https://datascience-project.eu/

Ficha de formación

Título	Ciencia de datos e impacto social: Conseguir resultados positivos
Palabras clave	Impacto social, Data for Good, métricas de equidad, seguimiento de medios sociales
Idioma	Español
Objetivos /Metas /Resultados de aprendizaje	**1. El uso de la ciencia de datos para el bien social** **2. Comprender los principales riesgos de la tecnología y ser capaz de nombrar ejemplos** **3. Ser capaz de enumerar las características de la "IA de confianza"** **4. Comprender los retos de medir la imparcialidad**
Curso de formación:	
Alfabetización en ciencia de datos	
Módulo de visualización de datos y análisis visual	
Introducción a la ciencia de datos para las ciencias humanas y sociales	
Ciencia de datos para el bien social	X
Periodismo de datos y Storytelling	
Descripción	En este curso, echaremos un vistazo a las muchas aplicaciones de la ciencia de datos que pueden hacer del mundo un lugar un poco mejor. A continuación, entraremos en detalle en el seguimiento de las redes sociales realizado en nombre de Amnistía Internacional Italia para comprender cómo puede funcionar una aplicación de este tipo. En la siguiente sección, exploraremos algunos de los efectos perjudiciales que pueden tener la ciencia de datos y la IA. Esto nos ayudará a comprender por qué es necesario que los sistemas de IA sean fiables. Por último, nos familiarizaremos con algunos de los retos de las métricas de imparcialidad y veremos lo que estas métricas pueden significar en la práctica.
Contenidos organizados en 3 niveles	**1. <u>Utilizar la ciencia de datos para el bien social</u>** A través de diferentes casos de uso, especialmente el de "Amnistía Italia", obtendrás una visión general de cómo la ciencia de datos puede utilizarse con buenos fines. **1.1 Visión general de la ciencia de datos para posibles casos de buen uso** La mejor manera de entender el impacto positivo que la ciencia de datos puede tener en las personas y el planeta es observar algunos ejemplos del pasado reciente. El rápido ritmo del cambio tecnológico también está provocando cambios en el mercado laboral: los antiguos empleos y profesiones están desapareciendo y están siendo sustituidos por otros nuevos. Esto provoca desempleo en algunos sectores, mientras que en otros los empresarios tienen dificultades

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

para encontrar empleados cualificados. Pero, de hecho, muchas competencias obtenidas en los sectores "en desaparición" podrían adaptarse y reutilizarse fácilmente en nuevos sectores. En el proyecto piloto SkillsFuture Singapur, la ciencia de datos se utiliza para detectar estas competencias "reutilizables" y ayudar a los desempleados con formación específica para que puedan adaptar sus competencias a las necesidades de los sectores industriales en expansión.

La IA también puede utilizarse para mejorar la capacidad predictiva de los gemelos digitales, por ejemplo para ayudar a que la cadena de suministro sea más resistente. Los gemelos digitales utilizan los datos de que dispone una empresa -ya sean datos generados internamente a través de procesos operativos, transaccionales o de otro tipo, o datos disponibles públicamente como la vigilancia meteorológica- para simular la cadena de suministro. A estos gemelos digitales pueden añadirse sistemas de IA entrenados con aprendizaje de refuerzo, lo que permite a las empresas explorar los efectos de varios escenarios hipotéticos, como el impacto de un tornado, y desarrollar medidas para reaccionar ante tales escenarios [2].

Los sistemas de IA pueden utilizarse de diversas formas para alcanzar los objetivos climáticos. Por ejemplo, Fero Labs utiliza la IA para ayudar a los fabricantes de acero a reducir el uso de ingredientes extraídos hasta en un 34%, lo que evita unas 450.000 toneladas de emisiones de CO2 al año, mientras que el proyecto Mapping the Andean Amazon utiliza la IA para vigilar la deforestación a través de imágenes por satélite para ayudar a descubrir la deforestación ilegal y apoyar las respuestas políticas [3].

Uno de los retos asociados a los vehículos eléctricos es que necesitan acceder a infraestructuras eléctricas específicamente diseñadas para ellos, es decir, estaciones de recarga. Si muchos coches necesitan la misma infraestructura al mismo tiempo, esto puede suponer un reto importante para la red eléctrica. Yendo más lejos, uno de los obstáculos para la adopción a gran escala de fuentes de energía renovables es la gran fluctuación en la disponibilidad de energía y la capacidad limitada para almacenar electricidad en los momentos de máxima disponibilidad, para luego dispensarla en los momentos de mayor uso. Las tecnologías "del vehículo a la red", que permiten utilizar los coches eléctricos como "almacén" del excedente de energía y que la red extraiga energía de los coches cuando éstos no estén en uso, pueden ayudar a mitigar el problema. Gracias a la inteligencia artificial, Caltech ha desarrollado un sistema de carga adaptable que programa cuándo cargar cada vehículo y cuándo y cuánta energía se puede devolver a la red, en función de las horas de salida indicadas por el conductor. De este modo se reduce la presión global sobre la red eléctrica y se abre la interesante posibilidad de que los coches eléctricos alivien parte de la carga que soportan las redes eléctricas [4].

