómo lo hagamos dependerá de cada tipo de plataforma.

En este contexto, deberemos implementar las reglas que nos permitan contar el número de registros informados (no nulos) y las reglas que nos permitan comparar los registros de las fechas de incorporación y salida en el lenguaje de cada uno de estos puntos, así:

a) Implantación de controles en Excel -> utilizaremos Macros de Excel y el lenguaje VBA.
b) Implantación de controles en SQL -> utilizaremos el propio lenguaje de SQL.
c) Implantación de controles en Python -> utilizaremos el propio lenguaje de Python.

D. Reportar los resultados de los controles

En las fases anteriores nos hemos encargado de definir e implementar los controles de calidad de los datos, es decir, hemos dispuesto los medios oportunos para la medición de la calidad de nuestros datos.

Llegados a este punto, la siguiente tarea será reportar los resultados de los controles y disponer estos resultados adecuadamente para llevar a cabo el análisis del grado de adecuación de los datos al propósito para el que han sido definidos o recabados, es decir, de su calidad.

Para ello, lo que tendremos que hacer es determinar cómo queremos reportar estos resultados y que información queremos incorporar en estos reportes. Para facilitar esta tarea, detallamos a continuación algunos de los elementos más relevantes que debe contener un adecuado reporte de calidad de los datos:

- **Contexto del reporte**: en esta sección explicaremos el ámbito y contexto del reporte y el objetivo que pretende cubrir.
- **Información que se reporta**: qué datos están incorporados, la definición de los controles de calidad reportados, los requisitos de calidad de cada control, la valoración de la calidad de los datos, así como cualquier información o detalle adicional que se considere necesario.
- **Información de tiempos de ejecución**: la fecha y hora en la que se ha ejecutado cada control, la fecha y hora en la que se elaboró el reporte, la periodicidad de ejecución de los controles, la periodicidad de ejecución del reporte, así como cualquier información relacionada con periodicidades y momentos de ejecución que se considere necesario.
- **Responsables en torno al reporte:** entre ellos destacan: responsable de la ejecución de cada control, responsable de la evaluación de los resultados de los controles, responsable de remediación, responsable de la elaboración del reporte, responsable de la emisión del reporte, responsable de la recepción y revisión del reporte.
- **Condiciones de almacenamiento del reporte**: deberá informarse la ubicación de almacenamiento del reporte, para cualquier consulta posterior.

Ejemplo

En este ejemplo trataremos de disponer, sin entrar en detalle, algunos ejemplos de cada uno de los elementos a incorporar. No pretende ser un reflejo exhaustivo de lo que sería un reporte completo de los resultados de la evaluación de la calidad de los datos.

Supongamos que seguimos con el ejemplo del sistema de IA relacionado con el caso de uso de la **promoción de empleados:**

- **Contexto del reporte:** este reporte representa un informe donde se recogen los resultados de los controles de calidad de los datos que alimentan el sistema de IA de promoción de empleados. El objetivo principal de este reporte es proporcionar una evaluación transparente y precisa de la confiabilidad de los datos utilizados por el sistema.
- **Información que se reporta:**

ID	Dato	Tipo de control	Descripción	Definición del control	Requisito de calidad mínima	Evaluación del control
0001	Situación del empleado (A)	Completitud	Ratio de datos informados (no nulos) en el set de datos	Registros no nulos del dato A / Registros totales del dato A	0,90	0,98
0002	Fecha de incorporación (B)	Completitud	Ratio de datos informados (no nulos) en el set de datos	Registros no nulos del dato B / Registros totales del dato B	0,90	0,92
0003	Fecha de salida (C)	Completitud	Ratio de datos informados (no nulos) en el set de datos	Registros no nulos del dato C / Registros totales del dato C	0,90	0,94
0004	Fecha de incorporación (B) y Fecha de salida (C)	Consistencia	Ratio de datos donde la regla de consistencia "B<C" se cumple correctamente	Registros donde se cumple B<C / Registros totales donde B y C están informados	0,90	0,97

- **Información de tiempos de ejecución**:
  - Fecha y hora de la última ejecución de cada control:
    - 0001: 28/09/2023 a las 16:45
    - 0002: 28/09/2023 a las 16:50
    - 0003: 28/09/2023 a las 16:55
    - 0004: 28/09/2023 a las 17:00
  - La fecha y hora en la que se elaboró el reporte: 28/09/2023
  - La periodicidad de ejecución de los controles: diaria.
  - La periodicidad de ejecución del reporte: semanal.
- **Responsables en torno al reporte:**
  - Responsable de la ejecución de cada control: departamento de IT.
  - Responsable de la evaluación de los resultados de los controles: el responsable del sistema de IA, es decir, recursos humanos.
  - Responsable de la elaboración del reporte: analista del área de recursos humanos.
  - Responsable de la emisión del reporte: emisión automática tras generación vía correo electrónico.
  - Responsable de la recepción y revisión del reporte: director de recursos humanos.

E. Desarrollar las medidas de mejora de la calidad de los datos

Llegados a este punto, habremos definido e implementado los controles de calidad de los datos y elaborado el reporte correspondiente que nos permite analizar si la calidad de nuestros datos es suficiente o no. Lo que debemos hacer ahora es determinar las medidas oportunas para remediar o corregir las incidencias de calidad de los datos identificadas.

En este punto, es posible que nos surjan dudas como, ¿Quién debe remediar las incidencias?, ¿En caso de haber identificado gran número de incidencias como remediaremos todas ellas?, ¿Cuál deberíamos remediar antes?, ¿Debemos tener en cuenta todos los afectados por estos datos?

Para tratar de dar respuesta a estas preguntas y facilitar la comprensión del lector del proceso de remediación de la calidad de los datos, a continuación, detallamos algunos de los elementos más relevantes que deben considerarse:

- En primer lugar, deberemos inventariar todas las incidencias de calidad de los datos identificadas.
- Después, deberemos determinar qué datos y que incidencias son las más críticas para el adecuado funcionamiento de nuestro sistema de IA.
- Seguidamente deberemos consultar a todas las partes interesadas (*stakeholders*) que pudieran verse afectadas por la incidencia de calidad de estos datos.
- A continuación, tras a ver analizado la criticidad de las incidencias y tras consensuarlo con las partes interesadas, debemos establecer un orden de prioridad de remediación de estas incidencias.
- Finalmente, debemos definir un plan de remediación para cada incidencia identificada, según el orden de prioridad establecido.

Ejemplo

En este ejemplo, trataremos de mostrar, sin entrar en detalle, algunos ejemplos de cada uno de los elementos que se deben incorporar. No pretende ser un reflejo exhaustivo de lo que sería un proceso completo de mejora de la calidad de los datos.

Supongamos que seguimos con el ejemplo del sistema de IA relacionado con el caso de uso de la promoción de empleados**:**

- En primer lugar, documentaremos todas las incidencias de calidad de los datos identificadas (por ejemplo, inventariándolas en una hoja Excel, donde les asignaremos un ID a cada una de ellas, identificaremos el control que la ha detectado, la persona responsable de dicho control, a que procesos impacta la calidad de ese dato, y toda la documentación que consideremos oportuna para ayudarnos a identificar la incidencia, categorizarla y priorizarla).
- Después, con la información dispuesta en el inventario de incidencias lo que haremos será, tras consensuarlo con los diferentes actores involucrados (por ejemplo, los propietarios de los datos y los desarrolladores de los sistemas de IA) determinar un orden de prioridad de dichas incidencias según el cual trataremos de remediar cada una de ellas.
- *Seguidamente, y de forma igualmente conjunta con los actores involucrados y partes interesadas (stakeholders) que pudieran verse afectadas por la incidencia de calidad de estos datos, acordaremos un plan de remedio.*

4.3.2 Transformación de los datos

¿Qué es?

Es el proceso de conversión de los datos a un formato homogeneizado.

¿Cuándo debo abordarlo?

Deberemos abordar este proceso cuando dispongamos de datos no homogéneos para su uso en nuestro sistema de IA. Los datos pueden no ser homogéneos por tener diferente formato, por estar expresados en diferentes unidades de medida o por tratarse de índices en diferente escala, entre otros motivos. Esto podrá suceder, por ejemplo, si estamos recabando datos de diversas fuentes diferentes.

¿Cómo debo abordarlo?

Mediante la homogeneización, normalización o escalado de los datos recabados.

Ejemplo

Seleccionemos para este ejemplo el sistema de IA relacionado con el caso de uso de la promoción de empleados.

Supongamos también que nuestra organización acaba de absorber a otra organización que desarrollaba una actividad similar a la nuestra. En este contexto, deberemos incorporar a los nuevos trabajadores a nuestro sistema de IA de promoción y ascenso para que también analice sus perfiles y su rendimiento.

Por último, supongamos que nuestra organización, ubicada en Europa incorpora la información relativa a los salarios de los empleados en euros, que la organización absorbida es japonesa e incorpora esta información en yenes. Si esta información es utilizada por nuestro sistema de IA para su análisis de promoción de empleados, deberemos homogeneizar este dato, por ejemplo, pasando la información de los nuevos trabajadores de yenes a euros.

4.3.3 Agregación de los datos

¿Qué es?

Es el proceso consistente en agrupar datos con el fin de poder analizarlos y facilitar observaciones para extraer conclusiones sobre la información que éstos representan.

¿Cuándo debo abordarlo?

Deberemos abordar este proceso cuando necesitemos conocer características de agrupaciones de los datos.

¿Cómo debo abordarlo?

Mediante la transformación de los conjuntos de datos originales en conjuntos de datos agrupados en función de las características que consideremos oportuno para el análisis

que queramos llevar a cabo. Los datos agregados o calculados deberían tener el mismo esquema de control de la calidad de datos.

Ejemplo

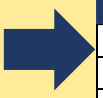
Seleccionemos para este ejemplo el mismo sistema de IA de la sección anterior, es decir, el relacionado con el caso de uso de la promoción de empleados. Supongamos ahora que estamos en la fase más incipiente del proceso de desarrollo del sistema de IA y que estamos seleccionando las características (proceso explicado en la sección 4.3.5) de los datos con los que entrenaremos nuestro sistema de IA.

Supongamos también que en este caso específico estamos desarrollando este sistema de IA en el contexto de una empresa de fabricación de un determinado tipo de piezas industriales.

En esta fase de selección de características, queremos incorporar el número medio de piezas procesado por cada empleado al día. Únicamente disponemos para ello de una base de datos que almacena un registro por cada empleado y día de trabajo con el número total de piezas procesadas en cada día.

En este contexto, si finalmente consideramos incorporar esta característica como dato de entrenamiento de nuestro sistema, el cual queremos entrenar con datos a nivel de empleado, deberemos previamente disponer del dato agregado por empleado. Para ello transformaremos la tabla previamente mencionada y la agregaremos por empleado calculando el número medio de piezas fabricado al día. A continuación, se dispone un ejemplo simple de este proceso:

ID empleado	Día	Piezas procesadas
E00001	1	234
E00001	2	259
E00001	3	289
E00002	1	301
E00002	2	297
E00002	3	278
E00003	1	223
E00003	2	245
E00003	3	237



ID empleado	Media de piezas procesadas al día
E00001	260,7
E00002	292,0
E00003	235,0

Teniendo en cuenta que debería aplicarse el esquema de control de la calidad de los datos, en este caso, se podría tener un control que verifique que el número de piezas procesadas al día esté dentro de un rango.

4.3.4 Muestreo de los datos

¿Qué es?

Es el proceso mediante el cual se extrae un subconjunto de datos desde un conjunto de datos más grande.

¿Cuándo debo abordarlo?

Generalmente haremos uso del muestreo de datos cuando necesitemos generar y ejecutar nuestro sistema de IA ágilmente, por ejemplo, para llevar a cabo una prueba que nos permita observar su funcionamiento sin tener que hacer uso del conjunto de datos completo.

También podremos hacer uso del muestreo de datos a la hora de seleccionar los conjuntos de datos de entrenamiento, validación y prueba.

¿Cómo debo abordarlo?

Existen dos técnicas principalmente conocidas y utilizadas para llevar a cabo un adecuado muestreo de datos:

- Muestreo aleatorio: cada muestra del conjunto de datos tiene la misma probabilidad de ser seleccionada.
- Muestreo estratificado: los datos se dividen en subgrupos en función de unas características determinadas. Este tipo de muestreo se utiliza, generalmente, para garantizar que cada subconjunto esté adecuadamente representado.

Ejemplo

Seleccionemos para este ejemplo el mismo sistema de IA de la sección anterior, es decir, el relacionado con el caso de uso de la **promoción de empleados.** También en este caso estamos desarrollando el sistema de IA en el mismo contexto anterior de una empresa de fabricación de un determinado tipo de piezas industriales.

Supongamos ahora que a la hora de analizar los datos para entrenar nuestro sistema de IA y seleccionar las características que utilizaremos, nos percatamos que las características que necesitamos recoger para entrenar nuestro sistema de IA no son las mismas para todos los grupos de empleados.

Por ejemplo, para los operarios que fabrican piezas industriales necesitamos la característica relacionada con el número de piezas que cada operario procesa. En cambio, para los empleados que trabajan en las oficinas no disponemos de estos datos, ya que no existen. Para este otro subconjunto, utilizaremos otras características para entrenar nuestro sistema, por ejemplo, el volumen de negocio que cada trabajador gestiona.

En este contexto, una solución será utilizar el muestreo estratificado, de forma que dividamos nuestro conjunto de datos en dos subconjuntos, uno con los datos de los operarios de fábrica y el otro con los datos de los operarios de oficina. De este modo podremos entrenar dos sistemas de IA diferentes y utilizaremos cada uno de ellos en los procesos de promoción de los empleados en función de si son de fábrica o de oficina.

4.3.5 Creación y selección de características

¿Qué es?

Son los procesos que nos permiten aumentar o reducir la dimensionalidad de nuestro conjunto de datos. Aumentaremos la dimensionalidad mediante la creación de nuevas características y la reduciremos mediante la selección de un subconjunto de las características disponibles.

¿Cuándo debo abordarlo?

Crearemos nuevas características cuando requiramos de ellas y no las dispongamos en nuestro conjunto de características. Por otro lado, seleccionaremos un subconjunto de las características que tenemos disponibles cuando determinemos que únicamente necesitamos algunas de ellas y no todas.

¿Cómo debo abordarlo?

El proceso de creación de características generalmente lo abordaremos mediante la combinación de algunas de las características disponibles. La selección de características la llevaremos a cabo determinando que características nos interesa utilizar para entrenar nuestro sistema.

También es posible que nos interese deshacernos de aquellas características que puedan ser redundantes o no aporten valor diferencial a nuestro sistema de IA. Para este escenario concreto existen técnicas como el Análisis de Componentes Principales (*PCA – Principal Component Analysis* en inglés). Esta técnica nos permite identificar aquellas características que están altamente correlacionadas y que podrían no aportar un valor diferencial si, por ejemplo, estamos entrenando un clasificador basado en redes neuronales.

Ejemplo (creación de características)

Seleccionemos para este ejemplo el **sistema de IA** de la **bomba de insulina inteligente** y supongamos que hemos determinado que las características que necesitamos para entrenar nuestro sistema son el nivel de azúcar en sangre, el ritmo cardíaco y volumen de oxígeno en sangre.
Supongamos también que por una limitación de hardware resulta muy difícil obtener el volumen de oxígeno en sangre, lo que resulta generalmente en mediciones imprecisas o incompletas. Por tanto, medir el volumen de oxígeno en sangre no es una solución viable, pero necesitamos obtener estos datos para poder determinar la insulina que necesita el paciente y, por consiguiente, para poder desarrollar nuestro sistema de IA. ¿Qué podemos hacer? Una opción es investigar si existe alguna forma de estimar estos datos.
Supongamos, en este contexto, que existiese una forma muy precisa de estimar el volumen de oxígeno en sangre mediante el uso de otros parámetros que, sí somos capaces de obtener, por ejemplo, el ritmo cardíaco y el volumen de oxígeno en los pulmones. Si tomamos la decisión de incorporar esta característica estimada del volumen de oxígeno en sangre estaremos creando, de esta forma, una nueva característica en nuestro conjunto de datos.

Ejemplo (selección de características)

Sigamos con el ejemplo desarrollado para la creación de características. Supongamos ahora que en nuestro conjunto de datos diseñado inicialmente disponíamos además de la característica que representaba el nivel de colesterol. Supongamos también que desarrollamos un Análisis de Componentes Principales sobre nuestro conjunto de datos y observamos que el nivel de colesterol es una característica totalmente correlacionada con el nivel de azúcar en sangre. En este escenario, generalmente, decidiremos prescindir de una de las dos características ya que no nos aportaría un valor diferencial en el entrenamiento de nuestro sistema de IA.

4.3.6 Enriquecimiento de los datos

¿Qué es?

