ManagementSolutions
Making things happen

Explainable artificial intelligence (XAI)
Desafíos en la interpretabilidad de los modelos

I+D

www.managementsolutions.com

Diseño y Maquetación
Dpto. Marketing y Comunicación
Management Solutions - España

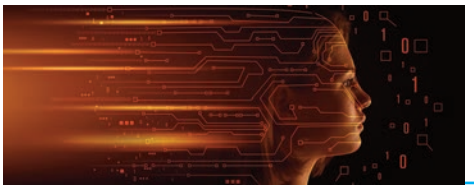
Fotografías
Archivo fotográfico de Management Solutions
iStock

Índice

Introducción 4

Resumen ejecutivo 8

Contexto y fundamentos de la XAI 12



Técnicas de interpretabilidad: estado del arte 22

Caso práctico de interpretabilidad 32

Conclusiones 38

Glosario 40

Bibliografía 42

Introducción

"La mayor parte de lo que hacemos con el aprendizaje automático ocurre bajo la superficie. Aunque no sea visible, gran parte del impacto del aprendizaje automático será así: una mejora silenciosa pero significativa de las operaciones esenciales".

Jeff Bezos[1]

"La inteligencia artificial (AI) es el campo de la ciencia y la ingeniería centrado en crear máquinas inteligentes, y en especial programas informáticos inteligentes. Está relacionada con la tarea similar de utilizar ordenadores para comprender la inteligencia humana, pero la AI no tiene por qué limitarse a métodos biológicamente observables"[2].

Esta fue la definición de AI que ofreció John McCarthy, profesor de la Universidad de Stanford, uno de los fundadores de esta disciplina y coautor del término "inteligencia artificial".

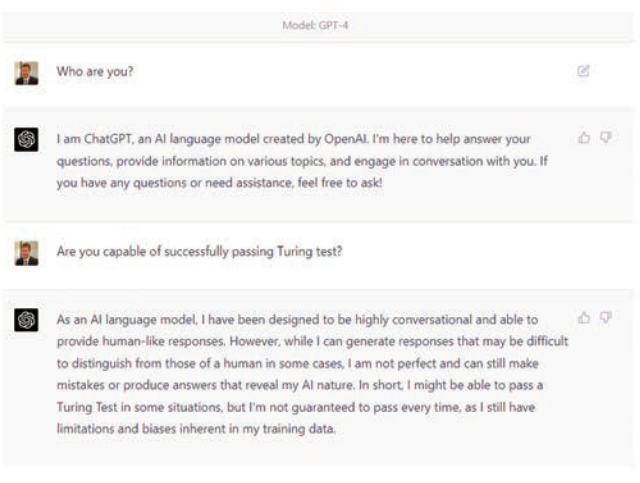
Sin embargo, ya en 1950 Alan Turing se preguntaba[3]: "¿pueden las máquinas pensar?", y formulaba lo que más tarde se conocería como "test de Turing": una prueba de la capacidad de una máquina para mostrar una inteligencia indistinguible de la de un ser humano. Turing propuso que un evaluador humano juzgara las conversaciones en lenguaje natural entre una persona y una máquina diseñada para generar respuestas similares a las humanas. Si el evaluador no era capaz de distinguir la máquina del humano, la máquina habría superado la prueba.

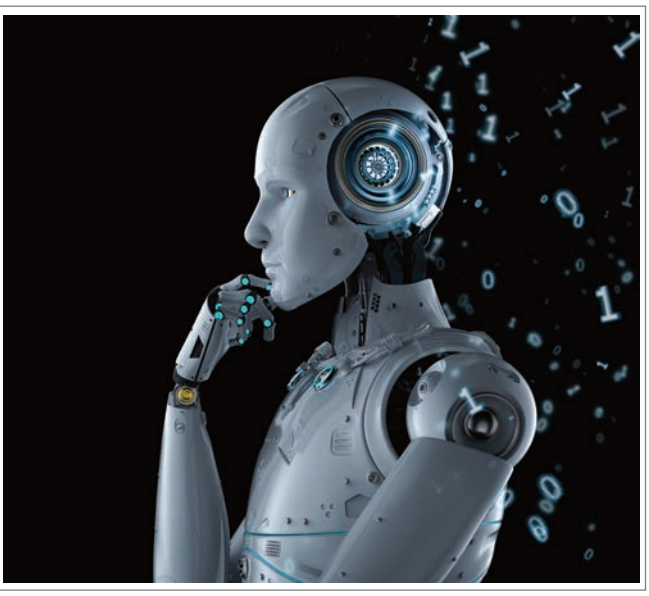
Aunque hay controversia al respecto[4], muchos autores consideran que ya hay inteligencias artificiales que podrían superar el test de Turing, como GPT-4, de la Open AI Foundation, aunque la misma GPT-4 no lo tiene totalmente claro (Fig. 1). Asimismo, existen tests más sofisticados, como la prueba de esquemas de Winograd, que consiste en la resolución de anáforas complejas que requieren conocimiento y sentido común[5], algo que la AI actual no parece ser capaz de hacer aún.

[1]Bezos (n. 1964), J., fundador, presidente ejecutivo y ex CEO de Amazon.
[2]McCarthy (2004). Profesor de Informática en la Universidad de Stanford.
[3]Turing (1950). Matemático, lógico, informático teórico, criptógrafo, filósofo y biólogo teórico británico.
[4]Harnad (2003). Catedrático de Psicología en la Universidad de Quebec en Montreal (UQAM) y en la Universidad McGill, y profesor emérito de Ciencias Cognitivas en la Universidad de Southampton.
[5]Un esquema de Winograd es una pregunta de opción binaria donde (i) hay dos partes mencionadas en la pregunta; (ii) se utilizan pronombres para referirse a ellas; (iii) hay una ambigüedad sobre a quién se refiere el pronombre; y (iv) hay palabras específicas que pueden cambiar la respuesta correcta. En un ejemplo del mismo Terry Winograd (profesor de Ciencias de la Computación en la Universidad de Stanford):
- Pregunta: los concejales de la ciudad negaron a los manifestantes la autorización porque [temían/defendían] la violencia. ¿Quién [teme/defiende] la violencia?
- Respuesta: [los concejales / los manifestantes].