Co-funded by the Erasmus+ Programme of the European Union

Las cadenas de suministro son increíblemente complejas, lo que supone un reto para legislaciones como la estadounidense Uyghur Forced Labor Prevention Act, que pretende imponer normas sociales o medioambientales más estrictas en los productos. El Atlas Altana combina información geolocalizada sobre la ubicación y las instalaciones de las empresas con datos sobre la propiedad corporativa para trazar las relaciones comerciales entre sectores. Esto ayuda a las empresas a cumplir más eficazmente dicha legislación y a tomar medidas por su cuenta contra problemas como el trabajo forzoso [5].

Los aerogeneradores son una importante fuente de energía renovable, pero su producción depende de un factor difícil de controlar: el viento. Esto supone un reto para la red energética, pero también para el departamento de ventas de los proveedores de energía eólica, ya que la energía que es más predecible también puede alcanzar precios más altos. Para respaldar el argumento comercial de los parques eólicos, DeepMind desarrolló una red neuronal entrenada con previsiones meteorológicas y datos operativos históricos que puede predecir la producción del parque eólico con 36 horas de antelación, logrando así un valor un 20 % superior para la energía producida [6].

1.2 Caso práctico de Amnistía Italia

Las redes sociales son una parte importante de la esfera pública. Para investigar cómo evoluciona el discurso político sobre cuestiones relacionadas con los derechos humanos y cómo repercute en los grupos desfavorecidos, Anmistía Internacional Italia lleva a cabo cada año un seguimiento denominado Barómetro del Odio (Barometre dell'Odio) utilizando técnicas de ciencia de datos.
Los datos se recopilan a través de las API públicas de Facebook y Twitter, a partir de una lista de cuentas y perfiles públicos facilitada por Amnistía. Normalmente, el periodo de seguimiento abarca entre cuatro y ocho semanas (en 2021 se amplió a 16 semanas). Durante este periodo, se toman muestras aleatorias de los comentarios de las cuentas más activas, lo que supone un conjunto de entre 30.000 y 100.000 comentarios, que son etiquetados por voluntarios formados de Amnistía en relación con el tema y el nivel de ofensividad. Todas las etiquetas se verifican de forma cruzada, lo que significa que cada comentario es etiquetado por dos o tres voluntarios y cualquier incoherencia es resuelta por el Consejo del Odio de Amnistía (Tavolo dell'Odio).

Ejemplo: Elecciones al Parlamento Europeo 2019

En el periodo previo a las Elecciones al Parlamento Europeo 2019, perfiles públicos de 461 candidatos en Twitter y Facebook en las seis semanas anteriores a las elecciones (15 de abril - 24 de mayo de 2019). En total, se

Co-funded by the Erasmus+ Programme of the European Union
With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

recopilaron inicialmente 27.000 publicaciones y 4 millones de comentarios. En un segundo paso, hubo que reducir el tamaño del conjunto de datos para hacerlo manejable para los voluntarios en función del alcance de la actividad en redes sociales de los perfiles, al tiempo que se garantizaba la representación general de todos los partidos, todas las regiones y al menos una mujer y un hombre por partido. De este modo, el conjunto de datos final incluía actividades en medios sociales relacionadas con 77 políticos: el 80% de las publicaciones fueron etiquetadas por 150 voluntarios de Amnistía junto con una muestra aleatoria de 100 mil comentarios.
Los resultados [8] muestran que el discurso del odio no se distribuye aleatoriamente, sino que se agrupa. Aunque se calcula que su prevalencia global en las plataformas de redes sociales es de alrededor del 1%, es más probable que se produzca en relación con grupos y temas específicos, y alcanza su punto máximo en determinados momentos. Por ejemplo, la incitación al odio es más probable cuando se habla de migración, romaníes, minorías religiosas o mujeres.
Profundizando en los datos, también se pueden observar ciertos patrones. El discurso de odio acumula más discursos de odio, pero también es más probable que reciba interacciones (como reacciones, comparticiones o comentarios). También puede utilizarse para excluir activamente a personas de las plataformas de las redes sociales: por ejemplo, durante la campaña de seguimiento de 2020, se observó cómo dos mujeres eran objetivo específico del discurso del odio y tres fueron expulsadas de las plataformas de las redes sociales [9].

2. La ciencia de datos no siempre es buena

Por desgracia, al igual que cualquier otra tecnología, la IA y la ciencia de datos también pueden utilizarse con malos fines o tener consecuencias no deseadas. Sin embargo, a diferencia de otras herramientas, la IA automatiza las decisiones por nosotros y, por lo tanto, tiene un potencial aún mayor de causar daños. Por lo tanto, también debemos ser conscientes de que la IA y la ciencia de datos pueden tener un impacto negativo en los seres humanos, la sociedad y el medio ambiente.