Este proceso consiste en ampliar las características y datos disponibles en nuestro conjunto de datos. El objetivo es tratar de aportar información de valor adicional a la información disponible.

¿Cuándo debo abordarlo?

El enriquecimiento de los datos supone la incorporación de datos de nuevas fuentes, y es un proceso estrechamente relacionado con el proceso de recopilación de datos detallado en la sección 4.2 (tal y como ya adelantábamos cuando introdujimos el concepto de enriquecimiento de datos en dicha sección). No obstante, cuando hablamos de enriquecimiento de los datos debemos interpretarlo como un proceso posterior a la recopilación de datos inicial, que abordaremos en caso de identificar necesidades de datos o características de éstos adicionales.

¿Cómo debo abordarlo?

Lo abordaremos de la misma forma que abordamos el proceso de recopilación de datos (sección 4.2) para las nuevas fuentes de datos que consideremos

4.3.7 Etiquetado de los datos

¿Qué es?

El etiquetado de datos o anotación de datos es el proceso mediante el cual asignamos etiquetas identificativas a los datos que hemos recabado. Por ejemplo, si hemos recabado datos para desarrollar un clasificador de imágenes de animales, es el proceso mediante el cual le asignamos la etiqueta que identifica el tipo de animal que contiene cada una de las imágenes.

¿Cuándo debo abordarlo?

Lo abordaremos en caso de que, para desarrollar nuestro sistema de IA, precisemos de datos etiquetados y no dispongamos de dichas etiquetas en las fuentes que alimentan nuestro conjunto de datos.

Si atendemos a los tipos más comunes de aprendizaje automático, los datos etiquetados son necesarios en el desarrollo de sistemas de IA basados en aprendizaje supervisado (por ejemplo, clasificadores de imagen). En cambio, los sistemas de IA basados en aprendizaje no supervisado no precisan de datos etiquetados (por ejemplo, segmentaciones de grupos de datos relacionados).

¿Cómo debo abordarlo?

Las tareas de etiquetado de datos son, generalmente, costosas en tiempo y recursos. Esto se debe a que normalmente precisamos de grandes cantidades de datos para desarrollar nuestros sistemas de IA. Es cierto que también existen herramientas para el etiquetado automático de los datos, no obstante, deberemos valorar si debemos incorporar algún tipo de vigilancia humana a dicho proceso (HITL – *Human In The Loop*) que nos garantice que el etiquetado se adecua a nuestras necesidades (para más detalle de los procesos de vigilancia humana, consultar guía del artículo vigilancia humana).

A continuación, detallamos algunos de los enfoques de etiquetado de datos más comunes:

- **Etiquetado interno**: en este caso abordarán el proceso de etiquetado los expertos encargados de los sistemas de IA dentro de la organización. Es una metodología que nos proporcionará garantías en precisión y calidad ya que controlamos internamente todo el proceso, pero también es el tipo de etiquetado que más recursos nos consumirá de forma directa.

- **Etiquetado externo**: en este caso contratamos expertos externos a nuestra organización para llevar a cabo las tareas de etiquetado de los datos. Esta opción puede ser especialmente interesante para procesos de etiquetado de alto volumen de trabajo en intervalos cortos de tiempo. No obstante, nos supondrá un coste económico mayor directo que la opción de etiquetado interno.

- **Etiquetado mediante *"crowdsourcing"***: este enfoque se basa en un proceso de etiquetado de datos distribuido en la web. Existen plataformas que se encargan de diseñar y ofrecer estos servicios. Esta solución es quizá la más eficiente y rentable de todas las descritas, no obstante, no nos aporta las mismas garantías de calidad y precisión que el etiquetado interno ni tenemos un contrato de por medio con una organización que responda en caso de que suceda algún tipo de incidencia. Adicionalmente, se debe contemplar que el envenenamiento de los datos de entrenamiento supone un posible vector de ataque a nuestro sistema (Ver Guía de Ciberseguridad).
- **Etiquetado automático**: Este enfoque se basa en un proceso automático que etiqueta los datos en base a un conjunto de reglas definido o en base a un valor almacenado y asociado al objetivo del modelo. La vigilancia humana se realiza en base a una muestra aleatoria y representativa de los registros etiquetados: si hay un número de discrepancias significativas entre lo que habría etiquetado un humano y

lo que etiqueta el proceso automático se revisa el sistema de reglas o la idoneidad del etiquetado automático. Dicho etiquetado debería acompañarse de suficiente validación y supervisión de forma que se mitigue el riesgo de sesgo y deriva. En este sentido la integración de agentes de IA en marcos de gobernanza de datos como la UNE 0085 y el Reglamento de Gobernanza de Datos (DGA) puede impulsar la innovación al automatizar tareas de gestión y cumplimiento. No obstante, exige una gobernanza específica que garantice transparencia, trazabilidad y supervisión humana. La literatura propone modelos híbridos humano-máquina y arquitecturas de alineación para auditar decisiones automatizadas.

Ejemplo

Supongamos que retomamos el ejemplo del sistema de IA relacionado con el caso de uso de la **promoción de empleados.** Supongamos que ya hemos seleccionado las características necesarias y recabado el conjunto de datos históricos de los empleados con el que pretendemos entrenar nuestro sistema de IA. Dicho sistema de IA consistirá en un clasificador que nos ayudará a determinar si un empleado debe ser promocionado o no.

En este supuesto, necesitamos incorporar a nuestro conjunto de datos una característica adicional que será la etiqueta utilizada por el sistema de aprendizaje supervisado que entrenaremos. Esta etiqueta consistirá en una marca binaria que identificará si el empleado debe ser promocionado o no.

¿A partir de que información construiremos esa etiqueta? Pues a partir de las características que hemos seleccionado para entrenar nuestro sistema por ser aquellas determinantes para decidir si un empleado debe ser promocionado. Suponemos que el proceso es suficientemente complejo como para necesitar de una evaluación específica de un experto para cada caso concreto (en caso contrario no necesitaríamos entrenar este tipo de sistema y podríamos definir un conjunto de reglas simples).

¿Quién abordará dicha evaluación de cada caso e incorporará la etiqueta de promoción? Aquí podríamos elegir entre cualquiera de las alternativas explicadas anteriormente. Por ejemplo, podemos suponer un escenario donde nuestra organización dispone de un gran equipo de científicos de datos y en este caso serán ellos quienes se hagan cargo de esta tarea tras alinear una serie de criterios con el departamento de recursos humanos quienes tienen el conocimiento específico de los datos utilizados para la evaluación de los empleados.

4.3.8 Análisis de sesgos en los datos

¿Qué es el sesgo en los datos?

El sesgo es la tendencia o potencial error sistemático al que podríamos incurrir si tuviéramos un desequilibrio en los datos. Éste se produce cuando se sobre ponderan o sobre representan determinados elementos de un conjunto de datos.

Ejemplo

Por ejemplo, si entrenásemos nuestro **sistema de IA**, previamente descrito, de **promoción de empleados** con datos históricos de cargos directivos entre los años 1900 y 2000, la gran mayoría de casos de promoción exitosa de los que aprendería el sistema sería de personas de género masculino. Si tratamos de utilizar este sistema (entrenado con los datos mencionados) en el contexto actual, podríamos obtener que todos los casos de promoción de empleados propuestos son de personas de género masculino, pues nuestro conjunto de datos estaba sesgado por esta característica y las decisiones de nuestro sistema podrían quedar condicionadas en esta dirección.

¿Cuándo debo abordarlo?

El análisis de sesgos en los datos, al igual que la medición y mejora de la calidad de los datos, es un proceso imprescindible en un adecuado gobierno de los datos. Como hemos visto en las fases anteriores, debemos atender a unas circunstancias específicas para determinar cuándo abordaremos cada una de ellas. No obstante, el análisis de sesgos en los datos no es un proceso opcional o que dependa de las circunstancias, sino que debe estar siempre presente y adecuadamente implementado en nuestro sistema de gobernanza de datos.

¿Cómo debo abordar el análisis del sesgo?

El objetivo del análisis del sesgo en los datos es tener la capacidad de identificar y evaluar el potencial sesgo que puedan presentar nuestros datos y cómo podría impactar en el desarrollo de nuestro sistema de IA. También, el poder determinar la necesidad de incorporar medidas que mitiguen dicho impacto.

Hacemos especial énfasis en definir este objetivo ya que es importante destacar que el sesgo en los datos no es un elemento puramente negativo o nocivo que debamos tratar siempre de eliminar o mitigar. Lo que es importante es tener la capacidad de identificarlo y evaluar su impacto.

Ejemplo

Siguiendo el ejemplo anterior, el impacto en el proceso de selección sería negativo para las personas de género diferente al masculino. Lo más importante es que debemos entender que este es un sesgo inherente de los datos históricos que hemos recabado, es decir, los datos no son incorrectos o contienen errores, simplemente representan un desequilibrio con respecto a una característica (en este caso el género).

¿Debemos corregir ese desequilibrio en nuestros datos? Esta es una pregunta que deberemos abordar en cada contexto y escenario concretos, no existe una respuesta o solución global para todos los casos.

Lo que sí podemos hacer es tratar de diseñar un proceso que describa los pasos que deberemos dar a la hora de enfrentarnos a un análisis de los sesgos y es lo que trataremos de abordar en esta sección de la guía. Este proceso lo dividimos en tres fases que detallamos a continuación:

1. **Análisis de las principales fuentes de sesgo en los datos:** Lo primero que debemos hacer es conocer y comprender cuales son las principales fuentes de sesgo en los datos, es decir, los elementos principales que pueden provocar que mis datos estén sesgados. Para facilitar esta tarea, en el Anexo C.1 se disponen y se detallan algunos ejemplos de las fuentes de sesgo más relevantes.
2. **Evaluación del sesgo en los datos:** En segundo lugar, deberemos hacer uso de las principales técnicas de evaluación del sesgo en los datos. Estas técnicas nos ayudarán identificar si existe algún tipo de sesgo en nuestros datos y medir y valorar su impacto. Del mismo modo que en la fase anterior, se facilita en el Anexo C.2 un listado detallado de las técnicas más relevantes de evaluación del sesgo en los datos.
3. **Tratamiento del sesgo en los datos:** Por último, deberemos determinar la necesidad de implementar medidas de tratamiento de los sesgos identificados y evaluados. Estas medidas nos ayudarán a mitigar el impacto de los sesgos presentes en nuestros datos. Así mismo, en el Anexo C.3 se detallan las medidas de tratamiento del sesgo más relevantes.

4.4 Disposición de los datos

¿Qué es?

Es el proceso consistente en proporcionar los datos, tras pasar por todos los procesos de preparación detallados en las secciones anteriores, para su uso en el sistema de IA.

¿Cómo debo abordarlo?

En la primera fase del proceso, determinamos los requisitos de información que necesitamos para desarrollar nuestro sistema de IA, después recabamos los datos que contienen y representan dicha información, posteriormente los procesamos y preparamos los datos y ahora debemos disponerlos para su adecuado uso en el desarrollo del sistema de IA.

La disposición de los datos consiste en proporcionarlos una vez los hemos preparado, mediante las herramientas técnicas oportunas. Del mismo modo que cuando implantábamos los controles de calidad (donde veíamos como dependía del tipo de plataforma y lenguaje sobre los puntos en los que habíamos definido los controles) el proporcionar los datos dependerá de donde vayamos a implementar nuestro sistema de IA.

Ejemplo

Por ejemplo, supongamos que en la fase de recopilación de datos partíamos de una descarga en una hoja Excel de una fuente de datos abierta, supongamos también que posteriormente cargábamos dicha hoja Excel en nuestra plataforma de Python y allí desarrollábamos todos los procesos de preparación de los datos mediante código en Python.

Supongamos ahora que hemos diseñado nuestro sistema de IA de forma que únicamente puede ser alimentado mediante una base de datos relacional en SQL para poder garantizar las medidas de seguridad e integridad de la información adecuadas. En este contexto, para disponer adecuadamente de los datos y que puedan ser utilizados por el encargado de desarrollar el sistema de IA (que podemos ser nosotros mismos) deberemos cargar los datos en esta base de datos relacional en SQL.

Adicionalmente, también existen otros elementos que deberemos considerar para el adecuado desarrollo del proceso de disposición de datos:

- Especificar el propio proceso de disposición de datos, dónde se almacenan los datos y cómo debemos acceder a ellos adecuadamente.
- Implementar los procesos de autenticación y acceso a los datos.
- Documentar y mantener un proceso de control de versiones de los procesos de disposición de los datos.
- Mantener un registro de las diferentes actualizaciones de los datos que se dispongan.
- Informar acerca de los procesos desarrollados hasta la disposición de los datos:
    - Evaluación de requisitos de la información.
    - Recopilación de datos.
    - Preparación de datos (procesos de medición y mejora de la calidad de datos, transformación, agregación y muestreo de los datos, creación y selección de características, enriquecimiento y etiquetado de los datos y análisis de sesgos).
- Informar acerca de la posible presencia de datos personales o de categorías especiales de datos personales (ver sección 5.1).
- Informar acerca de las medidas de seguridad y protección de la privacidad (como la seudonimización o la anonimización) que hayan sido implementadas (ver sección 5.1).
- Informar acerca de posibles usos indebidos y razonablemente previsibles de los datos dispuestos.
- Documentar todo detalle adicional de los datos y metadatos necesario para facilitar la comprensión de éstos y su uso adecuado para el desarrollo del sistema de IA (por ejemplo, la distribución y propiedades estadísticas de los datos, los tipos de datos, el formato de los datos o las fechas de adquisición y actualización de los datos).

4.5 Eliminación de los datos

¿Qué es?

Es la última fase del ciclo de vida de los datos, consistente en la retirada de los datos que han sido dispuestos y utilizados para desarrollar el sistema de IA.

¿Cómo debo abordarlo?

El proceso de eliminación de los datos que hayamos dispuesto y utilizado para el desarrollo de nuestro sistema de IA, generalmente lo llevaremos a cabo, bien mediante la eliminación de estos, o bien mediante su transferencia, tal y como detallamos a continuación.

Para la **trasferencia de datos**, deberemos:

- Identificar un destinatario que presente las garantías adecuadas para asumir la responsabilidad de custodiar los datos con las respectivas obligaciones legales y contractuales asociadas.
- Documentar un acuerdo expreso con el destinatario según el cual acepta hacerse cargo de los datos y las obligaciones asociadas.
- Asegurarnos de que la transferencia de datos no supone una violación de ninguna obligación legal, contractual o de retención de datos.
- Generar un informe con todo el detalle del proceso de transferencia de datos abordado.

Para la **eliminación de los datos**, deberemos:

- Asegurarnos de que no hay ningún usuario o actor involucrado en el desarrollo del sistema de IA que requiera todavía de acceso a los datos por un motivo o fin determinado.
- Asegurarnos de que el proceso de eliminación de datos que vayamos a llevar a cabo no viole ninguna obligación legal, contractual o de retención de datos.
- Garantizar que quedan eliminados de todas las ubicaciones donde pudieran estar almacenados.
- Evaluar si existe alguna posibilidad de restauración de los datos, parcial o total, a partir del sistema de IA entrenado con estos datos. En este caso, deberemos documentarlo adecuadamente y considerarlo en el proceso de eliminación de datos.
- Revisar si los datos tienen algún tipo de importancia cultural, social o histórica. En tal caso, debemos documentar y detallar por qué no optamos por la transferencia de los datos a un destinatario que pueda mantener todas las obligaciones de datos y su importancia.

- Revisar si es posible y aporta valor la donación de los datos al dominio público. Ésta puede beneficiar a diferentes campos de investigación.
- Generar un informe con todo el detalle del proceso de eliminación de datos abordado.

Adicionalmente, deberemos disponer de los medios adecuados para gestionar las posibles solicitudes específicas de eliminación parcial de datos. Como es conocido, algunas regulaciones como GDPR, pueden obligarnos a eliminar conjuntos específicos de datos en diferentes escenarios y contextos. Con respecto a este proceso de **eliminación parcial de datos**, deberemos:

- Evaluar el impacto de esta eliminación de datos parcial sobre el conjunto de datos restante, sus especificaciones y su calidad.
- Si los datos restantes han dejado de cumplir con sus especificaciones y propiedades, deberemos documentar e informar a todas las partes interesadas de este suceso y tratar de proporcionar medidas mitigadoras.
- Garantizar que los datos que han sido eliminados quedan efectivamente eliminados de todas las ubicaciones donde pudieran estar almacenados.
- Generar un informe con todo el detalle del proceso de eliminación parcial de datos abordado.

Ejemplo

Siguiendo con el ejemplo del sistema de IA relacionado con el caso de uso de la **promoción de empleados,** supongamos ahora que recibimos una solicitud de eliminación de datos por parte de un exempleado y, supongamos también, que la ley de protección de datos respalda esta solicitud en concreto en el contexto y circunstancias en las que sucede.