Con ello, se puede generar un test alternativo al Test de Turing, utilizando dichas preguntas y penalizando fuertemente las respuestas erróneas (véase Levesque (2014)).

Figura 1. Conversación con GPT-4 sobre su capacidad para pasar el test de Turing.





Aun así, aunque el campo de la AI no es nuevo, en los últimos años se han realizado avances vertiginosos, con aplicaciones que van desde los coches de conducción autónoma hasta el diagnóstico médico, pasando por el *trading* automático, el reconocimiento facial, la gestión de la energía, la ciberseguridad, la robótica o la traducción automática, por citar algunas.

Una característica diferencial de la AI actual está precisamente ligada con la definición de McCarthy antes mencionada: no se limita a métodos observables, y, cuando alcanza cierto nivel de sofisticación, plantea problemas de interpretabilidad. En otras palabras: los modelos de AI tienden a tener una elevada tasa de acierto, muy superior a los algoritmos tradicionales; pero en cada caso concreto puede resultar extremadamente complejo explicar por qué el modelo ha producido un resultado determinado.

Aunque hay aplicaciones de la AI en las que no es tan relevante ser capaces de comprender o explicar por qué el algoritmo ha arrojado un valor concreto, en muchos casos resulta esencial y es un requerimiento regulatorio. Por ejemplo, en la Unión Europea, de acuerdo con el Reglamento General de Protección de Datos (GDPR), los consumidores tienen lo que se conoce como el "derecho a una explicación"[6]:

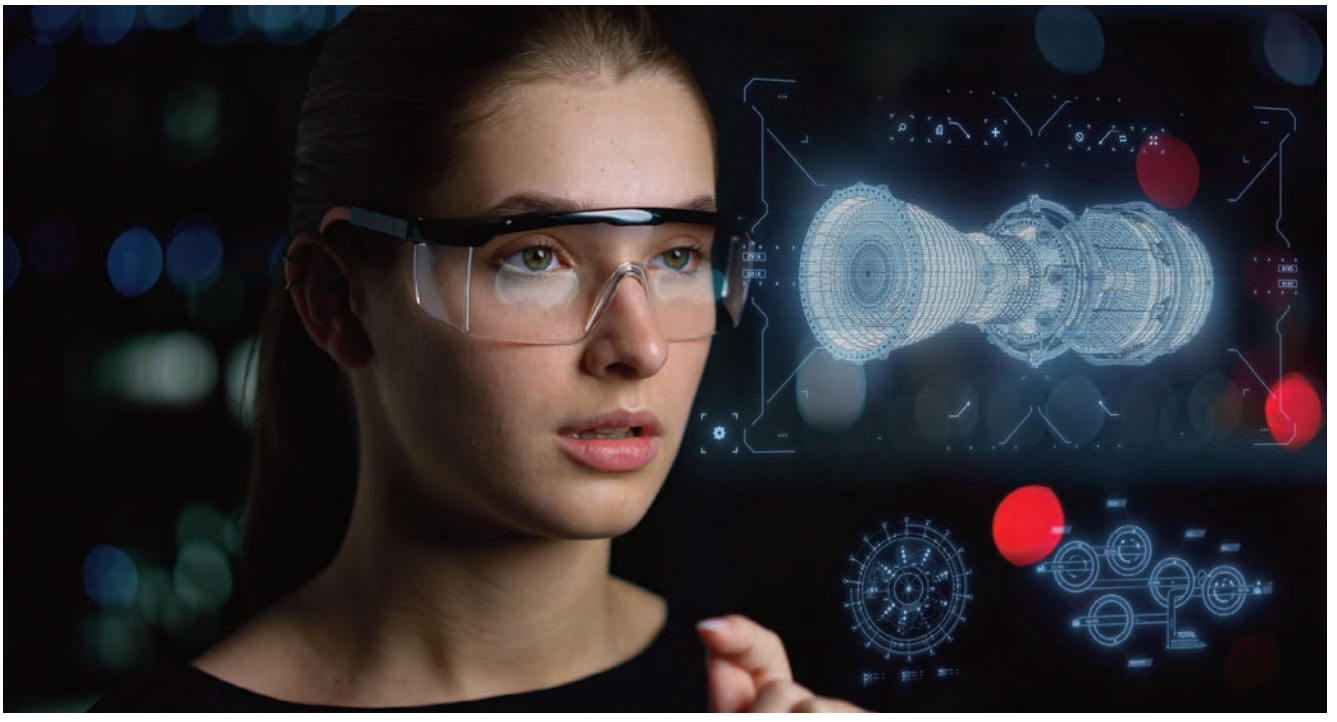
> [...] no ser objeto de una decisión [...] que se base únicamente en el tratamiento automatizado [...], como la denegación automática de una solicitud de crédito en línea, [...] [en la que] no medie intervención humana alguna", y tiene derecho "a recibir una explicación de la decisión tomada [...] y a impugnar la decisión".

Todo esto ha llevado al desarrollo de la disciplina de la inteligencia artificial explicable (XAI), que es el campo de estudio que pretende conseguir que los sistemas de AI resulten comprensibles para el ser humano[7], por contraposición a la noción de "caja negra" *(black box)*, que alude a los algoritmos en los que solo son observables los resultados y se desconoce el funcionamiento del modelo, o no se consigue explicar el fundamento por el cual se arrojan dichos resultados.

Se puede considerar[8] que un algoritmo se enmarca en la disciplina XAI si sigue tres principios: transparencia, interpretabilidad y explicabilidad. La transparencia se da si se pueden describir y justificar los procesos que calculan los parámetros del modelo y producen los resultados. La interpretabilidad describe la posibilidad de entender el modelo y presentar cómo toma decisiones de una manera comprensible para los humanos. La explicabilidad alude a la capacidad de descifrar por qué una determinada observación ha recibido un valor concreto. En la práctica, son tres términos muy ligados y con frecuencia se emplean de manera intercambiable, ante la falta de consenso sobre sus definiciones precisas[9].