2.1 Principales ejemplos conocidos

El objetivo de la ciencia de datos es ayudarnos a tomar mejores decisiones basadas en datos, haciendo posible procesar grandes cantidades o tipos muy diversos de información. Como hemos visto antes, la ciencia de datos puede utilizarse para controlar o mejorar procesos que contribuyan a hacer del mundo un lugar mejor. Sin embargo, la historia reciente nos ha demostrado que no podemos confiar ciegamente en los resultados de los algoritmos, especialmente cuando estos resultados pueden tener un grave impacto negativo en nuestras vidas.
Ejemplos bien conocidos de tales impactos negativos se produjeron en aplicaciones de la IA que van desde la sanidad al trabajo pasando por el medio ambiente:

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

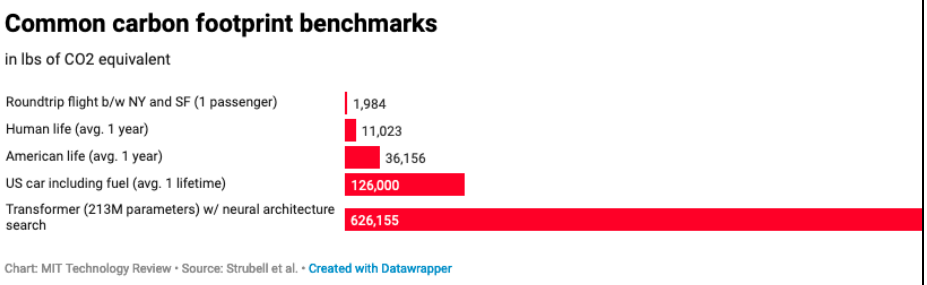
1. Los hospitales de EE.UU. recurren ahora a algoritmos que les ayudan a evaluar el grado de enfermedad de los pacientes para determinar si necesitan atención hospitalaria o ambulatoria. Un estudio descubrió que las evaluaciones de un sistema muy utilizado estaban sesgadas por motivos raciales: De hecho, los pacientes negros estaban más enfermos que los blancos que habían recibido la misma calificación de riesgo. Esto se debía probablemente a que el algoritmo utilizaba los costes sanitarios históricos como indicador de las necesidades sanitarias; sin embargo, dado que el sistema sanitario estadounidense ha estado históricamente plagado de desigualdad de trato, se gastaba menos dinero en cubrir las necesidades sanitarias de los pacientes negros. Por tanto, el algoritmo concluyó erróneamente que están más sanos que los pacientes blancos, que en realidad están igual de enfermos [10].
2. Amazon creó una herramienta de contratación con IA para ayudar al Departamento de Recursos Humanos a encontrar el personal adecuado para puestos técnicos, y la entrenó con los currículos enviados a la empresa durante los diez años anteriores. Sin embargo, dado que la mayoría de esas solicitudes procedían de hombres, Amazon pronto se dio cuenta de que su sistema de contratación no estaba puntuando a los candidatos de forma neutral desde el punto de vista del género. El sistema de IA penalizaba los currículos presentados por mujeres y que contenían palabras como "de mujer". El programa tuvo que ser retirado y hasta ahora no se ha vuelto a instalar [11].
3. Ya en 2015, el clasificador de imágenes de Google etiquetó a una persona negra como "gorila". Google se disculpó, pero optó por una solución rápida limitándose a censurar "gorila", "chimpancé", y "mono" de las búsquedas y etiquetas de imágenes. Seis años después, Facebook clasificó a un hombre negro en un vídeo como primate, recomendando a los usuarios que siguieran viendo vídeos de primates. [12]

Estos son solo algunos ejemplos que ilustran las posibles repercusiones negativas. La ciencia de los datos y la inteligencia artificial necesitan datos y, a menudo, estos datos son etiquetados o procesados de otro modo por trabajadores mal pagados, que trabajan en condiciones muy estresantes y a menudo están expuestos a contenidos violentos o perturbadores [13]. Los algoritmos pueden utilizarse para clasificar a empleados o contratistas de forma discriminatoria y con pérdida de oportunidades [14]. La ciencia de datos y la IA son costosas desde el punto de vista computacional, lo que significa que también consumen muchos recursos; esto es especialmente cierto en el caso de los modelos de gran tamaño y los modelos de ajuste fino,

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

como los transformadores incluidos en el gráfico comparativo que figura a continuación [15].



Ejercicio: Si quieres convertirte tú misma en una detective de prejuicios, sólo tienes que ir a Google Translate (o deepl.com) y traducir del inglés al alemán:
Inglés: Mi médico es inteligente. Inmediatamente encontró la solución
Google Alemán:

Inglés: Mi secretaria es inteligente. Encontró inmediatamente la solución
Google Alemán:

Google intentó solucionar este problema en 2018 tras una gran protesta por traducir a roles de género estereotipados desde idiomas de género neutro, pero como puedes descubrir por ti mismo, cinco años después, los problemas persisten.

2.2. Panorama de los principales riesgos

Desde usar bots para crear desnudos falsos en Telegram, generar avatares sexualizados de mujeres (pero no de hombres), no desarrollar funcionalidades útiles para un grupo específico de personas o socavar la identidad de género a través de la clasificación binaria, las aplicaciones de la ciencia de datos pueden causar daño.