En este escenario deberíamos:

- **Evaluación del impacto sobre el conjunto total de los datos y el funcionamiento del sistema**: en este caso se trata de una eliminación de datos específica de un único exempleado y se ha considerado que no tiene un impacto relevante sobre el conjunto total ni sobre el funcionamiento del sistema.
- **Documentación e información a las partes interesadas**: se envía un correo electrónico a las áreas que disponían de estos datos almacenados en sus sistemas. Estas son recursos humanos y el equipo de *Advanced Analytics*.
- **Eliminación de los datos en todas las ubicaciones**: se contacta con el área de recursos humanos (*owner* de los datos) para consultar la traza funcional e identificar todas las ubicaciones en las que el dato estaba almacenado. Después, en colaboración con el equipo de IT se accede a dichas bases de datos y se procede a eliminar los datos.
- **Generación del informe**: se genera un informe con todo el detalle del proceso de eliminación parcial de datos abordado (descripción de la solicitud, evaluación de

impacto, ubicaciones en las que se encontraban los datos, fechas y horas de las acciones realizadas).

5. Otros elementos a considerar

5.1 Tratamiento de las categorías especiales de datos personales

Como punto de partida, es necesario recordar que los proveedores y responsable del despliegue están obligados a cumplir con la normativa de protección de datos de carácter personal cuando lleven a cabo tratamientos de datos de carácter personal durante las distintas etapas del ciclo de vida de los sistemas de IA.

5.1.1 Las categorías especiales de datos personales en el Reglamento Europeo de la IA y su tratamiento

La normativa europea y nacional sobre protección de datos de carácter personal contempla categorías especiales de datos cuyo tratamiento puede suponer un especial riesgo para los derechos y libertades de las personas. Las categorías especiales de datos personales son las siguientes:

- Datos personales que revelan el origen racial o étnico;
- Datos personales que revelan opiniones políticas
- Datos personales que revelan convicciones religiosas o filosóficas;
- Datos personales que revelan la afiliación sindical;
- Datos genéticos;
- Datos biométricos dirigidos a identificar de manera unívoca a una persona física;
- Datos personales relativos a la salud, la vida sexual o la orientación sexual de una persona física.

El artículo 10.5 del Reglamento Europeo de la IA tiene como objetivo permitir a los proveedores de sistemas de IA de alto riesgo tratar categorías especiales de datos personales con el **objetivo de detectar y corregir los sesgos** que puedan estar presentes en estos sistemas de IA. Ello se permite para evitar discriminaciones provocadas por sesgos presentes en los sistemas de IA.

Así, el artículo 10. "***Datos y gobernanza de datos***" del Reglamento Europeo de la IA establece que:

AI Act

Art.10.5 – Datos y gobernanza de datos

En la medida en que sea estrictamente necesario para garantizar la detección y corrección de los sesgos asociados a los sistemas de IA de alto riesgo de conformidad con lo dispuesto en el apartado 2, letras f) y g), del presente artículo, los proveedores de dichos sistemas podrán tratar excepcionalmente las categorías especiales de datos personales siempre que ofrezcan las garantías adecuadas en relación con los derechos y las libertades

fundamentales de las personas físicas. Además de las disposiciones establecidas en los Reglamentos (UE) 2016/679 y (UE) 2018/1725 y la Directiva (UE) 2016/680, para que se produzca dicho tratamiento deben cumplirse todas las condiciones siguientes:

a) que el tratamiento de otros datos, como los sintéticos o los anonimizados, no permita efectuar de forma efectiva la detección y corrección de sesgos;

b) que las categorías especiales de datos personales estén sujetas a limitaciones técnicas relativas a la reutilización de los datos personales y a medidas punteras en materia de seguridad y protección de la intimidad, incluida la seudonimización;

c) que las categorías especiales de datos personales estén sujetas a medidas para garantizar que los datos personales tratados estén asegurados, protegidos y sujetos a garantías adecuadas, incluidos controles estrictos y documentación del acceso, a fin de evitar el uso indebido y garantizar que solo las personas autorizadas tengan acceso a dichos datos personales con obligaciones de confidencialidad adecuadas;

d) que las categorías especiales de datos personales no se transmitan ni transfieran a terceros y que estos no puedan acceder de ningún otro modo a ellos;

e) que las categorías especiales de datos personales se eliminen una vez que se haya corregido el sesgo o los datos personales hayan llegado al final de su período de conservación, si esta fecha es anterior;

f) que los registros de las actividades de tratamiento con arreglo a los Reglamentos (UE) 2016/679 y (UE) 2018/1725 y la Directiva (UE) 2016/680 incluyan las razones por las que el tratamiento de categorías especiales de datos personales era estrictamente necesario para detectar y corregir sesgos, y por las que ese objetivo no podía alcanzarse mediante el tratamiento de otros datos.

Este artículo sirve como **base de legitimación para poder tratar las categorías especiales de datos**. En este sentido, el artículo 9 apartado 1 del Reglamento (UE) 2016/679 (RGPD) permite el tratamiento de categorías especiales de datos cuando una norma de la UE o de un Estado Miembro de la UE así lo haya contemplado por razones de un interés público esencial. Así, el Reglamento Europeo de la IA, a través del artículo 10. "***Datos y gobernanza de datos***", es la **norma europea concreta que expresamente permite tratar estas categorías especiales de datos basado en un interés público esencial**, en este caso, **la detección y corrección de los sesgos para mitigar la discriminación efectuada por los sistemas de IA respecto de personas o colectivos**.

La normativa de protección de datos no sólo exige que una norma permita tratar las categorías especiales de datos. Además, esa norma a su vez debe establecer diferentes

medidas y garantías para proteger a los titulares de los datos en el marco de dicho tratamiento.

A continuación, con el fin de ayudar al lector a su adecuada comprensión, detallamos algunas de las medidas de seguridad y protección de la privacidad más relevantes:

a) Anonimización de datos

El Reglamento Europeo de la IA establece que siempre que no quede afectada de manera significativa la finalidad principal del tratamiento, esto es, la detección y corrección de sesgos, se ha de proceder a la anonimización de estos datos.

Por tanto, la premisa inicial es que **siempre que sea posible**, cuando se pretendan utilizar las categorías especiales de datos para detectar los sesgos, **se proceda a la anonimización** de estos.

Cuando el proveedor no anonimice estos datos personales por considerar que la anonimización de estos afecta o afectará significativamente al propio proceso de detección de sesgo, **éste deberá documentar y justificar** cómo el proceso de anonimización afecta a dicho proceso de detección de sesgos.

Es importante destacar que el Reglamento Europeo de la IA establece que el proceso de anonimización ha de afectar "significativamente", ello supone que, en principio, si la anonimización no afecta de una manera muy relevante al propio proceso de detección de sesgos, se ha de proceder a la anonimización de los datos a pesar de que la detección de sesgos sea menos precisa.

Los datos anonimizados quedan fuera del ámbito de aplicación de la normativa de protección de datos en la medida que sea posible demostrar objetivamente que no existe capacidad material para asociar los datos anonimizados a una persona física determinada, directa o indirectamente, ya sea mediante el uso de otros conjuntos de datos, informaciones o medidas técnicas y materiales que pudieran existir a disposición de terceros[2].

La AEPD en su página web pone a disposición de la ciudadanía, empresas y Administraciones Públicas todo un conjunto de herramientas y guías sobre los distintos procesos de anonimización de los datos de carácter personal[3].

b) Seudonimización de datos

Cuando el proveedor considere que la anonimización de los datos afectará significativamente a la detección y corrección de sesgos, se deberá proceder a la seudonimización de estos datos.

La seudonimización tiene como objetivo proteger los datos personales ocultando la identidad de las personas que son titulares de dichos datos sustituyendo uno o varios

[2] Fuente. Agencia Española de Protección de Datos Personales. Se puede consultar en: https://www.aepd.es/es/prensa-y-comunicacion/blog/anonimizacion-y-seudonimizacion
[3] Véase el espacio de contenidos dedicado a innovación y tecnologías y concretamente a los procesos de anonimización. Se puede consultar en:
https://www.aepd.es/es/areas-de-actuacion/innovacion-y-tecnologia

identificadores de datos personales por los denominados seudónimos y protegiendo adecuadamente el vínculo entre los seudónimos y la información adicional vinculada.

El conjunto de datos seudonimizados y la información adicional vinculada a dicho conjunto de datos entran dentro del ámbito de aplicación del RGPD y el resto de la normativa de protección de datos de carácter personal, ya que, si bien se reduce la identificación directa de las personas, no se elimina por completo dicha posibilidad.

Un ejemplo sencillo de seudonimización sería la sustitución de los datos identificativos de una persona como pueden ser el nombre y los apellidos por un código. De esta manera, por un lado, tendríamos un conjunto de datos seudonimizados a los que le asignamos distintos códigos y por otro lado la información adicional vinculada con dicho conjunto de datos.

Ejemplo

Historia clínica: (Dato personal)	Información adicional vinculada	Dato personal seudonimizado
Nombre: José Ruiz López	**Nombre:** José Ruiz López	**Código**: IGHJO98098A
Edad: 35	**Seudónimo:** IGHJO98098A	**Edad**: 35
Enfermedad: Cardiovascular		**Enfermedad**: Cardiovascular
Fumador: Sí		**Fumador**: Sí

Las garantías aplicables al conjunto de datos seudonimizados y a la información vinculada con dicho conjunto de datos serán entre otras[4]:

- El propio proceso de seudonimización, el cual ha de impedir la reidentificación sin disponer de la información adicional.
- Aplicación de los principios en materia de protección de datos a los datos seudonimizados. Por ejemplo: periodo de conservación de los datos, comunicación de los datos, etc.
- Despliegue de medidas de seguridad para evitar brechas de datos personales tanto sobre el conjunto seudonimizado como de la información adicional.

c) Otras medidas de garantía derivadas de la normativa de protección de datos de carácter personal

Además de las medidas previamente indicadas que expresamente se señalan en el Reglamento Europeo de la IA, se podrán implementar otra serie de garantías para asegurar

[4] Para más información sobre la seudonimización puede consultarse el espacio de contenidos dedicado a innovación y tecnologías y concretamente a los procesos de seudonimización. Se puede consultar en: https://www.aepd.es/es/areas-de-actuacion/innovacion-y-tecnologia

un cumplimiento adecuado de la normativa de protección de datos respecto del tratamiento definido en el artículo 10. "***Datos y gobernanza de datos***" del Reglamento Europeo de la IA.

Entre otras, podemos destacar las siguientes medidas que ha indicado la Agencia Española de Protección de Datos a la hora de detectar los sesgos:

- Definir procedimientos para identificar y eliminar, o al menos limitar, los sesgos en los datos utilizados para entrenar el modelo.
- Seguridad y acceso: Se han de establecer medida de identificación para el acceso a estos datos específicos, siendo conveniente establecer un proceso de gestión de derechos de acceso a esos datos. Por tanto, solo determinadas personas de la organización han de tener acceso a estos datos.
- Transmisión de datos a terceros: Resulta recomendable cifrar la información en tránsito.
- Verificar que en los datos de entrenamiento usados como entrada al modelo no existen sesgos históricos previos y que, en caso contrario, o bien se ha optado por otra fuente de datos de entrenamiento que no los contenga, o bien se ha realizado una limpieza y depuración adecuadas para su normalización.
- Adoptar medidas para evaluar la necesidad de disponer de datos adicionales de cara a mejorar la precisión o eliminar posibles sesgos.
- Implementar mecanismos de supervisión humana para controlar y asegurar la ausencia de sesgos en los resultados[5].

Además de poder incorporar las técnicas detalladas en esta sección, los proveedores deben implementar todas aquellas medidas de garantía generales que se derivan del cumplimiento general de la normativa de protección de datos de carácter personal[6].

5.1.2 La detección y/o corrección de sesgos de categorías especiales de datos en el marco del sandbox

Actualmente el Reglamento Europeo de la IA es una norma en paulatino proceso de aplicación, por tanto, mientras que esta norma no sea de aplicación en todos sus términos hasta el 2 de agosto de 2026, el artículo 10 "***Datos y gobernanza de datos***" no podrá utilizarse por parte de los proveedores para tratar las categorías especiales de datos tomando como referencia dicho precepto.

Es por ello por lo que, **en el marco del sandbox**, este precepto **no** podrá ponerse en práctica, hasta que el Reglamento Europeo de la IA sea es una norma vigente, de manera que no se permite el tratamiento de categorías especiales de datos para detectar sesgos basados en este precepto.

Dicho lo cual, cabe la posibilidad de acudir a otras vías para tratar de legitimar el tratamiento de categorías especiales de datos con la finalidad de detectar y/o corregir los

[5] Agencia Española de Protección de Datos. Guía Requisitos para Auditorías de Tratamientos que incluyan IA. 2021. Pág. 25.
[6] La Agencia Española de Protección de datos ha elaborado toda una serie de Guías centradas en el uso de sistemas de Inteligencia Artificial y cumplimiento de la normativa de protección de datos. Estos documentos son: Guías Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción (AEPD, 2020) o Requisitos para Auditorías de Tratamientos que incluyan IA (AEPD, 2021). Asimismo, 10 Malentendidos sobre el machine learning (AEPD, 2022).

sesgos, si bien, tal tratamiento no podría realizarse tomando como referencia el artículo "***Datos y gobernanza de datos***" del Reglamento Europeo de la IA.

Es importante señalar que las indicaciones realizadas en esta Guía a este precepto deberán atenderse con cautela. Sobre todo, los proveedores deberán revisar las posibles modificaciones que haya sufrido este precepto durante el desarrollo legislativo, así como las indicaciones e interpretaciones generales que hayan podido señalar las autoridades nacionales y europeas en materia de protección de datos sobre dicho artículo.

6. Documentación técnica

El Artículo 11 (Documentación Técnica) indica que se habrá de documentar el sistema de modo que demuestre que éste cumple los requisitos establecidos por el reglamento, proporcionando de manera clara y completa a las autoridades nacionales competentes y a los organismos notificados la información necesaria para evaluar la conformidad del sistema de IA con dichos requisitos.

El mencionado artículo indica que dicha documentación contendrá, como mínimo, los elementos contemplados en el anexo IV. [7] A continuación, se detallan aquellos aspectos a documentar en materia de gobernanza de datos de acuerdo con el citado anexo:

> *2. (d) cuando proceda, los requisitos en materia de datos, en forma de fichas técnicas que describan las metodologías y técnicas de entrenamiento, así como los conjuntos de datos de entrenamiento utilizados, e incluyan una descripción general de dichos conjuntos de datos e información acerca de su procedencia, su alcance y sus características principales; la manera en que se obtuvieron y seleccionaron los datos; los procedimientos de etiquetado (p. ej., para el aprendizaje supervisado) y las metodologías de depuración de datos (p. ej., la detección de anomalías);*
>
> *3. Información detallada acerca de la supervisión, el funcionamiento y el control del sistema de IA, en particular con respecto a sus capacidades y limitaciones de funcionamiento, incluidos los niveles de precisión para las personas o colectivos de personas específicos en relación con los que está previsto que se utilice el sistema y el nivel general de precisión esperado en relación con su finalidad prevista; los resultados no deseados previsibles y las fuentes de riesgo para la salud y la seguridad, los derechos fundamentales y la discriminación, en vista de la finalidad prevista del sistema de IA; las medidas de supervisión humana necesarias de conformidad con el artículo 14, incluidas las medidas técnicas establecidas para facilitar la interpretación de los resultados de salida de los sistemas de IA por parte de los responsables del despliegue; las especificaciones de los datos de entrada, según proceda.*

Estos dos puntos, por sí solos, podrían considerarse suficientes para cubrir los requisitos documentales esenciales del reglamento en materia de datos y gobernanza de datos. Sin embargo, se considera una buena práctica ampliar la documentación incorporando y justificando además los pasos indicados en la sección 4 junto con las consideraciones adicionales indicadas en la sección 5. En estas secciones se detalla cuáles son los elementos que deben implementarse para cumplir con los requisitos del artículo, así como también, se especifica cómo deberán implementarse cada uno de ellos. Para la

[7] Las pymes, incluidas las empresas emergentes, podrán facilitar los elementos de la documentación técnica especificada en el anexo IV de manera simplificada. A tal fin, la Comisión establecerá un formulario simplificado de documentación técnica orientado a las necesidades de las pequeñas empresas y las microempresas. Cuando una pyme, incluidas las empresas emergentes, opte por facilitar la información exigida en el anexo IV de manera simplificada, utilizará el formulario a que se refiere el presente apartado. Los organismos notificados aceptarán dicho formulario a efectos de la evaluación de la conformidad.

documentación de estos elementos, se debería detallar en un documento que acompañe al sistema, todas las medidas de gobernanza de datos descritas en la presente guía que hayan sido implementadas. Se debería especificar cada medida implementada y detallar cómo ha sido implementada, además se debería identificar al responsable de dicha implementación.