Con el objetivo de conseguir dichos principios se pueden establecer básicamente dos estrategias de abordaje: o bien desarrollar algoritmos que son interpretables y explicables por

[6] GDPR (2018), Recital 71.
[7] Vilone et al. (2021). Doctora en Inteligencia Artificial, School of Computer Science, Technological University Dublin.
[8] Roscher et al. (2020). Científico de Datos de la Universidad Técnica de Múnich.
[9] Marcinkevics et al. (2020). Investigador en el Departamento de Informática, ETH Zürich.



su naturaleza (como las regresiones lineales, los modelos logísticos o multinomiales, o ciertos tipos de redes neuronales profundas, entre otros), o bien utilizar técnicas de interpretabilidad como herramientas para conseguir cumplir con estos principios[10].

La XAI, por tanto, se ocupa tanto de las técnicas para intentar explicar el comportamiento de determinados modelos opacos *(black box)* como del diseño de algoritmos inherentemente interpretables *(white box)*[11].

La XAI es fundamental en el desarrollo de la AI, y por tanto para los profesionales que trabajan en contacto con ella, por al menos tres factores:

- Contribuye a generar confianza en la toma de decisiones basadas en modelos de AI; sin esta confianza, los usuarios de estos modelos podrían mostrar resistencia a su adopción.
- Es un requerimiento regulatorio en determinados ámbitos (p. ej., protección de datos, protección del consumidor, igualdad de oportunidades en la contratación de empleados, regulación de modelos en banca).
- Favorece la mejora y el robustecimiento de los modelos de AI (p. ej., mediante la identificación y la eliminación de sesgos, la comprensión de la información relevante para producir un determinado resultado o la anticipación de posibles errores en observaciones no contempladas en la muestra de entrenamiento del modelo). Todo ello revierte en el desarrollo de algoritmos éticos, y permite focalizar los esfuerzos en las organizaciones en la identificación y aseguramiento de la calidad de los datos que son relevantes en los procesos de decisión.

Aunque el desarrollo de sistemas de XAI está recibiendo gran atención por parte de la comunidad académica, la industria y los reguladores, todavía plantea numerosos desafíos.

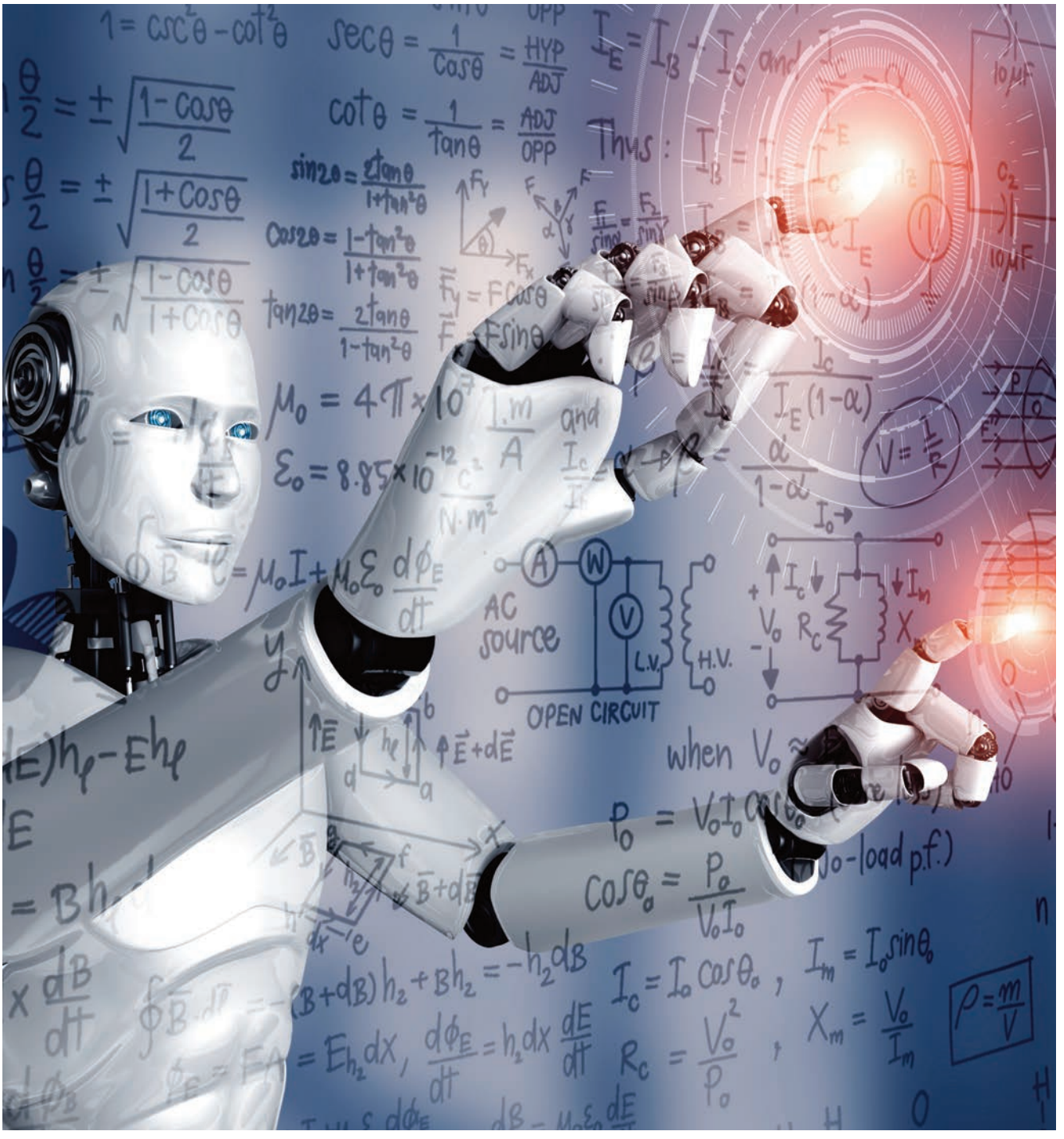
En este documento se repasan el contexto y los fundamentos de la XAI, incluyendo la normativa al respecto y sus implicaciones en la organización; el estado del arte y las principales técnicas de XAI; y los avances y retos sin resolver en la XAI. Por último, se proporciona un caso de estudio de XAI, para contribuir a ilustrar su aplicación práctica.