Uno de los principales riesgos de la IA y la ciencia de datos es que suponemos que la propia tecnología -como cualquier otra herramienta- está libre de juicios y errores humanos. Sin embargo, en esta teoría parecemos olvidar que somos nosotros los que creamos estos sistemas, los que elegimos los algoritmos, los que seleccionamos los datos y los que decidimos cómo utilizar y a quién desplegar el sistema. Por lo tanto, es fundamental comprender que las aplicaciones de la ciencia de datos -incluso con las mejores intenciones- no son ni objetivas, ni neutrales.

Reflexiona sobre lo que puede hacer tu aplicación, para qué se utiliza, a quién incluye/excluye y a quién puede afectar de distintas maneras: ¡las consecuencias pueden ser muy amplias!

En su estudio de 2018 [15], Joy Buolamwini y Timnit Gebru descubrieron que los algoritmos de clasificación de género que utilizan el reconocimiento facial clasifican erróneamente de forma rutinaria a las mujeres de piel más oscura con más frecuencia que a los hombres (y mujeres) de piel más clara. Esto se debe a que los conjuntos de datos en los que se entrenaron los modelos

Co-funded by the Erasmus+ Programme of the European Union
With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	investigados contenían una parte desproporcionada de imágenes de hombres y mujeres de piel clara. Dos estudios de 2019 mostraron que los algoritmos utilizados para detectar expresiones ofensivas en plataformas on line eran más propensos a clasificar como ofensivos patrones de expresión comunes entre los estadounidenses de raza negra, y los conjuntos de datos mostraban de forma similar un sesgo generalizado contra el inglés afroamericano [16]. Esto demuestra lo importante que es etiquetar el conjunto de datos: si los datos se etiquetan de forma sesgada, los resultados también lo estarán. -> Debemos reconocer que las aplicaciones de la ciencia de datos no son perfectas, y que sus errores no se distribuyen aleatoriamente: de hecho, estos sistemas tienden a fallar con mayor frecuencia en el caso de grupos demográficos históricamente marginados o vulnerables. Además, las aplicaciones de ciencia de datos pueden ser muy intensivas en datos, lo que conlleva problemas de - Privacidad: Los modelos de IA que se basan en cada vez más datos incentivan la recopilación de datos en distintos ámbitos. Esto significa que se acaban recopilando muchos datos sobre las personas, con importantes implicaciones para la privacidad. Por ejemplo, aunque a veces puede ser práctico desde la perspectiva del consumidor saber dónde está exactamente su paquete en ese momento, y desde la perspectiva de un proveedor de servicios postales puede ser práctico disponer de esos datos para optimizar las rutas, rastrear el vehículo en el que se entrega un paquete también significa rastrear a la persona que conduce el vehículo. - Protección de datos: Muchos de los datos recopilados pueden permitirte identificar a personas y, por tanto, se consideran datos de identificación personal, como el ejemplo del seguimiento de paquetes que acabamos de comentar. Estos datos no solo pueden utilizarse indebidamente más adelante, sino que también pueden utilizarse para restringir sus oportunidades, razón por la cual el Reglamento General de Protección de Datos de la UE establece una estricta política de minimización de datos. - Mala calidad de los datos: Es posible que hayas oído hablar de la frase "basura dentro, basura fuera" para describir cómo la mala calidad de los datos puede conducir a malos resultados. Esto significa que el simple hecho de tener muchos datos no mejorará su modelo ni sus resultados. Al contrario, un gran conjunto de datos mal etiquetado, mal procesado y lleno de datos irrelevantes empeorará los resultados. Ten en cuenta que la mayor parte del tiempo invertido en proyectos de ciencia de datos e IA se dedica a crear un conjunto de datos de alta calidad que pueda utilizar de forma fiable y repetida. ¡Haz que ese esfuerzo cuente!

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

- Con el fin de contrarrestar los riesgos derivados de la ciencia de datos y la IA, hasta la fecha se han desarrollado más de 80 directrices éticas diferentes: entre las más destacadas se encuentran las emitidas por organizaciones internacionales como la OCDE, la UNESCO, UNICEF; pero también por grandes empresas tecnológicas, como Google y Microsoft.

El problema de estas normas éticas es que no son jurídicamente vinculantes ni aplicables: su incumplimiento no tiene consecuencias. Las normas éticas nos ayudan a establecer la dirección correcta y nos orientan sobre lo que está mal y lo que está bien, pero el carácter voluntario de estas iniciativas significa que son algo que está bien tener, en lugar de algo que hay que hacer.

3. IA de confianza

En esta sección, examinaremos las características de la llamada "IA de confianza", analizaremos de dónde procede esta noción y por qué es importante. Nos centraremos en el tema de los sesgos no deseados que pueden conducir a la discriminación y en las formas de medir la imparcialidad con la ayuda de una matriz de confusión.

3.1 IA de confianza

La Unión Europea también ha creado sus propias normas éticas, las denominadas "Directrices éticas para una inteligencia artificial digna de confianza" [17]. Un documento elaborado por el Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial (AI HLEG), un grupo de expertos independientes que fue creado por la Comisión Europea en junio de 2018, como parte de la estrategia de IA de la UE.