7. Cuestionario de autoevaluación

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

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

8. Anexos

8.1 ANEXO A – Métodos de recopilación de datos

Los datos existen en muchas formas y podemos crearlos o recopilarlos diferentes maneras. En esta sección detallamos algunos de los métodos de recopilación de datos más comunes que podemos abordar [1][3].

- **Herramientas de recopilación automatizada de datos:** en este caso la recopilación de datos se logra utilizando aplicaciones dentro de la organización (por ejemplo, el correo electrónico) o externamente a través de un sitio web, una aplicación móvil o una aplicación similar. Consiste en utilizar programas informáticos para recopilar automáticamente datos de fuentes de datos en línea. Algunos métodos para automatizar la recogida de datos son: Web-scraping, web crawling, uso de API, etc.
- **Transacciones desde otros sistemas:** la recopilación o actualización de datos realizada en otros sistemas puede fluir hacia el sistema de la organización a través del intercambio electrónico de datos u otros procesos de interconexión.
- **Sensores:** cada vez se introducen más datos en la organización a través de sistemas mecánicos como los sensores. Los sensores abarcan una amplia gama de dispositivos de adquisición de datos, como registros de sitios web, fuentes de medios sociales y dispositivos de Internet de las cosas, que incluyen dispositivos cotidianos, desde simples sensores de temperatura hasta televisores, coches, semáforos y edificios. Los datos de los sensores también pueden incluir señales potencialmente urgentes, como alertas y alarmas.
- **Nuevos contextos:** los datos de diferentes reportes pueden combinarse con otros datos para proporcionar información adicional, que a su vez se retroalimenta con los datos de la organización. En muchos casos, estos datos adicionales aportan un nuevo contexto a los datos originales y puede ser necesario tratarlos de forma diferente. Los nuevos datos contextuales pueden proceder de decisiones que den relevancia o valor a los datos existentes.
- **Suscripción**: los datos pueden estar disponibles para la organización a través de una suscripción a una fuente de datos o a almacenes de datos virtuales.
- **Crowdsourcing público:** consiste en recopilar datos mediante la ayuda o aportación de personas independientes. Por ejemplo, si necesitamos recopilar imágenes de las calles para entrenar un sistema de reconocimiento de señales de tráfico podríamos obtenerlas a través del crowdsourcing público creando una plataforma para ello y proporcionando ciertas instrucciones a los participantes. Este método es habitualmente utilizado para el etiquetado de datos.
- **Crowdsourcing privado:** es un método similar al crowdsourcing público, la diferencia está en que las aportaciones en este caso las harán grupos de especialistas privados que se contratarán para llevar a cabo esta tarea.

8.2 ANEXO B – Calidad del dato

8.2.1 ANEXO B.1 – Dimensiones de la calidad de los datos

Las dimensiones y controles de calidad de los datos se utilizan para especificar y evaluar la calidad de los datos. Generalmente, se dispone de más de un control de calidad del dato para evaluar la calidad asociada a cada dimensión. En este Anexo detallamos las principales dimensiones de la calidad en torno a los datos [6]. En el Anexo B.2 se disponen algunos ejemplos de controles de calidad de los datos para cada una de estas dimensiones.

1. Accesibilidad: se refiere al grado en que se puede acceder a los datos en un contexto específico de uso, particularmente por parte de personas que necesitan tecnología de apoyo o configuración especial debido a alguna discapacidad.

2. Auditabilidad: se refiere al grado en que todo o parte del conjunto de datos ha sido sometido a una revisión de auditoría o, en su defecto, a la disponibilidad de todo o parte del conjunto de datos para ser sometido a una revisión de auditoría.

Por ejemplo, supongamos que un conjunto de datos utilizado para entrenar un sistema de IA de reconocimiento de imagen ha sido etiquetado por un especialista externo contratado. En esta situación, para garantizar el correcto etiquetado, podríamos contratar un tercero independiente para auditar un subconjunto de las imágenes etiquetadas.

3. Identificabilidad: se refiere la capacidad de identificar un principal de información de identificación personal (PII) directa o indirectamente sobre la base de un conjunto dado de PII. Es importante comprender si cualquier PII en un conjunto de datos se puede usar para identificar un principal de PII, ya que los requisitos legales en algunas jurisdicciones pueden restringir dicha actividad. Los procesos de desidentificación o anonimización se pueden aplicar a los datos de entrenamiento, validación y prueba para reducir la posibilidad de identificación.

Por ejemplo, supongamos que disponemos de un sistema de IA desarrollado para mostrar publicidad específica a cada usuario, entrenado mediante consultas de motores de búsqueda. El conjunto de datos incluye la dirección IP del usuario, que se considera PII en algunas jurisdicciones, entre ellas la española por doctrina del TS y del TJUE. La anonimización se aplica al conjunto de datos para eliminar la dirección IP antes de que el conjunto de datos se divida en conjuntos de datos de entrenamiento, validación y prueba, tratando, de esta manera, de reducir la posibilidad de identificación.

4. Portabilidad: se refiere a la capacidad de mover datos de un sistema a otro, dentro de un contexto específico, preservando su calidad.

Por ejemplo, supongamos que para el desarrollo de nuestro sistema de IA necesitamos procesar los datos en múltiples plataformas o sistemas. Necesitamos mantener la calidad de los datos cuando los transfiramos de una plataforma o sistema a otro.

5. Comprensibilidad: se refiere a la facilidad con la que los datos puedan ser interpretados por la organización. Los sistemas de IA parten de modelos y construcciones matemáticas cuyo lenguaje principal se basa en números y algoritmos. Si éstos no pudieran

ser interpretados por los usuarios, podrían llegar a perder su función y por ello es una dimensión de la calidad relevante.

6. Coherencia temporal: se refiere a la alineación de los datos con la realidad que representan actualmente en términos de su antigüedad.

Por ejemplo, los datos sobre las personas pueden estar incompletos para las poblaciones subrepresentadas antes de los cambios en las regulaciones y normas sociales. Por otro lado, los sistemas de IA basados en datos económicos recopilados durante varias décadas pueden ser incorrectos si los datos no se corrigen por inflación, tipos de cambio y otros factores que varían con el tiempo.

7. Eficacia: se refiere a la medida en la que el conjunto de datos cumple con una serie de requisitos para su uso en el desarrollo del sistema de IA en específico:

- Para un sistema de visión computacional, la eficacia del conjunto de datos puede ser la proporción aceptable más baja en la que la cantidad de imágenes con brillo o resolución sea inferior a un umbral requerido sobre todas las muestras del conjunto de datos.
- Para un sistema de clasificación de imágenes, la eficacia del conjunto de datos puede referirse a la proporción aceptable más baja de la cantidad de imágenes que pertenecen a una clase sobre la cantidad total de muestras en el conjunto de datos.
- Para un sistema de detección de objetos, la eficacia del conjunto de datos puede referirse a la proporción aceptable más baja en la que la cantidad de imágenes con números o áreas de cuadros delimitadores es inferior a un umbral requerido en todas las muestras del conjunto de datos.

8. Eficiencia: se refiere al grado en que los datos tienen atributos que pueden procesarse y proporcionar los niveles esperados de rendimiento mediante el uso de las cantidades y tipos apropiados de recursos en un contexto específico de uso.

9. Precisión: se refiere al grado en que cada dato del conjunto de datos tiene el valor correcto. Dicho de otra forma, es el grado en que los datos tienen atributos que representan correctamente el verdadero valor de los atributos previstos. Podemos diferenciar, principalmente, dos tipos de precisión:

- Precisión sintáctica que considera la proximidad de los datos a un conjunto de datos sintácticamente correctos en un dominio relevante
- Precisión semántica que considera la proximidad de los datos a un conjunto de datos semánticamente correctos en un dominio relevante

Un elemento de datos es sintácticamente correcto si su valor de datos tiene el mismo tipo que su tipo de datos explícito y semánticamente correcto si su valor de datos tiene un valor esperado correspondiente a la tarea ML (*Machine Learning* - Aprendizaje automático). Los modelos ML son construcciones matemáticas, lo que significa que la baja precisión sintáctica o semántica de los datos en conjuntos de datos de entrenamiento, validación o prueba puede hacer que el modelo en sí sea incorrecto o que las inferencias realizadas por el modelo puedan ser incorrectas.

Para un sistema de clasificación de aprendizaje supervisado, la corrección de las etiquetas puede afectar la precisión de inferencia de un modelo entrenado a partir de datos. Los factores que deben considerarse para medir la precisión del etiquetado incluyen:

- Corrección de las etiquetas de categoría
- Corrección de los cuadros delimitadores etiquetados

10. Completitud: se refiere al grado en el que los datos presentan valores para todas las características. La incompletitud de los conjuntos de datos puede tener un grave impacto en el desarrollo del sistema de IA.

La completitud de los datos etiquetados en un conjunto de datos es relativa. En diferentes escenarios, el significado de completitud puede ser diferente y debe considerarse con contextos de uso específicos. Los factores que deben considerarse para su medida incluyen:

- La completitud de un conjunto de datos que se utiliza para una clasificación de imágenes basada en ML debe comprobar las muestras no etiquetadas en un conjunto de datos, que no se pueden usar directamente en ML supervisado.
- La completitud de un conjunto de datos que se utiliza para una detección de objetos basada en ML debe comprobar la incompletitud de los cuadros delimitadores etiquetados en los objetos.

En particular, es común en la vida real que una muestra tenga múltiples objetos en varias categorías, ya que es difícil capturar una escena con un solo objeto aislado que ocupa todo el espacio de visión. En este caso, para medir la completitud del conjunto de datos para un reconocimiento de imágenes basado en ML, debe considerar si:

- Existe cualquier objeto de destino en una muestra
- Todos los objetos de destino están categorizados
- Todos los objetos de destino detectados se etiquetan con cuadros delimitadores u otros métodos

11. Cumplimiento normativo: se refiere al grado de cumplimiento con las regulaciones, estándares, normativas u otras reglas de cumplimiento en torno a los datos. Por ejemplo, los datos personales utilizados para el desarrollo de sistemas de IA pueden estar sujetos a requisitos legales y reglamentarios.

12. Credibilidad: se refiere al grado en que los datos tienen atributos que los usuarios consideran verdaderos y creíbles en un contexto específico de uso. La credibilidad es aplicable para datos individuales, para datos relacionados en un registro de datos y para todo el conjunto de datos. El contexto en el que se utilizan los datos puede afectar su veracidad y credibilidad percibidas. Los datos pueden ser perturbados durante el procesamiento por partes autorizadas y no autorizadas.

Una preocupación emergente para ML es que las partes no autorizadas perturban los datos de entrenamiento, validación, prueba y producción para hacer deliberadamente que los modelos entrenados sean inutilizables o para manipular las inferencias hechas por un modelo entrenado.

Los procesos utilizados en la preparación de datos pueden cambiar los datos sin cambiar su significado (por ejemplo, normalización, imputación, segmentación o combinación de características). En estos casos los datos mantienen su credibilidad.

13. Equilibrio: se refiere a la distribución de las muestras para todos los aspectos del conjunto de datos. Por ejemplo, si el conjunto de datos representa X número de categorías de elementos, el número de muestras por categoría debe distribuirse uniformemente para que este conjunto de datos se equilibre. Para un conjunto de datos de imagen, estos aspectos pueden incluir etiquetas de categoría significativas para la lógica empresarial, resolución de figuras, brillo de figuras, la relación anchura/altura de los cuadros delimitadores etiquetados, el tamaño de los cuadros delimitadores etiquetados y cualquier otro que pueda influir en el rendimiento del modelo de aprendizaje automático.

El equilibrio de un conjunto de datos puede afectar a una parte del rendimiento general de un modelo de aprendizaje automático. Para un sistema de visión artificial basado en ML, se debe considerar el equilibrio del conjunto de datos.

- Cuando existen diferencias considerables de brillo o resolución entre las muestras de un conjunto de datos de entrenamiento y los datos del mundo real, los modelos de ML pueden fallar debido a datos ruidosos introducidos por debilidad o vaguedad.
- En un sistema de clasificación basado en ML, la presencia de una categoría desequilibrada de población de muestra puede dar lugar al fracaso del descubrimiento y la clasificación de instancias raras. Tales casos pueden incluso clasificarse erróneamente o identificarse como datos ruidosos.
- En un sistema de detección de objetos basado en ML, las diferencias significativas en la relación anchura/altura o el tamaño de los cuadros delimitadores pueden dar lugar a la incoherencia de tamaño sobre los objetos que se van a detectar, dado un tamaño fijo de campo receptivo. En consecuencia, esto también puede causar una pérdida de generalización, si no se aplican enfoques adicionales de verificación o ajuste de objetos multi tamaño.

14. Consistencia: se refiere al grado en que los mismos datos representados en diferentes fuentes y datos diferentes que guarden una relación significativa no presenten contradicciones o incoherencias. La consistencia es un aspecto clave de los datos utilizados para el desarrollo de sistemas de IA, ya que las características utilizadas en los datos de entrenamiento deben proporcionar en conjunto un modelo que permita inferencias correctas en los datos de producción. Además, ML puede ser literal en su interpretación de los datos. Los registros duplicados pueden causar una sobre ponderación de ciertas características. Las contradicciones entre las características de los datos de entrenamiento pueden hacer que un modelo entrenado funcione por debajo de los requisitos.

La distribución de los datos para cada característica puede considerarse una medida de consistencia. En algunos casos, los modelos de ML requieren datos con una distribución normal para cumplir con los requisitos de rendimiento.

15. Diversidad: se refiere a la diferencia entre muestras en términos de los datos objetivo. En un conjunto de datos utilizado para un sistema de IA, es importante una diferencia adecuada entre las muestras. Si todos o la mayoría de los registros de un conjunto de datos

son iguales, un modelo de aprendizaje automático entrenado a partir de ese conjunto de datos puede tener el riesgo de sobre ajustarse y, en consecuencia, ser menos generalizable. La diversidad de un conjunto de datos representa un grado de cómo el conjunto de datos contiene varios dominios de valor, etiquetas, clústeres y distribuciones entre datos individuales. La mejora de datos mediante modelos generativos de aprendizaje automático puede mejorar la diversidad de datos, pero estos enfoques pueden fallar si la diversidad del conjunto de datos original es limitada. La diversidad está estrechamente relacionada con la representatividad y el equilibrio. Es una característica de calidad de datos que se puede utilizar para evaluar la fidelidad de un conjunto de datos.

La medición de la diversidad se puede realizar en el contexto de los datos objetivo-específicos, según lo determinado por los requisitos de la tarea ML.

16. Relevancia: se refiere al grado en que un conjunto de datos es adecuado para un contexto dado.

Para un sistema de IA, la relevancia puede significar que las características seleccionadas en los datos de entrenamiento y sus valores son buenos predictores para la variable objetivo. Por ejemplo, supongamos que se utiliza un sistema de IA para determinar la solvencia de las personas. Los datos de entrenamiento son representativos de la muestra de la población que se espera que aparezca en los datos de producción. Los datos de capacitación incluyen características relevantes como el historial crediticio previo, los ingresos, la permanencia en el trabajo y el patrimonio neto, que son buenos predictores de solvencia. Los datos de entrenamiento también incluyen la altura y el peso de cada persona. Las pruebas estadísticas no muestran correlación de altura y peso con el historial crediticio anterior y se consideran malos predictores del desempeño crediticio futuro. Para mejorar la relevancia general del conjunto de datos, se eliminan las características de altura y peso.

17. Representatividad: se refiere al grado en que una muestra refleja la población objetivo que se está estudiando. Para el ML supervisado, el conjunto de datos de entrenamiento puede considerarse como la muestra y los datos de producción como la población en estudio. Cuando los datos de entrenamiento no representan suficientemente los datos de producción, el sistema de IA entrenado puede no funcionar según lo requerido.

La representatividad de los datos está relacionada con la relevancia en el sentido de que es poco probable que un conjunto de datos que no represente a la población en estudio proporcione buenos predictores para la variable objetivo. Por ejemplo, para una aplicación de reconocimiento facial, es poco probable que un conjunto de datos de entrenamiento que no contenga imágenes de personas con piel oscura dé como resultado un sistema de IA entrenado que funcione correctamente en datos de producción que incluyan imágenes de personas con piel oscura.

18. Similitud: se refiere al grado de similitud entre muestras en términos de características interesadas. Esto es relevante para las tareas de clasificación que normalmente se implementan mediante aprendizaje supervisado. Esto también es relevante para las tareas de agrupación que normalmente se implementan mediante aprendizaje no supervisado.

Tanto las tareas de clasificación como las de agrupación requieren un nivel adecuado de diferencia entre las muestras para funcionar con éxito.