[10] iDanae (2022). Cátedra (inteligencia, datos, análisis y estrategia) en Big Data y Analytics, que surge gracias a la colaboración entre Management Solutions y la Universidad Politécnica de Madrid (UPM) en los campos formativo, científico y técnico, y tiene como objetivo promover la generación de conocimiento, difusión y transferencia de tecnología, y fomento de la I+D+i en el área de Análisis de Datos.

[11] Sudjianto et al. (2011). Responsable de Riesgo de Modelo de Wells Fargo.

Resumen ejecutivo

"Toda tecnología debería ir acompañada de un manual especial: no sobre cómo usarla, sino por qué, cuándo y para qué".
Alan Kay[12]



Contexto y fundamentos de la XAI

1. La transformación digital ha permitido el acceso y la explotación de una gran cantidad de datos estructurados y no estructurados, lo que ha impulsado el uso de técnicas de aprendizaje automático y la inteligencia artificial en todos los sectores.

2. Los modelos de AI proporcionan un mayor poder predictivo, pero también presentan riesgos, como la presencia de sesgos inadvertidos, la falta de comprensión del modelo o los errores en su aplicación derivados de causas como el sobre entrenamiento, todo lo cual puede generar desconfianza en el modelo. Esto plantea la pregunta de si es posible comprender suficientemente bien los resultados de los algoritmos de AI para tomar decisiones adecuadas.

3. La inteligencia artificial explicable (XAI) es un conjunto de procesos y métodos que permite a los usuarios comprender y confiar en los resultados y productos creados por algoritmos de aprendizaje automático. Esta disciplina es crucial para que una organización genere confianza a la hora de emplear modelos de AI, ayudando a caracterizar la precisión del modelo, la imparcialidad, la transparencia y el entendimiento de los resultados en la toma de decisiones basadas en AI.

4. El interés académico y profesional por la XAI ha aumentado exponencialmente en los últimos años, debido a la capacidad de esta disciplina para solucionar una serie de inquietudes de la industria en el uso de la AI, tales como requerimientos regulatorios, falta de confianza, potencial mal uso, impacto reputacional, impactos sociales o humanos y otros riesgos.

5. Esto ha llevado a reguladores y supervisores de distintas jurisdicciones a establecer reglamentos y directrices para el uso apropiado de la AI, incluyendo los aspectos de interpretabilidad de los modelos.

6. En Europa, el Reglamento General de Protección de Datos (GDPR) del Parlamento Europeo que entró en vigor en 2018 establece el "derecho a una explicación" de los ciudadanos, exigiendo que las compañías puedan explicar por qué un modelo de AI ha arrojado un determinado resultado. Esto tiene implicaciones críticas en el diseño y el análisis de interpretabilidad de los modelos de AI.

7. Por otra parte, el Parlamento Europeo propuso en 2021 el *Artificial Intelligence Act* (AI Act) para regular el uso de la inteligencia artificial en la Unión Europea. Esta propuesta de Reglamento establece un marco regulador para los sistemas de AI, incluyendo requisitos de desarrollo ético, transparencia, seguridad y precisión, así como un sistema de gobernanza y supervisión. El AI Act clasifica las aplicaciones de AI en niveles de riesgo (prácticas inaceptables, sistemas de alto riesgo y sistemas de riesgo bajo o limitado), y establece requerimientos de transparencia y vigilancia humana para los sistemas de alto riesgo, que serán de obligado cumplimiento en toda la Unión. Es probable que esto desencadene iniciativas de adaptación al Reglamento, como documentación exhaustiva de los modelos, técnicas de interpretabilidad, cuadros de mando de seguimiento y alertas sobre los modelos, entre otros.

8. Asimismo, la Comisión Europea formuló en 2019 las Directrices Éticas para una Inteligencia Artificial Fiable, que proponen siete requisitos clave para que los sistemas de AI sean considerados fiables: (i) acción y supervisión humanas, (ii) solidez técnica y seguridad, (iii) gestión de la privacidad y de los datos, (iv) transparencia, (v) diversidad, no discriminación y equidad, (vi) bienestar ambiental y social, y (vii) rendición de cuentas. Dentro del requisito de transparencia, se establece la necesidad de explicabilidad de los modelos de AI. Las Directrices proponen unos criterios para evaluar en qué medida un modelo de AI cumple con estos requisitos.

9. En Estados Unidos, en 2022 la Casa Blanca propuso un borrador de Declaración de Derechos sobre Inteligencia Artificial *(AI Bill of Rights)*, impulsado por el presidente Joe Biden. Esta declaración establece cinco principios o

[12] Alan Kay (n. 1940), informático estadounidense galardonado con el premio Turing, considerado el "padre de los ordenadores personales".

derechos de los ciudadanos en lo relativo a la AI, que incluyen sistemas seguros y efectivos, protección contra la discriminación de los algoritmos, privacidad de los datos, notificación y explicación, y evaluación y corrección por un ser humano en caso de fallo de la AI *(fallback)*. Estos principios incluyen la explicabilidad de los modelos de AI, que requiere una documentación en lenguaje sencillo, explicaciones técnicamente válidas, significativas y útiles, y notificaciones de uso demostrablemente claras, oportunas, comprensibles y accesibles.

10. Los Principios de la OCDE sobre Inteligencia Artificial, de 2019, promueven el uso de una AI digna de confianza y que respete los derechos humanos y los valores democráticos. Fueron adoptados por los 38 países miembros de la OCDE y requieren, entre otros, la transparencia y la divulgación responsable de los sistemas de AI para que los afectados por un sistema de AI puedan comprender el resultado.

11. El Discussion Paper on Machine Learning for IRB Models de la Autoridad Bancaria Europea (EBA), publicado en 2021, analiza la relevancia de los posibles obstáculos para la implementación de técnicas de aprendizaje automático en el ámbito del enfoque IRB de cálculo de capital en entidades financieras. El documento establece principios y recomendaciones para hacer compatible el uso de estas técnicas con el cumplimiento de la regulación europea sobre capital (CRR). Estas recomendaciones incluyen el análisis estadístico y económico de la relación entre las variables de entrada y la variable de salida, una documentación que explique de forma sencilla el modelo, y la necesidad de detección de posibles sesgos en el modelo.