El Grupo de Alto Nivel de la UE estableció las siguientes características de un sistema de IA fiable, basado en la Carta de los Derechos Fundamentales de la UE [1]:

(1) agencia y supervisión humanas: Los sistemas de IA deben ser comprensibles para los seres humanos en la medida en que sus decisiones puedan ser cuestionadas, y los seres humanos siempre deben poder intervenir en los sistemas de IA.

(2) solidez técnica y seguridad: Los sistemas de IA deben ser capaces de hacer frente a una serie de situaciones con las que podrían encontrarse razonablemente, así como a ataques malintencionados, y deben diseñarse teniendo en cuenta la seguridad y la protección.

(3) privacidad y gobernanza de datos: Los sistemas de IA no deben socavar el derecho a la intimidad de nadie, los titulares de los datos deben tener pleno control sobre cómo se utilizan sus datos y éstos no deben utilizarse para perjudicar o discriminar a los titulares de los datos. Además, debe existir un

[1] The Charter of Fundamental Rights of the European Union brings together the most important personal freedoms and rights enjoyed by citizens of the EU into one legally binding document. See, for example, https://fra.europa.eu/en/eu-charter

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

sistema adecuado de gobernanza de datos que garantice que el conjunto de datos es de alta calidad y no se puede acceder a él con fines ilegítimos.

(4) Transparencia: las decisiones tomadas por los sistemas de IA deben ser trazables y explicables a los humanos, y los límites del sistema de IA deben comunicarse con claridad.

(5) Diversidad, no discriminación e imparcialidad: los conjuntos de datos sesgados causan problemas, pero también son perjudiciales los modelos sesgados o los sistemas de IA que tienen efectos desproporcionados en grupos específicos y, por lo general, desfavorecidos. Por este motivo, la diversidad de representación y la participación en todas las fases del ciclo de desarrollo de la IA son fundamentales para detectar posibles daños en una fase temprana y desarrollar mecanismos adecuados de prevención y mitigación.

(6) bienestar medioambiental y social: Los sistemas de IA tienen un impacto real en la sociedad y en el medio ambiente, no sólo en los individuos. Esto significa que, en algunos ámbitos, el uso de los sistemas de IA debería estar bien reflexionado, y que todos los sistemas de IA deberían diseñarse de forma sostenible desde el punto de vista medioambiental y social.

(7) Rendición de cuentas: Los sistemas de IA deben ser auditables y los posibles efectos negativos, así como las compensaciones, deben identificarse y abordarse de antemano, ofreciendo la posibilidad de una reparación efectiva si se causa algún daño.
Aunque la guía HLEG de la UE va un paso más allá de las simples directrices éticas, al basar los principios en la Carta de los Derechos Fundamentales de la UE (un marco jurídico), en la próxima sección veremos, basándonos en el ejemplo de la equidad y la no discriminación (principio 5), que aún queda mucho camino por recorrer desde los principios hasta su aplicación.

3.2. Prejuicios, equidad, no discriminación
Todos tenemos derecho a un trato justo. Pero, ¿qué se entiende por equidad? Fundamentalmente, la imparcialidad es un concepto subjetivo y depende de la cultura y el contexto. En un intento de eludir esta delicada cuestión, muchas investigaciones se han centrado en el problema de la parcialidad en la IA.
Sin embargo, en el contexto de la ciencia de datos y el aprendizaje automático en general, chocan muchas definiciones diferentes de sesgo (uso coloquial vs. Estadística vs. aprendizaje profundo). Esto es un problema porque personas de diferentes disciplinas hablan de sesgo, pero en realidad no se refieren a lo mismo. En el contexto de la IA fiable, entenderemos por sesgo un prejuicio que favorece a un grupo en detrimento de otro.

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Existen muchos tipos diferentes de sesgos, como el sesgo social, el sesgo de confirmación, el sesgo intragrupo, el sesgo de automatización, el sesgo temporal, el sesgo de variable omitida, el sesgo de muestreo, el sesgo de representación, el sesgo de medición, el sesgo de evaluación y muchos más. Todos estos sesgos -en los datos, en el sistema de IA o derivados de la interacción de seres humanos con prejuicios con el sistema de IA- pueden conducir a un trato injusto y a la discriminación, es decir, al trato injusto o perjudicial de diferentes categorías de personas, por ejemplo, por motivos de etnia, edad, sexo o discapacidad.
Pero, ¿cómo detectar y medir el sesgo?
El primer paso es comprobar la calidad de los datos, que es una de las formas más comunes de que el sesgo se cuele en el conjunto de datos. Pero incluso si no hay defectos en los datos, el modelo puede estar sesgado.
Normalmente sólo se puede detectar el sesgo como un efecto sobre el resultado del modelo. Esto se hace con la llamada Métrica de Equidad, que es el tema de la siguiente sección. Como puede ver, el intento de evitar la definición de imparcialidad mediante el análisis del sesgo no llegó muy lejos.