Un sistema de IA entrenado en un conjunto de datos que contiene imágenes bastante similares (por ejemplo, que se generan mediante un ligero desplazamiento de píxeles basado en unas pocas imágenes semilla) puede tener el riesgo de sobreajuste y, en consecuencia, menos generalización. En este caso, es posible considerar la aplicación de enfoques de mejora de datos, como la rotación y el desplazamiento. que puede mejorar la generalización del sistema de IA. Estos enfoques no pueden funcionar si el número de imágenes semilla es limitado. En este caso, debe comprobarse la proporción de muestras similares. Otro enfoque es considerar algoritmos de agrupación con métodos de mitigación de deriva de temas.

Otras medidas identifican la similitud de datos a través del enfoque geométrico: es decir, un conjunto de datos de N registros y M características, se pueden representar como N vectores en un M-espacio dimensional, para que pueda ser analizado y comparado utilizando las herramientas de la geometría. En particular, la similitud puede asociarse a la posición mutua de los vectores en el espacio.

8.2.2 ANEXO B.2 – Controles de calidad de los datos

En este Anexo se disponen algunos ejemplos de controles de calidad de los datos para cada una de las dimensiones de calidad descritas [6].

1. Accesibilidad

- **Accesibilidad del usuario:** grado en que los usuarios consideran accesibles los valores de los datos.
  - **X = A/B**, donde:
  - **A=**número de datos relevantes para la tarea del usuario que son fácilmente accesibles.
  - **B=**número de datos relevantes para la tarea del usuario cuyo acceso es necesario.
- **Accesibilidad al formato de datos:** grado en que un dispositivo específico permite la accesibilidad.
  - **X =1- A/B**, donde:
  - **A=**número de datos que no son accesibles debido a su formato.
  - **B=**número de datos para los que se puede definir un formato accesible.

2. Auditabilidad:

- **Registros auditados:** Proporción de registros en el conjunto de datos que han sido auditados.
  - **X =A/B**, donde:
  - **A=**número de registros auditados.
  - **B=**número de registros total del conjunto de datos.
- **Registros auditables:** Proporción de los registros del conjunto de datos que están disponibles para ser auditados.
  - **X =A/B,** donde:

- **A=**número de registros disponibles para ser auditados.
- **B=**número de registros total del conjunto de datos.

3. Identificabilidad:

- **Índice de identificabilidad:** Proporción de registros en el conjunto de datos que se pueden usar para la identificación.
  - **X =A/B**, donde:
  - **A=**número de registros que contienen datos que se pueden utilizar para la identificación.
  - **B=** número de registros total del conjunto de datos.

4. Portabilidad:

- **Ratio de portabilidad de datos:** Proporción de datos que preservan su calidad tras la portabilidad.
  - **X =A/B**, donde:
  - **A=**número de datos que preservan su calidad tras la portabilidad.
  - **B=**número de datos que han sido portados.

5. Comprensibilidad:

- **Comprensibilidad de los símbolos:** Grado en el que son usados símbolos comprensibles.
  - **X =A/B**, donde:
  - **A=**número de datos representados por símbolos conocidos.
  - **B=**número de datos para los cuales es necesario comprender los símbolos.
- **Comprensibilidad semántica:** Proporción de términos definidos en el diccionario de datos.
  - **X =A/B**, donde:
  - **A=**número de datos definidos en el diccionario de datos utilizando un lenguaje comprendido por los usuarios.
  - **B=**número de datos definidos en el diccionario de datos.
- **Comprensibilidad de los valores de datos:** Grado en el que los valores de los datos son comprendidos por los usuarios.
  - **X =A/B,** donde:
  - **A=**número de datos fácilmente comprensibles para los usuarios.
  - **B=**número de datos que los usuarios necesitan comprender en un contexto concreto.
- **Comprensibilidad de la representación de datos:** Grado en que los datos son representados de forma comprensible para los usuarios en un sistema específico.
  - **X =A/B**, donde:
  - **A=**número de datos considerados comprensibles en un modelo de datos.
  - **B=**número de datos en el modelo de datos.

6. Coherencia temporal:

- **Coherencia temporal de las características:** Proporción de datos para una característica del conjunto de datos que se encuentran dentro del rango temporal requerido.

- **X =A/B**, donde:
- **A=**número de datos de la característica que se encuentran dentro del rango temporal requerido.
- **B=**número de datos de la característica.

- **Coherencia temporal de los registros:** Proporción de registros en el conjunto de datos donde todos los datos en el registro se encuentran dentro del rango temporal requerido.
  - **X =A/B**, donde:
  - **A=**número de registros que se encuentran dentro del rango temporal requerido.
  - **B=**número total de registros.

7. Eficacia:

- **Eficacia del brillo:** Relación de muestras con brillo inaceptable en un conjunto de datos de imagen.
  - **X =A/B**, donde:
  - **A=**número muestras con brillo inaceptable.
  - **B=**número total de muestras.
- **Eficacia de la resolución:** Proporción de muestras con resolución inaceptable en un conjunto de datos de imagen.
  - **X =A/B**, donde:
  - **A=**número muestras con resolución inaceptable.
  - **B=**número total de muestras.
- **Eficacia del tamaño de la categoría:** Proporción de categorías donde el número de muestras categorizadas son inferiores a un umbral.
  - **X =A/B**, donde:
  - **A=**número de categorías donde las muestras categorizadas son inferiores a un umbral.
  - **B=**número total de categorías.

8. Eficiencia:

- **Eficiencia del formato de datos:** Proporción de espacio ocupado de forma innecesaria debido a la definición del formato de los datos.
  - **X =1-A/B**, donde:
  - **A=**tamaño (en bytes) del registro de un fichero de datos ocupado innecesariamente debido a la definición del formato de los datos.
  - **B=** tamaño (en bytes) total ocupado del registro en un fichero de datos debido a la definición del formato de los datos.
- **Eficiencia en el procesamiento de datos:** Tiempo de trabajo perdido debido a la representación de los datos (formato de los datos).
  - **X =1-A/B**, donde:
  - **A=**tiempo perdido debido a la representación de los datos.
  - **B=**tiempo total de procesado.

9. Precisión:

- **Precisión sintáctica de los datos:** Relación de cercanía de los valores de datos a un conjunto de valores definidos en un dominio.
  - **X =A/B**, donde:
  - **A=**número de datos que presentan valores sintácticamente precisos.
  - **B=**número de datos para los cuales la precisión sintáctica es requerida.
- **Precisión de los datos semánticos:** Relación entre la precisión semántica de los valores de los datos en un contexto específico.
  - **X =A/B**, donde:
  - **A=**número de datos semánticamente precisos.
  - **B=**número de datos para los cuales la precisión semántica es requerida.
- **Riesgo de imprecisión del conjunto de datos:** El número de valores atípicos (*outliers*) indica un riesgo de inexactitud de los datos.
  - **X =A/B**, donde:
  - **A=**número de valores atípicos (*outliers*).
  - **B=**número de datos total del conjunto de datos.
- **Precisión del modelo de datos:** El modelo de datos describe el sistema con la precisión requerida.
  - **X =A/B**, donde:
  - **A=**número de datos del modelo que son descritos de forma precisa.
  - **B=**número de datos total utilizados en el desarrollo del modelo.
- **Precisión de etiquetado de clase:** Proporción de muestras dentro de clases incorrectas en un conjunto de datos.
  - **X =A/B**, donde:
  - **A=**número de muestras, en el que cada una está etiquetada con una clase incorrecta.
  - **B=**número de muestras total.

10. Completitud:

- **Completitud del valor:** Proporción de datos sin presencia de nulos en un conjunto de datos.
  - **X =A/B**, donde:
  - **A=**número de datos cuyo valor no es vacío.
  - **B=**número total de datos.
- **Completitud de ocurrencia de valor:** Relación entre el número de apariciones de un valor de datos determinado y el número esperado de apariciones del valor en datos con el mismo dominio en un conjunto de datos.
  - **X =A/B**, donde:
  - **A=**número de apariciones de un valor de datos determinado
  - **B=**número esperado de apariciones del valor en datos con el mismo dominio en un conjunto de datos.
- **Completitud de las características:** Proporción de datos sin presencia de nulos para una característica determinada en un conjunto de datos.
  - **X =A/B**, donde:
  - **A=**número de datos sin presencia de nulos para una característica determinada en un conjunto de datos.
  - **B=**número de datos total asociado a dicha característica.

- **Completitud del registro:** Ratio de registros sin presencia de datos vacíos en un conjunto de datos.
  - **X =A/B**, donde:
  - **A=**número de registros que tienen todos los datos informados.
  - **B=**número de datos total.
- **Completitud de la etiqueta de categoría:** Proporción de muestras con categorías sin etiquetar o etiquetadas de forma incompleta en un conjunto de datos.
  - **X =A/B**, donde:
  - **A=**número de muestras con categorías sin etiquetar o etiquetadas de forma incompleta.
  - **B=**número de muestras.

11. Cumplimiento normativo:

- **Cumplimiento de elementos de datos:** Grado en que los datos cumplen los requisitos de cumplimiento normativo.
  - **X =A/B**, donde:
  - **A=**número de datos que cumplen los requisitos de cumplimiento.
  - **B=**número total de datos

12. Credibilidad:

- **Credibilidad de los valores:** Grado en que los elementos de información se consideran verdaderos, reales y creíbles.
  - **X =A/B,** donde:
  - **A=**número de datos cuyos valores han sido validados satisfactoriamente por un proceso específico.
  - **B=**número de datos que han sido sometidos a dicho proceso.

- **Credibilidad de la fuente:** Grado en que los valores son proporcionados por una organización cualificada.
  - **X =A/B,** donde:
  - **A=**número de datos validados por una organización cualificada.
  - **B=**número de datos para los cuales la credibilidad de la fuente puede ser definida.
- **Credibilidad del diccionario de datos:** Grado en que el diccionario de datos proporciona información creíble.
  - **X =A/B,** donde:
  - **A=**número de datos en el diccionario cuyos valores han sido validados satisfactoriamente por un proceso específico.
  - **B=**número de datos total en el diccionario.
- **Credibilidad del modelo de datos:** Grado en que el modelo de datos proporciona información creíble.
  - **X =A/B,** donde:
  - **A=**número de datos en el modelo cuyos valores han sido validados satisfactoriamente por un proceso específico.
  - **B=**número de datos total del modelo.

13. Equilibrio:

- **Equilibrio de brillo:** Relación máxima de la diferencia de luminosidad de una muestra de imagen sobre la luminosidad media de las muestras de un conjunto de datos.
  - **X =A/B**, donde:
  - **A=** valor medio de brillo de las muestras.
  - **B=** valor máximo de las diferencias absolutas entre el valor de brillo de cada imagen de la muestra y A.
- **Equilibrio de resolución:** Relación máxima entre la diferencia de resolución de una muestra de imagen y la resolución media de las muestras de un conjunto de datos.
  - **X =A/B**, donde:
  - **A=** valor medio de resolución de las muestras.
  - **B=** valor máximo de las diferencias absolutas entre el valor de resolución de cada imagen de la muestra y A.
- **Equilibrio de imágenes entre categorías:** Relación máxima entre la diferencia de tamaño de categoría (número de muestras contenidas) y el tamaño medio de categoría de un conjunto de datos.
  - **X =A/B**, donde:
  - **A=** tamaño medio de la categoría del conjunto de datos.
  - **B=**valor máximo de las diferencias absolutas entre el tamaño de cada categoría del conjunto de datos y A.

14. Consistencia:

- **Consistencia del registro de datos:** Proporción de registros duplicados en el conjunto de datos.
  - **X =A/B**, donde:
  - **A=**número de registros duplicados en el conjunto de datos.
  - **B=**número total de registros.
- **Consistencia del formato de datos:** Consistencia del formato del dato para un mismo elemento de datos.
  - **X =A/B,** donde:
  - **A=**número de datos donde el formato de todas las características es consistente en las diferentes ubicaciones donde se almacenan.
  - **B=**número de datos para los que se ha definido un formato determinado.
- **Consistencia semántica:** Grado en que las reglas semánticas son respetadas.
  - **X =A/B**, donde:
  - **A=**número de datos semánticamente correctos.
  - **B=**número de datos para los que se ha definido reglas semánticas.

15. Diversidad:

- **Riqueza de etiquetas:** El número de etiquetas diferentes en un conjunto de datos.
  - **X =A**, donde:
  - **A=**número de etiquetas diferentes en el conjunto de datos.
- **Abundancia relativa de etiquetas:** La proporción del número de datos que tiene la misma etiqueta en un conjunto de datos.

- **X =A/B**, donde:
- **A=**número de datos que comparten la misma etiqueta en un conjunto de datos.
- **B=**número de datos totales.

16. Relevancia:

- **Relevancia de las características:** Proporción de características en el conjunto de datos que son relevantes para el contexto dado.
  - **X =A/B**, donde:
  - **A=**número de datos considerados relevantes en un contexto dado.
  - **B=**número de datos totales.
- **Relevancia del registro:** Proporción de registros en el conjunto de datos que son relevantes para el contexto dado.
  - **X =A/B**, donde:
  - **A=**número de registros considerados relevantes en un contexto dado.
  - **B=**número de registros totales.

17. Representatividad:

- **Ratio de representatividad:** Proporción de atributos relevantes encontrados en una muestra con respecto a los atributos que se encuentran en la muestra.
  - **X =A/B**, donde:
  - **A=**número de atributos relevantes encontrados en una muestra.
  - **B=**número de atributos de la muestra.

18. Similitud:

- **Similitud de las muestras:** Proporción de muestras similares en un conjunto de datos (cuanto más bajo, mejor).
  - **X =A-B/A,** donde:
  - **A=**número de muestras en el conjunto de datos.
  - **B=**número de grupos de muestras procesadas por un algoritmo de agrupamiento.
- **Independencia de las muestras:** Ratio entre Análisis de Componentes Principales (PCA) y la dimensión del conjunto de datos.
  - **X =1-A/B**, donde:
  - **A=**número de componentes principales.
  - **B=**número total de componentes (dimensión).

8.3 ANEXO C – Sesgos

8.3.1 ANEXO C.1 – Fuentes de sesgo

En este Anexo detallamos algunas de las fuentes de sesgo más comunes que pueden afectar al funcionamiento deseado de nuestro sistema de IA [4]:

1. Sesgos inherentes a los datos: una de las fuentes más comunes de sesgo es el introducido por los datos utilizados para desarrollar los sistemas de IA, en alguna de las

etapas de su ciclo de vida. A continuación, analizaremos los diferentes tipos de sesgos que podemos diferenciar en esta categoría.

1.1 Sesgo en la selección de los datos: este tipo de sesgo está asociado al desajuste entre los datos seleccionados y su distribución en el mundo real. Éste puede estar estrechamente relacionado con el sesgo cognitivo humano introducido en el proceso de selección de los datos.

1.1.1 Sesgo en el muestreo de los datos: este tipo de sesgo está relacionado con la recopilación no aleatoria de los datos dentro de una población. Por ejemplo, si entrenamos un sistema de reconocimiento de imagen para el desarrollo de coches autónomos únicamente con imágenes dentro de una ciudad, es posible que éste no sea capaz de reaccionar adecuadamente en otros entornos (una autovía, una carretera nacional, un camino, etc.).

1.1.2 Sesgo de cobertura: este tipo de sesgo está relacionado con un desajuste entre la población representada en el conjunto de datos con el que se ha entrenado el sistema de IA y la representada en el conjunto de datos sobre el que se están haciendo las predicciones. Por ejemplo, si entrenamos un sistema de recomendación de música entrevistando únicamente a personas de un rango de edad entre los 65 y los 80 años, estaremos introduciendo un sesgo de cobertura.

1.1.3 Sesgo de participación: este tipo de sesgo se genera cuando un grupo de población determinado tiene un índice de participación en una encuesta muy diferente al de otro grupo. Por ejemplo, supongamos que queremos entrenar un sistema de IA para predecir el resultado de unas elecciones. Para ello, recabamos datos entrevistando a la población y resulta que únicamente conseguimos recabar información de los votantes de una ideología ya que los de la ideología contraria se niegan a participar en la encuesta.

1.2 Sesgo en el etiquetado de los datos: el propio proceso de etiquetado de datos puede introducir sesgos al tratar de discretizar determinadas variables del conjunto de datos. Por ejemplo, si tratamos de etiquetar un conjunto de datos partiendo de la variable "edad" mediante las categorías "joven", "adulto" y "anciano" estaremos discretizando la población de este conjunto de datos en categorías que no siempre serán un reflejo fiel de la realidad que representan. Por otro lado, el propio proceso de etiquetado puede verse afectado por sesgos cognitivos de la persona que realiza este proceso.

1.3 Sesgo por incompletitud de los datos: en muchas ocasiones los datos que recopilemos podrán presentar incompletitudes en alguna de sus características. Este hecho es muy común cuando utilizamos datos recopilados del mundo real. Si existe un desequilibrio considerable de completitud entre poblaciones diferentes, nuestros datos podrían quedar sesgados. Por ejemplo, los datos médicos que se disponen de los pacientes a menudo varían mucho entre grupos de población de diferente capacidad adquisitiva especialmente en sociedades donde la sanidad es privada.