12. Un principio básico de la XAI es la necesidad de integrar la interpretabilidad y explicabilidad en la organización y los procesos de una compañía. Esto se lleva a cabo a través de un marco de XAI compuesto por cuatro elementos: técnicas de interpretabilidad de los modelos de AI, integración en los procesos de gestión del riesgo de modelo (MRM), soporte tecnológico y factor humano.

13. Técnicas: el núcleo del marco de XAI se basa en tres aspectos principales de interpretabilidad: la explicación del diseño del modelo, la explicación de los resultados del modelo, y otros aspectos como la detección de sesgos y el seguimiento periódico del modelo.

14. MRM: la interpretabilidad de los modelos de AI es una característica que afecta a toda la cadena del ciclo de vida de los modelos, y por tanto a la gestión del riesgo de modelo. Para incorporar los elementos propios de XAI, se debe revisar y actualizar el marco de organización y gobierno, las políticas y procedimientos de desarrollo, seguimiento, validación, implementación y uso de los modelos, así como el marco de auditoría.

15. Soporte tecnológico: para implementar un marco de XAI, se requieren soluciones tecnológicas profesionales para dar soporte a los aspectos propios de la interpretabilidad de los modelos de AI, como son las herramientas de interpretabilidad y de gobierno de modelos, los sistemas de análisis de datos, APIs, mecanismos de seguridad y auditoría, y los protocolos para garantizar el cumplimiento de los estándares de calidad y explicabilidad.

16. Factor humano: la integración de XAI debe considerar el factor humano, incluyendo la captación y retención de talento especializado, programas de formación, la creación de una cultura que potencie el uso de la explicabilidad y la interpretabilidad de los modelos de AI, y programas de gestión del cambio para asegurar la adecuada adopción de XAI.

17. Adicionalmente, un quinto elemento central para la AI y la XAI son los datos, por cuanto su buen gobierno, calidad, integridad, consistencia, trazabilidad y ausencia de sesgos determinan la calidad del modelo de AI, y en último término de las decisiones que se toman basadas en él. No obstante, los aspectos relativos a los datos y su relevancia en los modelos no son objeto de este documento, puesto que ya han sido abordados extensivamente en publicaciones anteriores[13].

Técnicas de interpretabilidad: estado del arte

18. El uso de técnicas de AI se ha extendido a todas las industrias y ámbitos, ofreciendo un mayor poder predictivo a cambio de mayor complejidad. Esto ha generado la necesidad de explicar los resultados de los modelos de AI, lo que ha llevado a la aparición de técnicas cada vez más sofisticadas de interpretabilidad local y global. Estas técnicas no resuelven por completo el problema, por lo que se siguen desarrollando distintos enfoques para garantizar la interpretabilidad de los modelos de AI, como el desarrollo de modelos inherentemente interpretables *(white boxes)*.

19. Los enfoques más comunes para abordar el problema de la interpretabilidad se pueden clasificar en dos grupos: interpretabilidad post-hoc (técnicas de interpretabilidad global y local) y modelos inherentemente interpretables. Además, existen estrategias complementarias, como la simplificación del modelo, el uso de variables con sentido de negocio, el análisis de datos para identificar sesgos o falta de imparcialidad, y el análisis de la reproducibilidad del desarrollo del modelo.

20. LIME *(Local Interpretable Model-agnostic Explanations)* permite explicar un modelo de manera local y agnóstica; es decir, puede generar explicaciones para una predicción específica sin tener que entender el modelo subyacente.

21. SHAP *(SHapley Additive exPlanations)* explica el modelo de manera global mediante la evaluación de la contribución de cada variable de entrada a la predicción de salida.

[13] Véanse Management Solutions (2020, 2018 y 2015): "Auto machine learning, hacia la automatización de los modelos", "Machine learning, una pieza clave en la transformación de los modelos de negocio" y "Data science y la transformación del sector financiero".

22. Los PDP (Gráficos de Dependencia Parcial) se utilizan para visualizar cómo cambia la salida de un modelo cuando se modifican los valores de las variables de entrada.

23. Los modelos *white box* se basan en el desarrollo de algoritmos que, por diseño, son inherentemente interpretables. Estos modelos se agrupan según el tipo de algoritmo empleado, y se suelen limitar los parámetros a optimizar para conseguir una mayor interpretabilidad. Con ello, se obtienen resultados más precisos, ya que permiten una mayor comprensión de la información, lo que a su vez resulta en una mejor toma de decisiones, especialmente en aquellos sectores en los que la interpretabilidad es un factor crítico.

24. A pesar de los avances en la interpretabilidad de los modelos de AI, todavía existen retos como la reproducibilidad de los resultados, la explicación de la secuencia de predicciones más probables, los sesgos en los datos de entrada, la imparcialidad *(fairness)* y la exactitud de la explicación. Además, hay margen de mejora en el desarrollo de los modelos *white box* para competir en precisión con los modelos *black box* en problemas complejos, así como en el desarrollo de nuevas técnicas para explicar los modelos más complejos.

Caso práctico de interpretabilidad

25. Con el objetivo de mostrar la aplicación de las técnicas de interpretabilidad descritas, se realiza un ejercicio ilustrativo empleando datos ficticios generados por IBM y publicados en Kaggle[14]. El objetivo del estudio es comprender las causas que llevan a los empleados a abandonar su puesto de trabajo, y para ello emplear técnicas de AI y XAI sobre los datos ficticios propuestos.

26. El ejercicio se ha realizado con ayuda de un sistema de modelización por componentes, ModelCraft™[15], que contiene múltiples técnicas relevantes de AI y XAI, lo que ha permitido completar el estudio en un tiempo muy inferior a lo habitual, y sin necesidad de escribir código.

27. Para explicar el abandono de los empleados, se han entrenado y validado distintos modelos, entre los que el algoritmo *random forest* ha arrojado la mejor capacidad predictiva.

28. Para explicar los resultados del modelo, se han aplicado las técnicas de interpretabilidad SHAP, LIME y PDP, que han permitido comprender qué variables explican mejor el abandono de los empleados, cómo impactan los cambios en las variables más importantes para distintos rangos de población, y los resultados del modelo en casos individuales.