3.3. Métrica de equidad
Dado que no existe una definición única y perfecta de imparcialidad, tampoco existe una única métrica correcta para medir la imparcialidad, y es imposible una solución única para todos los casos. En su lugar, hay muchos tipos diferentes de imparcialidad y formas de medirla, incluyendo la imparcialidad de grupo, la paridad estadística condicional, el equilibrio de la tasa de errores falsos positivos, el equilibrio de la tasa de errores falsos negativos, la igualdad de precisión de uso condicional, la igualdad de precisión global, la imparcialidad de prueba, la buena calibración, la imparcialidad a través de la falta de conocimiento, la imparcialidad contrafactual y muchos más.

Desgraciadamente, no se puede simplemente probarlas todas para asegurarse de que el algoritmo es justo, ya que es probable que estas métricas conduzcan a resultados contradictorios. Por ejemplo, es matemáticamente imposible cumplir los requisitos tanto de paridad predictiva como de probabilidades igualadas. Consideremos la siguiente fórmula, derivada en [18]:

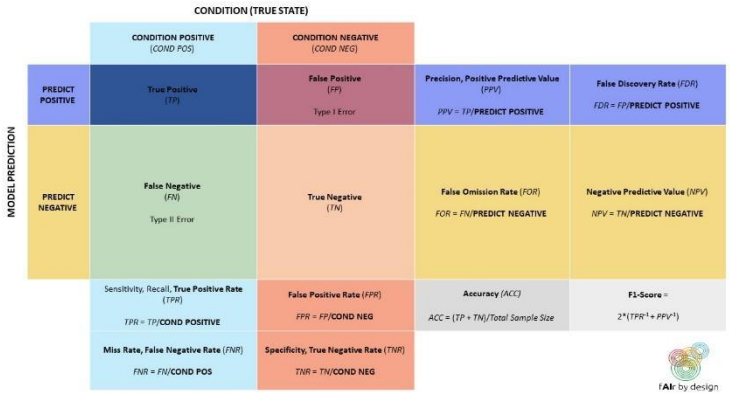
FPR = (1 – FNR)

La p de la fórmula se refiere a la prevalencia de la clase POSITIVA, y puede utilizar la matriz de confusión que aparece a continuación para entender los demás términos. Supón ahora que tienes dos grupos demográficos, G1 y G2, con prevalencia p1 y p2. Si se mantiene la igualdad de probabilidades, entonces FPR y FNR son iguales para ambos grupos. Si se mantiene la paridad predictiva, entonces también el VPP es el mismo para ambos grupos.

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Si introducimos toda esta información en la fórmula anterior, obtendremos dos ecuaciones, una para G1 y otra para G2. Un poco de álgebra te mostrará entonces que p1 y p2 también deben ser iguales.
En resumen: si tanto las probabilidades igualadas como la paridad predictiva son ciertas, entonces la prevalencia **debe** ser la misma para ambos grupos. Por el contrario, si la prevalencia no es la misma para ambos grupos, ¡entonces las probabilidades igualadas y la paridad predictiva **no pueden** ser verdaderas!



La imposibilidad matemática de satisfacer todas las métricas de imparcialidad simultáneamente significa que hay que decidir qué definición de imparcialidad debe aplicarse. Lamentablemente, en la actualidad no existe un marco jurídico ni ejemplos de buenas prácticas, lo que significa que hay que considerar detenidamente el contexto de la aplicación de IA antes de elegir la métrica para evaluar su impacto en términos de imparcialidad.
Para entender las implicaciones de tener múltiples definiciones de imparcialidad que no son compatibles, y la importancia de ponerse de acuerdo sobre una definición antes de desplegar tales sistemas, echaremos un vistazo a un ejemplo de la vida real que desencadenó gran parte de la investigación y el debate sobre el sesgo en los algoritmos en la comunidad de la ciencia de datos y ML.

COMPAS es un sistema de IA desarrollado por una empresa llamada Northpointe, y se utiliza en el sistema de justicia penal de Estados Unidos para estimar el riesgo de reincidencia de un acusado (en otras palabras, para calificar el riesgo de un acusado de cometer otro delito en el futuro). Esta puntuación de riesgo se utiliza después para tomar decisiones sobre la libertad condicional o la puesta en libertad anticipada.
Para producir sus resultados, el sistema de IA se basó en los registros históricos de delincuencia, que rastreaban a los delincuentes pasados y si habían vuelto a ser detenidos por otro delito tras su puesta en libertad; es decir, contenía información sobre la probabilidad de que determinados tipos de acusados volvieran a delinquir (¡y de que los pillaran haciéndolo!). Estos

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

registros se utilizaron para entrenar el modelo de predicción del riesgo de reincidencia de los acusados que no formaban parte del conjunto de datos, una vez que el sistema entró en funcionamiento. Esto significa que se calculó la probabilidad de reincidencia de cada acusado y se les clasificó como de bajo o alto riesgo.