1.4 Sesgo en la preparación de los datos: determinadas acciones o decisiones en las etapas de preparación de los datos pueden introducir sesgos. Por ejemplo, en la etapa

de medición y mejora de la calidad podemos encontrarnos con muchos valores no informados de una característica que decidimos imputar. El abordar este proceso de imputación y cómo decidamos abordarlo puede suponer la introducción de un sesgo en nuestro conjunto de datos.

1.5 Paradoja de Simpson: esta paradoja plantea un escenario en el cual una tendencia presente en varios conjuntos de datos (por separado) desaparece y se invierte cuando estos grupos se combinan. Un ejemplo que se presenta habitualmente para ilustrar esa situación es la comparación de las tasas de mortalidad de dos hospitales, que pueden favorecer de forma global al hospital A frente al B, y sin embargo al analizarlas por procedimientos se descubre que cambia el signo de la diferencia, debido a que los pacientes con peor pronóstico y patologías más graves son internados en el hospital B con mayor frecuencia.

1.6 Variables de confusión: una variable de confusión es una variable que influye tanto en la variable dependiente como en la variable independiente causando una asociación espuria. Debido a esto, una relación percibida entre dos variables podría demostrarse como parcial o totalmente falsa.

1.7 No normalidad: la mayoría de los métodos estadísticos asumen que los datos están sujetos a una distribución normal. Sin embargo, si los datos están sujetos a una distribución diferente (por ejemplo, Chi-Square, Beta, Lorentz, Cauchy, Weibull o Pareto), los resultados pueden ser sesgados y engañosos.

1.8 Agregación de los datos: agregar datos que cubren diferentes grupos de objetos que pueden tener diferentes distribuciones estadísticas puede introducir sesgos en los datos utilizados para entrenar sistemas de IA. Esto podría ser causado por sesgos cognitivos humanos, como el sesgo de homogeneidad fuera del grupo.

1.9 Otras fuentes de sesgo en los datos: los datos también pueden estar sesgados por influencias perturbadoras externas. Este sesgo sería considerado por un algoritmo de IA como parte del modelo a generalizar y, por lo tanto, conduciría a resultados no deseados. Un ejemplo son los "*outliers*" o valores extremos que representan eventos de muy baja probabilidad. Otro ejemplo son las componentes de ruido, éste es causado por procesos estocásticos y no puede describirse de manera determinista. El ruido puede tener una influencia negativa en el modelo si se produce un sobreajuste. Además, el ruido generado artificialmente se puede utilizar para crear ejemplos contradictorios que causarán resultados no deseados.

2. Sesgos cognitivos humanos: el pensamiento a menudo se basa en procesos opacos y subjetivos que nos llevan a tomar decisiones sin saber siempre qué conduce a ellas. Estos sesgos cognitivos humanos pueden afectar las decisiones sobre la recopilación y la preparación de los datos y el desarrollo y uso del sistema de IA.

2.1 Sesgo de automatización: este tipo de sesgo se genera cuando la persona que toma las decisiones favorece las recomendaciones hechas por un sistema automatizado de toma de decisiones sobre la información hecha sin automatización, incluso cuando la automatización comete errores.

2.2 Sesgo de atribución de grupo: este tipo de sesgo se genera cuando asumimos que lo que es cierto para una muestra también es cierto para todas las muestras en ese grupo. Los efectos del sesgo de atribución grupal pueden exacerbarse si se utiliza una muestra de conveniencia para la recopilación de datos. En una muestra no representativa, se pueden hacer atribuciones que no reflejan la realidad.

2.3 Sesgo implícito: este tipo de sesgo se genera cuando hacemos asociaciones o suposiciones basadas en nuestros modelos mentales y recuerdos. Por ejemplo, para un clasificador de imagen de identificación de fotos de bodas buscar la presencia de vestidos blancos, pues no en todas las épocas y culturas el vestido de bodas es blanco.

2.4 Sesgo dentro del grupo: este tipo de sesgo se genera cuando mostramos parcialidad hacia el propio grupo al que pertenecemos. Es bastante común encontrar tendencias a hacer más atribuciones internas (disposicionales) para eventos que se reflejen positivamente en los grupos a los que pertenecemos y más externas (situacionales) para eventos que se reflejen negativamente en estos grupos.

2.5 Sesgo de homogeneidad fuera del grupo: este tipo de sesgo se genera cuando tenemos la percepción de que los miembros del grupo externo guardan similitudes más cercanas que los de nuestro propio grupo. Por ejemplo, los europeos podrían ser percibidos como un grupo más homogéneo por los estadounidenses y viceversa.

2.6 Sesgo de auto confirmación: este tipo de sesgo se genera cuando tendemos a recabar información con la intención única de confirmar nuestras creencias o hipótesis pasando por alto argumentos opuestos que puedan contradecirlas.

2.7 Sesgo social: este tipo de sesgo se genera cuando uno o más sesgos cognitivos similares (conscientes o inconscientes) están siendo sostenidos por muchos individuos en la sociedad. Se manifiesta en los sistemas de IA cuando éstos aprenden o amplifican patrones históricos preexistentes de sesgo en conjuntos de datos. El sesgo social también se manifiesta cuando se aplican supuestos culturales sobre los datos sin tener en cuenta la variación intercultural. Por ejemplo, cuando los datos históricos son inapropiados para las inferencias que se hacen, posiblemente reforzando puntos de vista sociales comunes pero inexactos.

8.3.2 ANEXO C.2 – Técnicas de evaluación del sesgo

En el desarrollo de un sistema de IA es muy importante tener en cuenta los posibles sesgos que podrían conducir a comportamientos no deseados y poco equitativos del sistema. Una manera de identificar posibles sesgos es evaluando los resultados del sistema de IA utilizando métricas desarrolladas para ello. Estas métricas, tratan de evaluar las diferencias entre los valores promedio observados y los valores verdaderos [4].

La mayor parte del trabajo desarrollado en torno a las métricas mencionadas se ha centrado en los sistemas de IA basados en clasificación o regresión con respecto a grupos definidos en términos de uno o más atributos biográficos, demográficos o de comportamiento. En esta sección se presentan los enfoques más relevantes para evaluar el sesgo de los sistemas de IA basados en la clasificación.

1. Sistemas de clasificación: el sesgo en los sistemas de clasificación puede detectarse mediante mediciones de diferentes tipos de errores con respecto a varios grupos. El enfoque de dividir los datos en conjuntos de entrenamiento, validación y prueba se complementa subdividiendo cada uno de esos conjuntos en función de las características con respecto a las cuales se espera que el sistema sea justo. Si hay múltiples características relevantes para detectar posibles sesgos en un sistema en particular, entonces esas características podrían considerarse independientes o interseccionales. Por ejemplo, un sistema que es imparcial con respecto al género y la raza independientemente podría estar sesgado hacia una combinación específica de los dos.

Una vez que los datos se han dividido adecuadamente, se calculan métricas de sesgo en cada grupo y se realizan comparaciones entre grupos relacionados. Un sistema de clasificación podría ser "imparcial" con respecto a las características relevantes si las mediciones basadas en métricas entre grupos están dentro de un delta suficientemente pequeño.

2. Sistemas de aprendizaje por refuerzo: el sesgo en los sistemas de aprendizaje por refuerzo difiere del de los sistemas de clasificación. En este caso el sesgo se centra en la serie de acciones que va tomando el sistema. Si bien probar el sesgo de los sistemas de aprendizaje por refuerzo es una práctica menos desarrollada, están surgiendo algunos métodos. Los sistemas de aprendizaje por refuerzo generalmente se entrenan en función de una medición de utilidad siguiendo sus acciones. Las métricas de sesgo se pueden aplicar a esta medición de utilidad en todos los grupos.

Además de usar la utilidad como análogo a la precisión, el sesgo en los sistemas de aprendizaje continuo que utilizan el aprendizaje por refuerzo puede manifestarse como diferentes "*path attractors*" para diferentes grupos. Este podría ser particularmente el caso en los sistemas de aprendizaje por refuerzo que realizan recomendaciones de contenido. En tales sistemas, el "*ground truth*" es un concepto menos definido. Si, por ejemplo, un asistente de IA para guiar a los estudiantes universitarios a lo largo de su curso de estudio condujo consistentemente a las mujeres hacia profesiones de cuidado y a los hombres hacia profesiones de ingeniería, entonces ese sistema está potencialmente sesgado independientemente de las métricas de utilidad.

3. Matriz de confusión: una matriz de confusión (figura 1) es un método utilizado para descubrir el sesgo de los clasificadores. Informa el número de falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos e incluye otros criterios de rendimiento derivados de estos valores. Dado que una matriz de confusión contiene y compara múltiples métricas, permite un análisis detallado del rendimiento de un clasificador y es útil para eludir o descubrir las debilidades de las métricas individuales. Por ejemplo, el uso de métricas individuales para medir el rendimiento de un clasificador daría lugar a resultados engañosos si los conjuntos de datos subyacentes estuvieran desequilibrados y, por lo tanto, no pueden proporcionar resultados confiables.

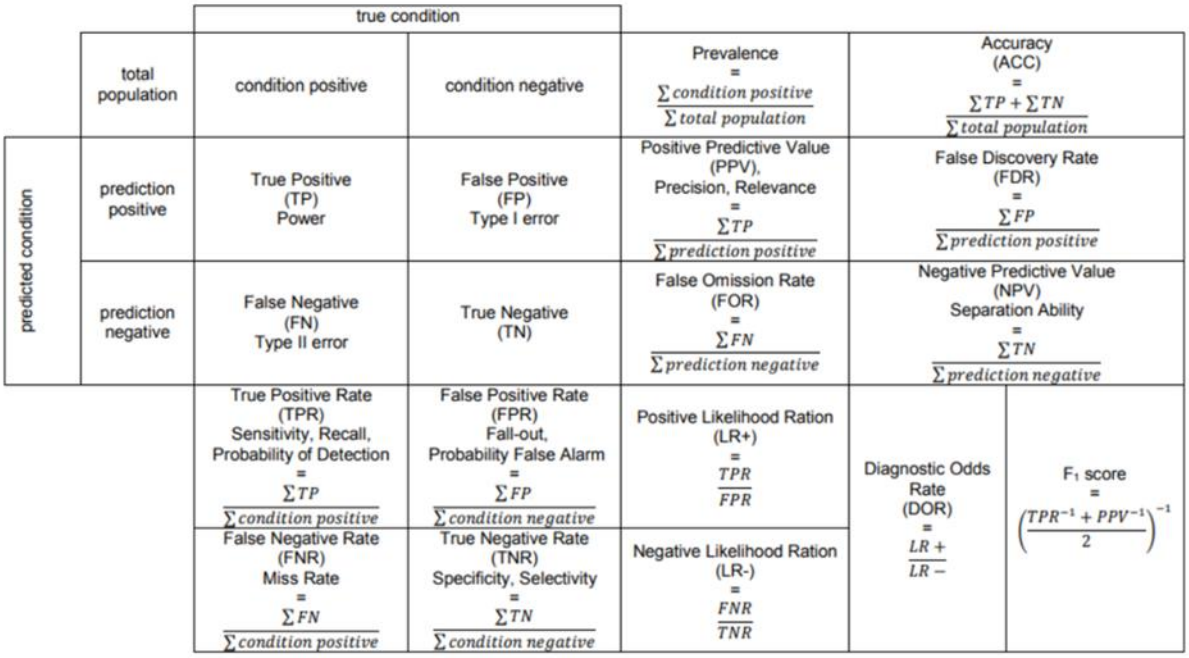


Figura 1 – Matriz de confusión y métricas de rendimiento de clasificación derivadas [17]

4. Igualdad de probabilidades: también conocida por su término en inglés *"equalized odds"* significa que las decisiones de un algoritmo son independientes de una categoría *A* dada una entrada *Y*. Un predictor $\hat{Y}$ satisface las probabilidades igualadas con respecto a la categoría *A* y la salida *Y*, si $\hat{Y}$ y *A* son independientes condicionales a *Y*:

$$P(\hat{Y} = \hat{y} | Y = y, A = m) = P(\hat{Y} = \hat{y} | Y = y, A = n), \forall\, Y, m, n \in A$$

Esto implica que las tasas de verdaderos positivos (TPR) son iguales en todas las categorías demográficas y las tasas de falsos positivos (FPR) son iguales en todas las categorías demográficas.

Debemos considerar que esta definición permite que los modelos tengan en cuenta la información demográfica. TPR es igual a 1 - Tasa de falsos negativos (FNR), por lo que esto también fomenta tasas de falsos negativos iguales en todas las categorías demográficas. Para ver las compensaciones entre falsos negativos y falsos positivos, comparar la tasa de falsos negativos y la tasa de falsos positivos podría ayudar.

5. Igualdad de oportunidades: también conocida por su término en inglés *"equality of opportunity"* significa que las decisiones de un algoritmo donde $\hat{Y}=1$ son independientes de una categoría *A* dada la entrada *Y=1*.

Un predictor binario $\hat{Y}$ satisface la igualdad de oportunidades con respecto a *A* y *Y* si $\hat{Y}=1$ y *A* son independientes condicionales a *Y=1*. Formalmente:

$$P(\hat{Y}=1|Y=1, A=m) = P(\hat{Y}=1|Y=1, A=n), \forall\, m, n \in A$$

Esto implica tasas positivas verdaderas (TPR) iguales en todas las categorías demográficas.

6. Paridad: La paridad estadística significa que hay tasas de predicción iguales entre las categorías. La paridad demográfica dice que hay tasas de predicción iguales entre las categorías demográficas, como la raza. La paridad demográfica, que es un caso de paridad estadística, significa que una decisión, como aceptar o rechazar una solicitud de préstamo, debe ser independiente de un atributo demográfico. Formalmente, dada la variable demográfica A:

$$P(\hat{Y} = \hat{y}|A = m) = P(\hat{Y} = \hat{y}|A = n), \forall\, m, n \in A$$

La paridad no captura los casos en los que la decisión de salida está correlacionada con uno de los grupos o atributos que se evalúan, y no hay garantía de que las predicciones realizadas sean igualmente buenas para cada categoría.

7. Igualdad predictiva: la igualdad predictiva implica tasas de falsos positivos (FPR) iguales en todas las categorías demográficas. Formalmente:

$$P(\hat{Y} = 1|Y = 0, A = m) = P(\hat{Y} = 1|Y = 0, A = n), \forall\, m, n \in A$$

8.3.3 ANEXO C.3 – Medidas de tratamiento del sesgo

En este Anexo detallamos algunas de las medidas de tratamiento del sesgo más comunes que pueden implementarse para mitigar el impacto de los posibles sesgos que hayamos identificado y afectar al funcionamiento deseado de nuestro sistema de IA. Las medidas detalladas a continuación van más allá del contexto de los datos, se trata de un enfoque que trata de abarcar el tratamiento de los sesgos en las diferentes etapas del ciclo de vida de un sistema de IA [4]:

1. Análisis del contexto y los requisitos del sistema

El análisis de los requisitos del sistema es una actividad importante para mitigar el sesgo. Es la etapa donde se analizan los requisitos internos y externos, se determinan las partes interesadas del sistema y se evalúan los objetivos del sistema. Para este hito, se han identificado los riesgos planteados por el sistema, se ha evaluado el impacto para las partes interesadas identificadas y se han definido los niveles de participación de las partes interesadas.

1.1 Requisitos externos

La identificación de requisitos externos como parte de la actividad de análisis del sistema es una parte normal del ciclo de vida de desarrollo y adquisición de sistemas. Durante este proceso podría prestarse especial atención a los siguientes marcos reglamentarios:

- Instrumentos internacionales de derechos humanos e igualdad.
- Leyes y orientaciones específicas relacionadas con la provisión de soluciones técnicas.
- La legislación sobre protección de datos y privacidad podría incluir disposiciones relativas a la toma de decisiones automatizada.
- Derecho de la competencia y de los negocios.

A continuación, se presentan ejemplos de posibles tipos de obligaciones para la entidad responsable:

- La necesidad de una evaluación de riesgos, que podría incluir las preocupaciones sociales desde la perspectiva de las partes interesadas afectadas.
- Notificación a los usuarios de que están sujetos a una decisión automatizada, el requisito de obtener el consentimiento explícito y proporcionar una alternativa no automatizada cuando no se otorga el consentimiento.
- Asegurar un cierto nivel de auditabilidad o explicabilidad en la solución, para apoyar el análisis de una decisión o evento en particular.
- Actividades para cuantificar o mitigar los riesgos, como la recopilación de metadatos sobre fuentes de datos para comprender la procedencia y la calidad.
- Provisión para la participación significativa de un ser humano en el proceso de toma de decisiones.