29. La correcta aplicación e interpretación del modelo en este caso de estudio permitiría anticipar y prevenir el abandono de empleados, crear perfiles con distinta propensión al abandono e identificar las características de estos empleados con antelación para tomar las medidas adecuadas. Además, este caso de uso pone de manifiesto las limitaciones y dificultades en la aplicación de las técnicas de interpretabilidad *post-hoc,* así como el hecho de que emplear modelos de AI junto con un módulo de interpretabilidad puede potenciar la capacidad predictiva del modelo.

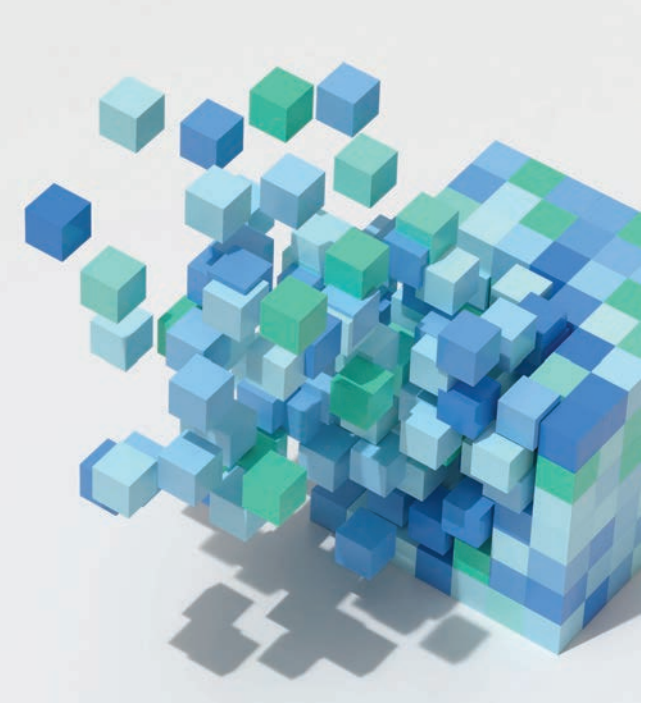
Conclusión

30. La inteligencia artificial explicable (XAI) es una disciplina emergente que busca mejorar la interpretabilidad de los modelos de AI mediante el uso de técnicas específicas para entender y explicar los resultados de los modelos de AI, y es especialmente importante en ámbitos de alta sensibilidad, como la salud, la seguridad, los servicios financieros y la energía, entre otros.

31. La XAI se ha convertido en una prioridad para muchos sectores, ya que los modelos de AI se vuelven cada vez más complejos y cada vez hay más regulación que requiere su interpretabilidad. Un caso práctico desarrollado con ModelCraft™ ha demostrado cómo se pueden emplear estas técnicas para entender y explicar los modelos de AI.

32. En los próximos años, es esperable que la XAI continúe desarrollándose y creciendo en importancia a medida que los modelos de AI se vuelvan más complejos, la regulación siga proliferando, y su uso se extienda a más ámbitos de alta sensibilidad.

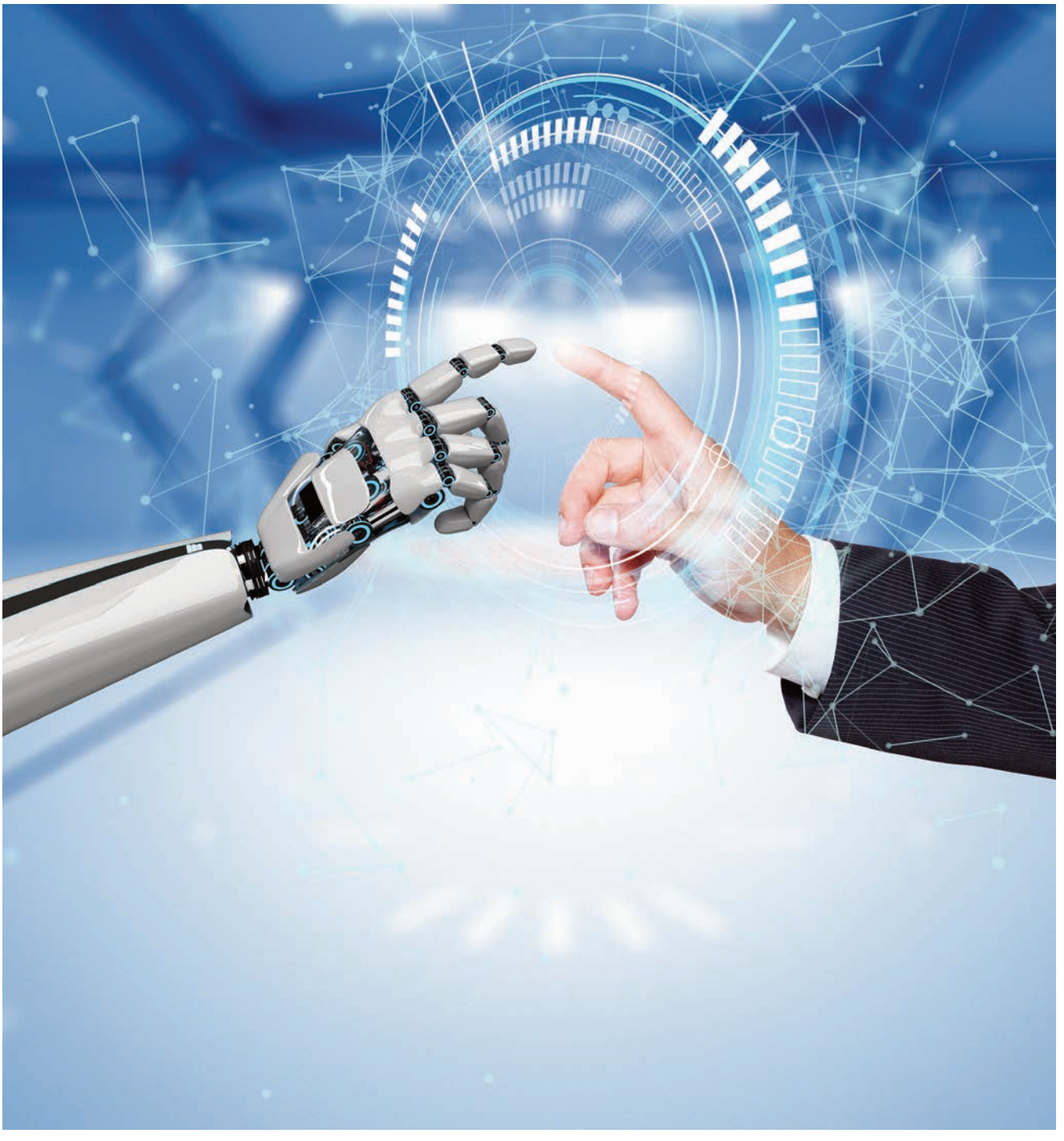
[14] Kaggle (2017). IBM HR Analytics Employee Attrition & Performance.
[15] Herramienta de AutoML y modelización por componentes propietaria de Management Solutions. Véase Management Solutions (2023).