En mayo de 2016, ProPublica publicó un artículo en el que indicaba que las predicciones de este modelo de modelización de la reincidencia estaban sesgadas [18; véase también 19, 20]: ProPublica demostró que la fórmula del sistema de IA era particularmente propensa a marcar falsamente a los acusados negros como de alto riesgo de reincidencia, etiquetándolos erróneamente de esta manera en casi el doble de la tasa que a los acusados blancos (42% frente a 22%); al mismo tiempo, los acusados blancos fueron etiquetados erróneamente como de bajo riesgo con más frecuencia que los acusados negros.[2]

Si nos fijamos en la matriz de confusión anterior, podemos ver que ProPublica estaba diciendo que COMPAS era injusto porque FPR y FNR no eran iguales para los acusados negros frente a los acusados blancos. Resulta que se trata de la métrica de equidad Equalized Odds:

1. Equalized odds

Equalized odds significa que, dentro de cada categoría de riesgo verdadero, el porcentaje de predicciones falsas negativas y falsas positivas es igual para cada grupo demográfico. La pregunta ya no se centra en la precisión global del modelo, sino que desglosa los tipos de error que puede cometer el modelo (falsos positivos y falsos negativos), y exige que los errores del modelo sean comparables: FPR es igual en todos los grupos, y FNR es igual en todos los grupos.

Northpointe defendió su sistema COMPAS contra la acusación de parcialidad, señalando que, si el modelo predecía que un acusado era de alto riesgo, la probabilidad de que reincidiera era la misma, independientemente del grupo demográfico al que perteneciera el acusado. Northpointe está diciendo: la probabilidad de un verdadero positivo, dado que el modelo predijo positivo, es la misma para todos los grupos. Esto se conoce como métrica de equidad de Paridad Predictiva.

2. Paridad Predictiva

La paridad predictiva significa que la proporción de acusados de alto riesgo correctamente predichos es la misma con independencia del grupo demográfico. En otras palabras, la paridad predictiva se refiere al concepto en ML e IA de que los modelos predictivos utilizados deben producir el mismo Valor Predictivo Positivo (VPP) para diferentes grupos, independientemente de su pertenencia a una clase protegida (por ejemplo, raza, sexo, edad, etc.). El VPP es una métrica utilizada para evaluar la proporción de predicciones positivas verdaderas (instancias positivas correctamente clasificadas) entre todas las instancias en las que el modelo

[2] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

predijo positivo. Sin embargo, esta métrica no tiene en cuenta la prevalencia global de los casos en un conjunto de datos.
Para decirlo de otro modo, la paridad predictiva considera la imparcialidad teniendo en cuenta los errores relativos a la clase predicha, mientras que las probabilidades igualadas tienen en cuenta los errores relativos a la clase verdadera. Si es más importante optimizar el VPP (y, por tanto, preferiría la equidad de la paridad predictiva), o si prefiere minimizar el RPF (y, por tanto, preferiría las probabilidades igualadas) es en gran medida una cuestión de perspectiva. Por ejemplo, ¿qué métrica de error es más importante para ti si has recibido un diagnóstico médico de un sistema de IA? ¿Y qué métrica de error es más importante en un algoritmo de contratación utilizado para contratar para un puesto de trabajo al que te has presentado? ¿Se te ocurren situaciones en las que considerarías más importante el VPP y otras en las que preferirías un VPF bajo?
Si quieres saber más sobre las distintas definiciones de imparcialidad (en realidad, actualmente hay más de 21), cómo medirlas y las diferencias entre ellas, consulta "Definiciones de imparcialidad explicadas" [22].
Reflexionar: Volviendo al ejemplo del COMPAS, ¿qué definición considerarías justa?
Pregunta: ¿Es posible satisfacer ambas definiciones de justicia?
Respuesta: Debemos comprender la prevalencia de la reincidencia. En EE.UU., la tasa global de reincidencia de los acusados negros es superior a la de los acusados blancos (52% frente a 39%). Según la fórmula que vimos anteriormente, esto significa que no es posible que se cumplan las dos definiciones de equidad.
Este caso del COMPAS ejemplifica cómo las cuestiones sociales influyen en los datos disponibles en primer lugar. El exceso de vigilancia de las comunidades negras hace que la probabilidad de que se produzcan detenciones o se registren incidentes sea mayor en estas comunidades. Como resultado, se introducen datos sesgados en los modelos. Y lo que es aún más sutil, esto significa que el índice de reincidencia percibido para las dos poblaciones no es el mismo, lo que obliga a tomar decisiones muy difíciles sobre qué métrica de justicia utilizar, es decir, qué es justo en este contexto.
El problema real es que existen sesgos sistémicos en el sistema judicial y de aplicación de la ley (en EE.UU., pero también en otros lugares), que no pueden solucionarse simplemente introduciendo más datos (casos históricos) en el sistema. En "Dirty Data, Bad Predictions. How Civil Rights Violations Impact Police Data: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice" [22].
Los sesgos sistémicos también afectan a otros ámbitos de aplicación, ya se trate de la sanidad, la educación o el precio que se paga por productos o servicios. A veces, podemos elegir las herramientas adecuadas para tener en cuenta esos sesgos sistémicos. Y, a veces, tenemos que admitir que no se dan las condiciones adecuadas para un uso seguro de los algoritmos. Estas