La provisión y el precio equivalentes de servicios para grupos de personas con ciertas características. Esto podría requerir la capacidad de demostrar que la igualdad se logra en la práctica.

1.2 Requisitos internos

Además de los requisitos reglamentarios, muchos otros factores podrían contribuir al deseo de una parte interesada de mitigar el sesgo, tales como:

- Metas, estrategias y políticas internas de una organización.
- Valores morales o culturales.
- Evitar preocupaciones sociales o daños a la reputación.

El proceso de análisis puede prestar especial atención a cinco áreas específicas: inclusión de expertos transdisciplinarios; la identificación de las partes interesadas; la selección de fuentes de datos; cambio externo; y especificación de los criterios de aceptación, incluidos los niveles aceptables de sesgo.

1.3 Expertos transdisciplinarios

Si bien el sesgo no deseado es un tema relativamente nuevo en el contexto de la tecnología, es un tema bien entendido en las ciencias sociales. Como parte del proceso de análisis de requisitos (y, de hecho, de todo el ciclo de vida del sistema), es relevante considerar la experiencia que podría requerirse para mitigar completamente las preocupaciones sociales sobre el sesgo y tener en cuenta diversas perspectivas. Esto podría incluir:

- Científicos sociales y especialistas en ética;
- Científicos de datos y especialistas en calidad;
- Expertos legales y de privacidad de datos;
- Representantes de usuarios o grupos de partes interesadas externas.

Por ejemplo, los diseñadores de un sistema de reconocimiento facial podrían dar importancia a la característica de contorno facial en su diseño y pasar por alto el hecho de que el contorno podría estar (parcial / completamente) cubierto para personas con antecedentes culturales / religiosos particulares. Es más probable que un equipo suficientemente diverso identifique tales limitaciones con diseños, suposiciones y conjuntos de datos.

1.4 Identificación de las partes interesadas

El análisis tradicional de los requisitos incluye la identificación de las partes interesadas. Sin embargo, para cumplir con los aspectos de los marcos regulatorios antes mencionados y mitigar adecuadamente las preocupaciones sociales, podría ser necesario ampliar esta definición tradicional de parte interesada para incluir a aquellos directa e indirectamente afectados por el sistema implementado.

Según los tipos de datos que se utilizan para tomar decisiones automatizadas, podría ser necesario descomponer aún más las listas de partes interesadas en grupos de personas que podrían verse afectadas de manera diferente por el sesgo en el sistema, tener diferentes habilidades en el uso del sistema o tener diferentes niveles de conocimiento y acceso. Es importante considerar qué sesgos, experiencias negativas o resultados discriminatorios podrían ocurrir.

Estos pueden considerarse a través de una variedad de diseño participativo o métodos etnográficos, los cuales implican un alcance activo y una discusión con los grupos afectados. Por lo general, es insuficiente señalar quién podría verse afectado teóricamente sin el aporte directo de esos grupos. A menudo, tampoco es suficiente que un miembro del grupo afectado (que posiblemente también trabaje como parte del equipo de diseño) evalúe cómo se ve afectado ese grupo. Los grupos no son monolitos, y una persona no siempre puede representar adecuadamente la gama de perspectivas posibles.

La identificación y el compromiso de las partes interesadas podrían incluirse en una descripción formal y documentación de las áreas de preocupación anticipadas y las posibles consecuencias para los grupos afectados, positivas y negativas. De estos, se pueden derivar requisitos más calificados y cuantitativos. Una evaluación de impacto humano realizada en fases posteriores puede revisar estas áreas de preocupación y evaluar si la preocupación se mitigó con éxito.

1.5 Selección y documentación de fuentes de datos

La selección de datos utilizados para formar reglas explícitas dentro de un sistema experto basado en reglas, o datos que se utilizan para entrenar modelos de ML, es una actividad esencial que tiene una influencia significativa en el sesgo.

En el caso del conocimiento explícito, es importante considerar los sesgos cognitivos humanos ya presentes en aquellos que especifican el conocimiento. El sesgo cognitivo humano podría estar presente en un juicio humano, que luego se codifica en un sistema basado en reglas, y luego se propaga durante toda la vida útil del sistema a una escala mayor. Si bien podría considerarse aceptable que tal sesgo cognitivo esté presente en una sola decisión humana, propagar ese sesgo a través de la toma de decisiones automatizada podría tener un impacto mucho mayor.

Los sistemas estadísticos de IA que aprenden de los datos sin que se especifique el conocimiento explícito sufren muchos riesgos. Se podría considerar la fuente de datos individual para determinar:

- **Integridad**. Un origen de datos que excluye determinados registros porque no contienen las mismas características para todos los registros podría no proporcionar

una imagen completa y provocar defectos en el proceso de entrenamiento. Es poco probable que los datos disponibles públicamente (por ejemplo, de Internet) tengan una distribución equivalente entre grupos de personas.

- **Exactitud**. Un origen de datos que contenga datos inexactos propagará esas imprecisiones en un modelo de aprendizaje automático. Esto podría dar lugar a problemas generales de precisión, pero estos problemas también podrían estar sesgados hacia ciertos grupos de personas, por ejemplo, aquellos con menos historial de crédito.
- **Procedimientos de recogida**. Es importante comprender el linaje de los datos, cómo se recopilan, cómo se ingresan y si estos procesos afectan la integridad y la precisión. Se puede considerar la ubicación y el entorno del que se obtienen los datos.
- **Coherencia temporal.** La frecuencia con la que se recopilan o actualizan los datos puede ser relevante para garantizar su exactitud. Por el contrario, la actualización puede requerir una nueva evaluación. Un sistema que pasa una auditoría puede quedar fuera de conformidad, especialmente si las actualizaciones provienen de un proceso diferente al de los datos iniciales.
- **Consistencia**. La consistencia con la que se determinan los datos de entrada (o etiquetas) puede ser importante. Por ejemplo, si un humano está categorizando elementos que no tienen un límite claro entre categorías, diferentes categorías o etiquetas podrían resultar de los mismos datos.

1.6 Cambio externo

Tal vez sea necesario prestar atención a la forma en que podrían producirse cambios en la utilización del sistema desarrollado o adquirido, ya que ello podría requerir una reevaluación completa de un sistema.

Algunos ejemplos son:

- El despliegue de un sistema existente en un entorno diferente, incluidos diferentes usuarios, mercados objetivo y fuentes de datos, puede cambiar los riesgos y requerir que se vuelva a evaluar un sistema.
- Con el tiempo, la relación entre las entradas y salidas del sistema puede cambiar. Por ejemplo, un sistema que utiliza un modelo de ML para tomar decisiones basadas en correlaciones establecidas durante el entrenamiento inicial del sistema podría sufrir si esas correlaciones cambian con el tiempo.

Los casos de uso para el sistema pueden desarrollarse, ya sea deliberada u orgánicamente, requiriendo una reevaluación de los riesgos.

- Las normas sociales cambian con el tiempo (por ejemplo, las actitudes hacia las normas de comportamiento de género, la forma corporal ideal o el tabaquismo). Es posible que sea necesario reevaluar el sesgo en los sistemas de IA para considerar los cambios resultantes (como métricas, riesgos, partes interesadas o requisitos) y abordarse en consecuencia.

1.7 Criterios de aceptación

Los requisitos efectivos son comprobables, ya que cuando se evalúan es posible determinar si un sistema los cumple. A menudo, el rendimiento del sistema de IA se compara con el de los humanos. Aun así, es beneficioso poder especificar el rendimiento de manera estadística. Por ejemplo, las partes interesadas pueden especificar un límite para las decisiones falsas positivas o falsas negativas, además de una métrica de precisión general.

Establecer la aceptación en términos de características específicas del sistema y el grado de su cumplimiento por adelantado permite una evaluación y toma de decisiones efectivas.

Los criterios de falla pueden ser el límite inferior de aceptación, estableciendo límites claros para el rendimiento aceptable de un modelo. Sin que se establezcan y monitoreen estos criterios, un sistema de IA podría desviarse de tal manera que surjan sesgos no deseados sin ser notados o remediados. La forma en que un sistema falla también podría necesitar ser considerada cuidadosamente como parte del proceso de diseño para evitar que ocurran casos extremos de sesgo.

2. Diseño y desarrollo

Los modelos en sí mismos pueden estar sesgados si no se tiene cuidado para evitar esto. Los sesgos humanos pueden codificarse en los sistemas de ML a través de suposiciones implícitas que se abren camino en el diseño. Por lo tanto, ayuda a identificar y hacer explícitas las suposiciones implícitas.

2.1 Representación y etiquetado de datos

Un paso clave en el desarrollo de un sistema de aprendizaje automático es decidir cómo representar mejor los datos de entrenamiento en características que el modelo pueda interpretar. Esto también se denomina ingeniería de características, y el Anexo C.1 describe algunos tipos de sesgo de datos que podrían afectar este proceso. Hay varios criterios a menudo implícitos que entran en este proceso, incluidos los criterios por los cuales los datos se consideran "buenos" o "malos" (por ejemplo, si una foto sobreexpuesta podría mantenerse en un conjunto de datos). Estos criterios sobre qué datos se incluyen en los datos de entrenamiento y qué características se seleccionan pueden hacerse explícitos. Es importante considerar cómo los datos se relacionan con el propósito de construir el sistema, el proceso por el cual se eligen las características y las personas que eligen las características y su justificación.

Es importante evaluar las características elegidas para cualquier dato y sesgo cognitivo humano, como valores de entidad faltantes, valores de características inesperadas o sesgo de datos. Cualquiera de estos podría indicar que ciertos grupos o características no están representados con precisión en los datos.

La falta de valores de entidad podría ser el resultado de sesgos implícitos en el proceso de recopilación de datos, que podrían necesitar ser modificados.

En los algoritmos de aprendizaje profundo donde las características se crean durante el entrenamiento, las etiquetas correctas son críticas. Las anotaciones realizadas por humanos para crear etiquetas para los datos pueden ser propensas a sesgos debido a sesgos cognitivos humanos o errores humanos debido a dificultades en la tarea de etiquetado en

sí. Es importante garantizar que las etiquetas sean correctas mediante la evaluación tanto de los etiquetadores como de los datos finales etiquetados.

Los anotadores de datos a menudo anotan y etiquetan los datos que se utilizarán para el entrenamiento de ML supervisado. Los errores consistentes o los sesgos cognitivos humanos que se manifiestan durante ese proceso se propagan en un modelo entrenado.

Cuando se utilizan anotadores de datos para el etiquetado de datos, podría ser útil comprender la diversidad y los objetivos de las personas que anotan los datos. Por ejemplo, podría considerarse cómo se ve el éxito para diferentes trabajadores, y las compensaciones entre el tiempo dedicado a la tarea y el disfrute de la tarea.

Los responsables del despliegue pueden desarrollar tareas que tengan en cuenta las diferencias humanas (y los sesgos cognitivos) en la anotación, por ejemplo, mediante el uso de preguntas de prueba de oro con respuestas conocidas u otras formas de preselección de participantes.

La claridad de las instrucciones, así como la obtención de comentarios de los trabajadores de la multitud sobre tareas potencialmente confusas, podrían ser importantes para reducir el sesgo no deseado. La variabilidad humana, incluida la accesibilidad, la memoria muscular y los sesgos cognitivos humanos en la anotación, podría explicarse mediante el uso de un conjunto estándar de preguntas con respuestas conocidas.

2.2 Herramientas de transparencia

Para aclarar los casos de uso previstos de los modelos de ML y minimizar su uso en contextos para los que no son adecuados, los modelos publicados pueden ir acompañados de documentación que detalle sus características de rendimiento. Las herramientas de transparencia del modelo podrían proporcionar un marco para la presentación de informes transparentes sobre la procedencia del modelo de ML, el uso y la evaluación basada en la imparcialidad. La documentación del modelo puede incluir:

- Información cualitativa, como consideraciones éticas, usuarios objetivo y casos de uso;
- Información cuantitativa, que consiste en una evaluación desagregada del modelo (dividida en los diferentes subgrupos objetivo) e interseccional (incluida la evaluación de múltiples subgrupos en combinación, por ejemplo, raza + género).
- Información de datos, si es posible, que podría formalizarse como una herramienta de transparencia de datos.

La utilidad y precisión de una herramienta de transparencia depende de la integridad del creador o creadores de la propia herramienta y pueden almacenarse como documentación o metadatos asociados con cada modelo.

2.3 Entrenamiento

2.3.1 Datos de entrenamiento

Un enfoque aparentemente sencillo para la mitigación del sesgo es eliminar las características relevantes que podrían ser responsables del sesgo directamente. Para ilustrar eso, en un caso de uso que preselecciona automáticamente a los candidatos en función de la información del currículum, ejemplos de tales características son la raza, el género y la edad. Al mismo tiempo, las características que son relevantes para el caso de uso incluyen experiencia, habilidades, calificación, certificación y membresía profesional. Si bien la eliminación de la raza, el género y la edad podría abordar el problema, otras características y variables indirectas podrían reflejar indirectamente sesgos. Por ejemplo, características como el saludo/prefijo (Sr./Sra.) o la ocupación pueden revelar el género. Por lo tanto, eliminar solo las características asociadas con el sesgo puede no funcionar siempre.

Se pueden usar métodos basados en datos para mitigar el sesgo en los datos de entrenamiento. La reponderación de datos, por ejemplo, podría aumentar el peso de las muestras que se alinean con un objetivo. Tales técnicas incluyen:

- Técnicas de muestreo para medir la representatividad de muestras de diferentes fuentes para evaluar el sesgo de selección.
- Muestreo de estratificación para superar un fenómeno de rareza. El muestreo estratificado podría utilizarse aumentando la frecuencia relativa de los casos positivos en comparación con los negativos.
- Selección cuidadosa de las características en los casos en que las características de la muestra tienen una fuerte correlación con el sesgo a excluir (por ejemplo, género o color).

Otro enfoque es averiguar la cantidad de sesgo presente en los datos y compensar el sesgo del resultado. Usando una serie de pasos, es posible averiguar la contribución de la característica y la importancia relativa de cada característica en la predicción del modelo. Es posible compensar toda la influencia por una característica que causa el sesgo. El proceso de determinar la importancia relativa de las características puede incluir lo siguiente:

- Proyección ortogonal iterativa de características. Dada la entrada y salida de un modelo de ML, el método busca producir una clasificación de entrada que corresponda a la dependencia del sistema de aprendizaje automático de cada entrada en su proceso de toma de decisiones y, por lo tanto, podría detectar sesgos que involucran ciertas características.
- Redundancia mínima, relevancia máxima. La selección de características identifica subconjuntos de datos que son relevantes para los parámetros utilizados. Un esquema es seleccionar las características que se correlacionan más fuertemente con la variable de clasificación y al mismo tiempo están mutuamente distantes entre sí. Se ha descubierto que este esquema, denominado selección de redundancia mínima y relevancia máxima, es más potente que otros modos de selección de características.
- Regresión Ridge/Regresión LASSO. LASSO y Ridge son métodos de regresión lineal con regularización para evitar el sobreajuste a los datos de entrenamiento. Estas técnicas también se utilizan para ayudar con la selección de características.

- Random Forest. Random Forest es un enfoque que combina varios árboles de decisión aleatorios y agrega sus predicciones promediando. Esta técnica ha demostrado un excelente rendimiento en entornos donde el número de variables es mucho mayor que el número de observaciones.

2.3.2 Testeo

La evaluación del sesgo es una tarea difícil. Los métodos de agrupación y visualización de los datos de entrenamiento, las características derivadas o las predicciones resultantes pueden ayudar a detectar desequilibrios o posibles sesgos en los datos o el sistema de entrenamiento. En muchos casos, la preparación o conservación de un conjunto de datos equilibrado puede ocupar la mayor parte del tiempo de desarrollo en sistemas basados en el aprendizaje profundo. Tener equipos separados trabajando en la capacitación y la evaluación también podría evitar la influencia de los sesgos cognitivos individuales.

La investigación de defectos aparentes en el modelo podría revelar por qué no está maximizando la precisión general. La resolución de estos defectos podría mejorar la precisión general. Los conjuntos de datos sub representativos de ciertos grupos podrían necesitar datos de entrenamiento adicionales para mejorar la precisión en la toma de decisiones y reducir los resultados sesgados.

2.3.3 Ajuste

Se han creado algoritmos de mitigación de sesgos para lograr varios objetivos. Los algoritmos de mitigación de sesgo (también denominados a veces algoritmos justos) se pueden clasificar de la siguiente manera:

- Métodos basados en datos, como el muestreo ascendente de poblaciones subrepresentadas o el uso de datos sintéticos.
- Métodos basados en modelos, como la adición de términos de regularización o restricciones que imponen un objetivo durante la optimización, o el aprendizaje de la representación para ocultar o reducir el efecto de una variable específica.
- Métodos post-hoc, como identificar umbrales de decisión específicos del grupo basados en resultados previstos para igualar las tasas de falsos positivos u otras métricas relevantes.