Contexto y fundamentos de la XAI

"Comprender la inteligencia artificial es un desafío que requiere una enorme capacidad intelectual; afortunadamente, contamos con la inteligencia artificial para abordarlo".

GPT-4[16]



Contexto

Una de las características más notables de la transformación digital es que está poniendo a disposición de todas las industrias una cantidad masiva de datos estructurados y no estructurados provenientes de múltiples aplicaciones; por ejemplo:

- Datos de comercio minorista procedentes de acciones de compra, transacciones y comentarios de los clientes.
- Datos financieros procedentes de fuentes bancarias, de inversión y comerciales.
- Datos de redes sociales, incluidos análisis de opiniones y análisis predictivos.
- Sensores digitales IoT (Internet de las Cosas) que miden la temperatura, la presión y otros datos del entorno.
- Datos sanitarios, como historiales médicos, diagnósticos, imágenes e información genómica.
- *Wearables*, como rastreadores de actividad, sensores de salud y relojes inteligentes.
- Sistemas de reconocimiento de voz que permiten a las máquinas entender y responder al lenguaje natural.
- Satélites y otros sensores espaciales que proporcionan información sobre el tiempo y el clima.
- Sistemas de vigilancia inteligentes que utilizan el reconocimiento facial y la detección de objetos.
- Sensores de vehículos autónomos como cámaras, lidar, radar y sensores ultrasónicos.

La disponibilidad de estos datos, junto con la presencia de enormes capacidades de almacenamiento y procesamiento computacional a coste reducido, ha impulsado un mayor apetito por la modelización avanzada, que se manifiesta en el uso de una amplia gama de técnicas de aprendizaje automático y en el desarrollo de la inteligencia artificial (AI) en prácticamente todos los sectores y ámbitos[17].

Aunque hay consenso sobre el hecho de que los modelos de AI proporcionan en general un mayor poder predictivo que los modelos tradicionales[18], también introducen una mayor complejidad y puede resultar difícil interpretarlos y explicar sus resultados.

Esto genera riesgos vinculados con el uso de estos modelos, como la falta de comprensión del modelo, la presencia de sesgos inadvertidos o la dificultad para determinar si el modelo está sobreentrenado (global o localmente), lo que puede dar lugar a una escasa capacidad de generalización y a potenciales errores en las decisiones basadas en él, y como consecuencia, derivar en una falta de confianza en el modelo.

Todo ello lleva a la pregunta de si es posible comprender lo suficientemente bien los resultados que arrojan los algoritmos de AI, especialmente cuando tienen impacto en decisiones críticas, como el diagnóstico médico, la conducción autónoma o la detección del fraude, entre otras muchas.

[16] GPT-4, Generative Pre-Trained Transformer, red neuronal profunda diseñada por la Fundación OpenAI para realizar tareas de procesamiento del lenguaje natural (NLP). En este caso, se le pidió "Inventa 10 citas ingeniosas sobre la inteligencia artificial y cómo de difícil y necesario es ser capaz de interpretar y explicar los modelos de AI". La cita presentada fue la tercera.

[17] Aunque hay diferencias, dada la falta de consenso sobre su definición, en este documento se emplearán de forma indistinta los términos "aprendizaje automático", "machine learning (ML)", "inteligencia artificial (AI)" y "modelización avanzada". Asimismo, se utilizará la abreviatura "AI" para "inteligencia artificial", por consistencia con las siglas "XAI" (que habitualmente no se traducen), incluso en las citas de publicaciones en español.

[18] LeCun, Y. et al (2015). Investigador en Facebook AI Research y la Universidad de Nueva York.

Figura 2. Número de publicaciones científicas por año sobre Explainable Artificial Intelligence (XAI).

Definición

La disciplina de XAI es relativamente nueva y, por tanto, no hay todavía una doctrina asentada que estandarice su terminología. Pese a algunos esfuerzos notables para definir los términos[19], la aproximación a la XAI es o bien heterogénea (según la fuente académica consultada), o bien intuitiva (más frecuente en la práctica industrial).

En todo caso, para la mayor parte de usos en la práctica puede ser suficiente definir XAI del siguiente modo[20]:

> La inteligencia artificial explicable (XAI) es el conjunto de procesos y métodos que permiten a los usuarios humanos comprender y confiar en los resultados y productos creados por algoritmos de aprendizaje automático. La XAI se utiliza para describir un modelo de AI, su impacto previsto y sus posibles sesgos. Ayuda a caracterizar la precisión del modelo, la imparcialidad, la transparencia y los resultados en la toma de decisiones basada en AI. La XAI es crucial para que una organización genere confianza a la hora de poner en producción modelos de AI. La explicabilidad de la AI también ayuda a una organización a adoptar un enfoque responsable del desarrollo de la AI.

Relevancia de la XAI

Un aspecto en el que hay consenso entre académicos y profesionales de la industria es en la relevancia creciente de la XAI como disciplina complementaria a la AI.

Las herramientas de análisis de publicaciones científicas identifican más de 77.000 artículos sobre XAI entre 2014 y 2022, y en tendencia exponencialmente creciente, con más de 20.000 artículos solo en 2022 (Fig. 2)[21].