Co-funded by the Erasmus+ Programme of the European Union
With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	decisiones, sin embargo, no deben dejarse solo en manos del científico de datos, sino que deben implicar a una multitud de partes interesadas y muchos conocimientos diferentes, incluyendo, por ejemplo, la sociología, la psicología, el derecho y los expertos en dominios específicos del contexto. La IA y la ciencia de datos no pueden hacer milagros y resolver nuestros problemas sociales, pero podemos utilizar la tecnología como herramienta para sacar a la luz estos problemas sistémicos y abordarlos como sociedad en su conjunto. Porque "la IA solo funciona, si funciona para todos nosotros"[24]. 4. Conclusión Resumamos lo que hemos aprendido: Por un lado, la ciencia de datos y la IA tienen una enorme variedad de aplicaciones con un impacto social positivo. Por ejemplo, la ciencia de datos es útil para investigar el impacto de las redes sociales en los derechos humanos. Por otro lado, las aplicaciones de la ciencia de datos y la IA también conllevan riesgos para la salud, la seguridad, el medio ambiente y los derechos humanos. Los prejuicios y la discriminación, los problemas de privacidad y los impactos medioambientales perjudiciales son sólo algunos de los posibles efectos. La imparcialidad de los resultados en las aplicaciones de la ciencia de datos y la IA puede medirse de muchas maneras diferentes. La creación de aplicaciones de IA fiables requiere una intensa colaboración interdisciplinar: si nos aseguramos de que nuestros procesos de desarrollo son inclusivos y permiten una amplia participación, podremos crear mejores aplicaciones.
Autoevaluación (preguntas y respuestas de elección múltiple)	1. Nombra tres casos diferentes de uso de la ciencia de datos para el bien A) tarificación adaptativa B) adecuación de competencias C) seguimiento de las repercusiones de los derechos humanos en las redes sociales 2. ¿Cuál de los siguientes **no** es uno de los principios HLEG de la IA fiable? A) Robustez B) Reproducibilidad C) Transparencia 3. La métrica de equidad Equalized Odds exige que: A) el TPR sea igual en todos los grupos demográficos B) la FPR sea igual en todos los grupos demográficos C) Todas las anteriores

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Recursos, (vídeos, enlaces de referencia)	- [1] Skills adjacency detection and targeted training of missing skills: SkillsFuture Singapore, https://www.skillsfuture.gov.sg/AboutSkillsFuture - [2] AI & digital twins - simulating and practicing for resilience in the supply chain: https://www.technologyreview.com/2021/10/26/1038643/ai-reinforcement-learning-digital-twins-can-solve-supply-chain-shortages-and-save-christmas/ - [3] Reducing the footprint of recycled steel: Fero Labs uses AI to help steel manufacturers reduce the use of mined ingredients by up to 34%, preventing an estimated 450,000 tons of CO2 emissions per year: https://gpai.ai/projects/responsible-ai/environment/climate-change-and-ai.pdf - [4] Adaptive charging breaks down barriers to electric vehicle adoption. Bi-directional charging & Vehicle to Grid technologies need smart scheduling algorithms. https://ev.caltech.edu/info - [5] Using AI to detect forced labor in the supply chain: https://www.altana.ai/blog/illuminating-xinjiang-forced-labor-ecosystem - [6] Machine learning can boost the value of wind energy: https://www.deepmind.com/blog/machine-learning-can-boost-the-value-of-wind-energy - [7] Barometre dell'Odio:. https://www.amnesty.it/campagne/contrasto-allhate-speech-online/ - [8] Barometre dell'Odio: Elezioni europee. https://d21zrvtkxtd6ae.cloudfront.net/public/uploads/2020/01/Amnesty-barometro-odio-2019.pdf - [9] Barometre dell'Odio: sessimo da tastiera. https://www.amnesty.it/barometro-dellodio-sessismo-da-tastiera/#sintesi - [10] Ziad Obermeyer et al. Dissecting racial bias in an algorithm used to manage the health of populations. https://science.sciencemag.org/content/366/6464/447 - [11] The Guardian, Amazon ditched AI recruiting tool that favored men for technical jobs, October, 2018. https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine - [12] After Google's Gorillas comes Facebook's Primates: Facebook Apologizes After A.I. Puts 'Primates' Label on Video of Black Men, September 2021. https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html - [13] - [14] - [15] Joy Buolamwini & Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf - [16] The algorithms that detect hate speech online are biased against Black people. August 2019. https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter - [17] EU HLEG Guidelines for trustworthy AI: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai - [18] Chouldechova A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data. 2017 Jun;5(2):153-163. - [19] Machine bias. There's software used across the country to predict future criminals. And it's biased against blacks. May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing - [20] A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. October 2016. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/ - [21] Julia Dressl and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. January 2018. https://www.science.org/doi/10.1126/sciadv.aao5580 - [22] Sahil Verma, Julia Rubin: „**Fairness Definitions Explained",** 2018 ACM/IEEE International Workshop on Software Fairness; https://dl.acm.org/doi/10.1145/3194770.3194776 - [23] Richardson, R. et al, "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice"; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423 - [24] D. Raji, "How our data encodes systematic racism", MIT Technology Review. https://www.technologyreview.com/2020/12/10/1013617/racism-data-science-artificial-intelligence-ai-opinion/
Material relacionado	
PPT Relacionado	
Bibliografía	
Proporcionado por	[Women in AI Austria]

Co-funded by the Erasmus+ Programme of the European Union