Ejemplos de algoritmos de mitigación de sesgo que se aplican son:

- Removedor de impacto dispar: Una técnica de preprocesamiento que edita valores que se utilizarán como características de tal manera que se reduzcan los diferentes tratamientos entre los grupos;
- Detector y eliminador de sesgo individual: Una técnica que crea un nuevo modelo de ML para individuos en el grupo desfavorecido que reciben una decisión diferente en comparación con individuos similares en el grupo favorecido. A veces puede aplicar diferentes umbrales para la clasificación positiva entre los grupos;
- Clasificadores desacoplados: Una técnica para entrenar un clasificador separado en cada grupo. Los clasificadores separados podrían considerarse equivalentemente como un único clasificador que se ramifica en la función de grupo.

- Función de pérdida articular: Una técnica para capturar la paridad de grupo mediante el uso de una función de pérdida conjunta que penaliza las diferencias en las estadísticas de clasificación entre grupos.
- Aprendizaje por transferencia: Una técnica para mitigar los problemas de bajo volumen de datos para grupos donde hay una población de datos más pequeña.

2.4 Evaluación del sesgo

El sesgo en los sistemas de IA se mide de manera comparable a cómo se miden otras propiedades, como el rendimiento agregado. Sin embargo, es posible que las métricas de rendimiento agregadas en todo el conjunto de validación no indiquen si hay sesgo en el modelo. Las métricas generales en la matriz de confusión pueden parecer que funcionan bien en todo el conjunto. Sin embargo, calcular la precisión y el recuerdo en subconjuntos de categorías demográficamente importantes o ciertas a menudo puede revelar sesgos como una menor precisión para las mujeres que para los hombres, o una menor precisión para un grupo demográfico específico. Estas diferencias en el rendimiento probablemente indican que los sesgos no detectados están presentes en las primeras etapas del proceso de desarrollo. Por ejemplo, un determinado grupo podría estar infrarrepresentado en los datos de entrenamiento.

2.5 Métodos adversarios para mitigar el sesgo

Un método para mitigar el sesgo es incorporar una unidad adversarial en la arquitectura del modelo. En estos métodos, un "adversario" está prediciendo alguna propiedad o característica que define grupos hacia los cuales se desea la equidad. La salida del modelo para el cual se está mitigando el sesgo es la entrada al modelo adversarial. La actualización de peso para ese modelo se modifica para que, además de estar optimizada para la tarea que está realizando, también reduzca la cantidad de información que pone a disposición del adversario útil para su predicción. El efecto neto de este sistema es que el sistema aprende a realizar su tarea de manera ortogonal a las características para las cuales el sesgo no es deseado.

3. Verificación y validación

La verificación y validación de un modelo de ML recientemente desarrollado podría identificar y mitigar posibles sesgos antes de la implementación. Un conjunto de datos de retención obtenido de un origen de datos independiente de los datos de entrenamiento se utiliza normalmente en la comprobación y validación. Esta salvaguarda para la generalización del modelo también es importante para protegerse contra cualquier sesgo implícito en el conjunto de datos de entrenamiento. En general, cualquier paso tomado durante el procesamiento de conjuntos de datos y la capacitación de modelos sería beneficioso para aplicar a los datos y procedimientos de validación cuando corresponda.

Si bien la verificación de los sistemas de ML se lleva a cabo de forma intensiva utilizando conjuntos de datos de entrenamiento y prueba, se limita a la verificación de los resultados basada en la selección y variación de los datos disponibles. Es posible que sea necesario evaluar un sistema de IA en un contexto específico.

Las pruebas de software tradicionalmente se basan en un "cuerpo de conocimiento utilizado como base para el diseño de pruebas y casos de prueba". El éxito de cualquier

actividad de prueba empírica suele estar limitado por el grado en que los requisitos circundantes o los procesos de gestión de riesgos han identificado explícitamente posibles sesgos o fuentes de sesgo.

Las técnicas descritas en esta sección están destinadas a llevarse a cabo a una escala estadísticamente significativa. Las técnicas generalmente miden la sensibilidad del resultado a un subgrupo no incluido explícitamente en los datos de entrada.

Esta sección está destinada a aplicarse al desarrollo de nuevos sistemas, al despliegue de sistemas existentes y a la evaluación de si los sistemas mantienen la calidad a lo largo del tiempo. Un cambio en la relación entre los datos de entrada esperados y reales puede ser motivo de evaluación. La evaluación también puede incluir los resultados de la implementación en responsables del despliegue del sistema y espectadores (como personas u objetos que están presentes incidentalmente pero que no son el objetivo o el sujeto de una implementación del sistema de IA). Por ejemplo, un sistema que es injusto con respecto al género y la raza independientemente podría ser justo con una combinación específica de los dos.

3.1 Análisis estático de datos de entrenamiento y preparación de datos

El análisis de los datos de capacitación y entrada podría proporcionar información útil y detectar los tipos de sesgo de datos descritos en el Anexo C.1.

Los evaluadores pueden identificar el perfil de los datos de capacitación y entrada y validar si la propagación de una determinada variable representa el conjunto de datos real esperado. Un ejemplo de esto es identificar que los registros de un determinado grupo de edad se han utilizado para el entrenamiento, cuando se espera una distribución diferente de edades en conjuntos de datos reales. Esta actividad podría tener como objetivo validar el potencial de sesgo de selección, sesgo de muestreo y sesgo de cobertura, pero no puede hacerlo de manera exhaustiva, ya que está limitada por el conocimiento del evaluador.

Los evaluadores podrían identificar etapas en el proceso de preparación de datos que podrían introducir sesgos a través de "datos faltantes". Por ejemplo, si datos específicos no están disponibles de forma coherente en un conjunto de datos de entrada, los ingenieros pueden imputar esa información para los registros restantes o pueden eliminarla. Si la ausencia de ese elemento de datos se correlaciona con grupos específicos de registros, esto podría dar lugar a un sesgo que normalmente no se detectaría en las pruebas del modelo.

3.2 Muestras de controles de etiquetas

El riesgo de etiquetado incorrecto descrito en el Anexo C.1, es decir, etiquetadoras humanas que especifican incorrectamente las etiquetas para un conjunto de datos de entrada que luego se utilizan para entrenar el modelo, podría evaluarse mediante controles de muestra de las etiquetas presentadas.

El etiquetado basado en la opinión de expertos podría ser más complicado de evaluar. Puede ser necesario realizar revisiones doble ciego, o evaluar la evaluación de múltiples expertos, para evaluar la calidad de la etiqueta inicial.

3.3 Pruebas de validez interna

Las pruebas de validez interna evalúan la correlación entre los datos de entrada individuales y las salidas del sistema. Las pruebas de validez interna luego revisan si estas correlaciones son adversas en el contexto de requisitos específicos o criterios de aceptación.

Este proceso se basa en los datos que causan que se incluya un sesgo en el dominio de datos de entrada. Podría detectar sesgos en los modelos y su interacción.

Esto podría incluir la evaluación en un entorno totalmente integrado, con el fin de detectar cualquier sesgo en las actividades de recopilación o preparación de datos utilizadas durante el desarrollo del sistema de IA. La integración también puede detectar muestreos no representativos. Por ejemplo, los datos recopilados en un entorno integrado pueden tener características variables, como los niveles de iluminación o la frecuencia de actualización del sensor. Esas variaciones pueden influir en los datos de entrada al sistema de IA.

3.4 Pruebas de validez externas

Las pruebas de validez externas pueden implicar la reevaluación de observaciones previas utilizando fuentes de datos externas. Esta es una técnica útil porque puede detectar muchos tipos de sesgo que se han descrito en este documento, incluido el sesgo indirecto. El aspecto de los datos de entrada al que se refiere el sesgo indirecto no está contenido explícitamente dentro de los datos de entrada, sino que es una derivación de segundo orden.

Por ejemplo, algunos informes de los medios de comunicación sobre el sesgo de la IA se han centrado en la investigación que correlaciona los resultados del modelo con los datos del censo o del código postal con los resultados para ilustrar la disparidad de los resultados.

Las pruebas de validez externas también pueden incluir la integración de nuevos datos de entrada y la validación de que los resultados son coherentes con las pruebas de validez internas.

Las pruebas de validez externa son particularmente importantes para los sesgos indirectos introducidos por variables proxy. Si un diseñador de modelos intenta mitigar el sesgo simplemente eliminando la información demográfica de la entrada, es probable que el sesgo siga existiendo a través de variables proxy. Por ejemplo, el modelo podría perpetuar el racismo "daltónico", un concepto sociológico que describe cómo las afirmaciones de no "ver" el color de la piel de una persona impiden comprender y abordar la persistencia de la desigualdad racial en la sociedad. Para evitar este resultado, las pruebas de validez externas podrían necesitar incluir datos demográficos excluidos originalmente, o una exploración de los efectos de la variable proxy. Una investigación más profunda podría ser necesaria para comprender por qué existen tales proxies y si el propósito puede cumplirse sin ellos. Las pruebas de validez externas también podrían necesitar incluir datos cualitativos que demuestren impactos dispares de la misma clasificación. Por ejemplo, si se utiliza un determinado modelo para identificar a las personas antes de abordar un avión,

el daño emocional de un falso negativo podría ser mayor para aquellos grupos estereotipados como probables "terroristas" que para otros grupos.

En este contexto, podría ser necesario integrar conjuntos de datos de entrada con puntos de datos adicionales para evaluar adecuadamente el sesgo del sistema.

3.5 Pruebas de usuario

Las pruebas con diferentes tipos de usuarios finales pueden ser útiles cuando la interacción de un usuario con el sistema influye en los resultados y las predicciones de una manera correlacionada con la pertenencia del usuario a un grupo.

Evaluar la experiencia del usuario en escenarios del mundo real en un amplio espectro de usuarios, casos de uso y contextos de uso es una técnica útil para detectar sesgos en las interacciones del modelo, problemas de procesamiento de datos y problemas con las etiquetas de datos.

3.6 Pruebas exploratorias

Los desarrolladores de sistemas podrían organizar un grupo de probadores confiables y diversos que podrían probar adversariamente el sistema e incorporar una variedad de entradas potencialmente dañinas en pruebas unitarias o funcionales. Esto podría ayudar a descubrir formas imprevistas en que un sistema podría estar sesgado.

4. Despliegue

Una vez implementado, la capacitación y el soporte adecuados para el sistema de IA son importantes para que los responsables del despliegue permitan un uso efectivo del producto. Esto incluye orientación para los desarrolladores de sistemas sobre lo que constituye una implementación apropiada e inapropiada de un modelo dentro de un sistema de software. Por ejemplo, un sistema de seguimiento de la atención podría percibirse como injusto si se utiliza en un sistema educativo para monitorear el comportamiento de los estudiantes, pero ese podría no ser el caso si el mismo sistema se utiliza como herramienta de investigación en un experimento de psicología.

Los sistemas implementados también pueden incluir instrucciones para los usuarios finales del sistema. Por ejemplo, es deseable que los reclutadores que utilizan un sistema de recomendación de contratación comprendan las capacidades y limitaciones del sistema. Es posible que tanto los desarrolladores de sistemas como los usuarios finales de software deban ser conscientes de las áreas conocidas de sesgo. Esto se puede lograr a través de una herramienta de transparencia, que contiene información sobre los datos en los que se entrenó el modelo, distribuciones para poblaciones de sus errores falsos positivos y falsos negativos, y otra información asociada.

Los interesados, las personas a las que se refieren los datos de entrenamiento, no son necesariamente responsables del despliegue del sistema. No necesitan ser entrenados, pero es posible que necesiten ser informados de cualquier sesgo dentro del sistema que pueda afectarlos, en un lenguaje apropiado para el contexto. Los fallos en la capacitación o el soporte pueden dar lugar a sesgos adicionales que pueden ser difíciles de detectar antes.

4.1 Validación continua

Los modelos pueden degradarse y perder rendimiento con el tiempo. La degradación del rendimiento puede atribuirse a los cambios en el entorno, los nuevos comportamientos emergentes, los cambios en la composición de la población de insumos y los cambios en los requisitos. Además, un sistema podría estar sesgado hacia una posición histórica.

Es posible que sea necesario supervisar el rendimiento continuo cuando se implemente el sistema. Esto incluye verificar el rendimiento del sistema, especialmente los resultados atípicos, tanto manualmente como utilizando diferentes métricas y técnicas para evaluar el sesgo y la equidad. Si hay indicios de sesgo, es posible que sea necesario volver a entrenar o rediseñar el sistema.

El monitoreo es un proceso conocido en muchas industrias que aprovechan la toma de decisiones automatizada en sus procesos. Por ejemplo, en la banca, se están desarrollando e introduciendo modelos de cuadros de mando junto con sus procesos de monitoreo aprobados. Los procesos de monitoreo no solo podrían aplicarse a la precisión y rendimiento de modelos/sistemas, pero también podría utilizarse para la identificación y el seguimiento de sesgos en sistemas o modelos.

9. Referencias, estándares y normas

Para el desarrollo de esta guía se han consultado y utilizado especialmente las normas y estándares siguientes:

[1] ISO-IEC 38505-1 - Information technology - Governance of IT - Governance of data. Part 1: Application of ISO/IEC 38500 to the governance of data

[2] ISO-IEC 38507 - Information technology - Governance of IT - Governance implications of the use of artificial intelligence by organizations

[3] ISO-IEC 5338 - Information technology - Artificial intelligence - AI system life cycle processes

[4] ISO-IEC 24027 - AI Algorithmic bias - Information technology - Artificial Intelligence (AI) - Bias in AI systems and AI-aided decision making

[5] ISO-IEC 5259-1 - Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 1: Overview, terminology, and examples

[6] ISO-IEC 5259-2 - Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 2: Data quality measures

[7] ISO-IEC 5259-3 - Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 3: Data quality management requirements and guidelines

[8] ISO-IEC 5259-4 - Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 4: Data quality process framework

[9] ISO-IEC 5259-5 - Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 5: Data quality governance

[10] ISO-IEC 8183 - Information technology — Artificial intelligence — Data life cycle framework

[11] ISO-IEC 42001 - Information technology - Artificial intelligence - Management system

[12] ISO-IEC 22989 - Information technology - Artificial intelligence - Artificial intelligence concepts and terminology

[13] prEN 18229 AI Trustworthiness Framework

[14] prEN XXX Quality and governance of datasets in AI

[15] prEN XXX Concepts, measures and requirements for managing bias in AI Systems

[16] Especificación UNE 0085:2024

[17] VERMA, Sahil and RUBIN, Julia, 2018. Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness - FairWare '18 [online]. Gothenburg, Sweden: ACM Press. 2018. p. 1–7. [Accessed 6 May 2019]. Fairness definitions explained | Proceedings of the International Workshop on Software Fairness (acm.org)

[18] MITCHELL, Shira, POTASH, Eric, BAROCAS, Solon, D'AMOUR, Alexander and LUM, Kristian, 2020. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. arXiv:1811.07867 [stat] [online]. 24 April 2020. [1811.07867] Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions (arxiv.org)

[19] GAJANE, Pratik and PECHENIZKIY, Mykola, 2018. On Formalizing Fairness in Prediction with Machine Learning. arXiv:1710.03184 [cs, stat] [online]. 28 May 2018. [Accessed 30 October 2020]. [1710.03184] On Formalizing Fairness in Prediction with Machine Learning (arxiv.org)

[20] AGARWAL, Alekh, DUDÍK, Miroslav and WU, Zhiwei Steven, 2019. Fair Regression: Quantitative Definitions and Reduction-based Algorithms. arXiv:1905.12843 [cs, stat] [online]. 29 May 2019. [Accessed 30 October 2020]. [1905.12843] Fair Regression: Quantitative Definitions and Reduction-based Algorithms (arxiv.org)

[21] Confusion matrix, 2020. Wikipedia [online]. [Accessed 30 October 2020]. Confusion matrix - Wikipedia

[22] Agencia Española de Protección de Datos. "Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una Introducción". 2020.

[23] Agencia Española de Protección de Datos. "Requisitos para Auditorías de Tratamientos que incluyan IA". 2021.

[24] ENISA. Tecnicas Seudonimizacion Sector Sanitario. 2022.

[25] Asociación Española de Normalización. Guía de evaluación del Gobierno, Gestión y Gestión de la Calidad del Dato (UNE 0080).

[26] Asociación Española de Normalización. Gestión de Calidad del Dato (UNE 0079).

[27] Asociación Española de Normalización. Gestión del Dato (UNE 0078).

[28] Asociación Española de Normalización. Gobierno del Dato (UNE 0077).

La versión del Reglamento Europeo de la IA que se ha tomado como referencia ha sido la publicada por el Consejo de la Comisión Europea el 13 de junio de 2024.

Financiado por
la Unión Europea
NextGenerationEU

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

Plan de
Recuperación,
Transformación
y Resiliencia

España | digital 2026