Más allá del interés académico, la atención que recibe la XAI se explica por su capacidad para dar solución a una serie de inquietudes de la industria en el uso de la AI (Fig. 3); entre ellas:

- **Requerimientos regulatorios:** la obligación de cumplir con la regulación emergente sobre el uso de AI.
- **Falta de confianza:** la necesidad de generar confianza sobre el modelo de AI y los resultados que arroja en los usuarios, los validadores y auditores, y en última instancia el público en general.
- **Potencial mal uso:** la conveniencia de evitar el mal uso de los modelos debido a la falta de comprensión sobre su funcionamiento, lo que puede conllevar costes e incluso sanciones.
- **Impacto reputacional:** la prevención de impactos reputacionales sobre la compañía debidos a sesgos, decisiones discriminatorias, uso inapropiado o simplemente predicciones erróneas del modelo.
- **Impactos sociales o humanos:** la prevención de impactos sociales o humanos en usos críticos como la AI para el diagnóstico de enfermedades médicas, sentencias judiciales, identificación biométrica, polígrafos, etc.
- **Otros:** la mitigación de otros riesgos que emanan de la falta de comprensión sobre el modelo, como ciberseguridad, protección de datos, fraude, riesgo de modelo, etc.

Pese a todo lo anterior, hay casos en los que los modelos de AI no necesitan ser particularmente interpretables, porque los usos no están regulados, porque no tienen impactos potenciales relevantes o simplemente porque no es necesario interpretarlos, como los sistemas de recomendación automática de cine y música, o los algoritmos que juegan al ajedrez, por ejemplo.

[19] Marcinkevics et al. (2020). Departamento de Computer Science, ETH Zürich.
[20] IBM (2022).
[21] Dimensions (2022).

Figura 3. Inquietudes de la industria que la XAI contribuye a solucionar.

Regulación

La XAI, por tanto, se está posicionando como una disciplina de relevancia creciente; y esto está llevando a reguladores y supervisores de distintas jurisdicciones a establecer reglamentos y directrices para el uso apropiado de la AI, incluyendo los aspectos de interpretabilidad de los modelos.

En este contexto, posiblemente las referencias regulatorias más relevantes a la fecha de redacción de este documento son las siguientes:

1. GDPR (Parlamento Europeo)

En Europa, el Reglamento General de Protección de Datos, que entró en vigor en 2018, establece el "derecho a una explicación" de los ciudadanos, según el cual[22]:

> El interesado debe tener **derecho a no ser objeto de una decisión**, que puede incluir una medida, que evalúe aspectos personales relativos a él, y **que se base únicamente en el tratamiento automatizado** y produzca efectos jurídicos en él o le afecte significativamente de modo similar, como la denegación automática de una solicitud de crédito en línea o los servicios de contratación en red en los que no medie intervención humana alguna. [...]
>
> En cualquier caso, dicho tratamiento debe estar sujeto a las garantías apropiadas, entre las que se deben incluir la información específica al interesado y el derecho a obtener intervención humana, a expresar su punto de vista, a **recibir una explicación de la decisión tomada después de tal evaluación y a impugnar la decisión.**

Esto tiene implicaciones críticas en el uso de la AI, y puede llevar a cuestionar su viabilidad. No obstante, en palabras del Parlamento Europeo[23]:

> Ciertamente existe una tensión entre los principios tradicionales de protección de datos –limitación de la finalidad, minimización de los datos, tratamiento especial de los "datos sensibles", limitación de las decisiones automatizadas– y el pleno despliegue del poder de la AI y *big data*. Estos últimos implican la recopilación de cantidades ingentes de datos relativos a las personas y sus relaciones sociales y su tratamiento para fines que no estaban totalmente determinados en el momento de la recopilación. Sin embargo, hay formas de interpretar, aplicar y desarrollar los principios de protección de datos que son coherentes con los usos beneficiosos de la AI y de *big data*.

Y esto está en línea con el cuarto principio para el uso ético de la AI establecido por el Grupo de Alto Nivel sobre inteligencia artificial de la Comisión Europea[24]:

> Explicabilidad: los procesos algorítmicos deben ser transparentes, las capacidades y objetivos de los sistemas de AI deben comunicarse abiertamente, y las decisiones deben poder explicarse a los afectados directa e indirectamente.

En todo caso, GDPR tiene impactos relevantes en el uso de la AI, en el sentido de que las compañías están legalmente obligadas a poder explicar por qué un modelo de AI ha arrojado un determinado resultado, y esto tiene implicaciones críticas en el diseño y el análisis de interpretabilidad de los modelos de AI[25].

2. Artificial intelligence act (Parlamento Europeo)

El borrador de Reglamento de inteligencia artificial o *artificial intelligence act* (AI Act), publicada en 2021, es una propuesta para el uso de la inteligencia artificial en la Unión Europea que pretende garantizar un alto nivel de confianza en la AI y sus aplicaciones, al tiempo que sienta las bases para la innovación.

[22] GDPR (2018), Cons. 71.
[23] *European Parliamentary Research Service* (2020).
[24] Ibíd.
[25] En algunos países europeos se está analizando el nivel de cumplimiento de este tipo de IA (en particular, de los denominados *Large Language Models*) con la regulación de protección de datos, y en ciertos casos se ha prohibido el uso de algunos de estos modelos de forma provisional.

El Reglamento establece un marco regulador para los sistemas de AI en la UE, e incluye requisitos de desarrollo ético, transparencia, seguridad y precisión. También establece un sistema de gobernanza y supervisión de los sistemas de AI, así como normas de protección y gobernanza de datos.

Al tratarse de un Reglamento, cuando sea aprobado será de aplicación directa en los 27 países de la Unión[26], sin necesidad de ser traspuesto al ordenamiento jurídico de cada país.

Una de sus características fundamentales es que clasifica las aplicaciones de AI en niveles de riesgo[27]:

- **Prácticas prohibidas,** que denotan la categoría de mayor riesgo; estos sistemas están totalmente prohibidos. Entre ellos se incluyen:
  - Sistemas biométricos en tiempo real que pueden utilizarse para cualquier tipo de vigilancia, aunque se aplican excepciones para la prevención de delitos y las investigaciones criminales en contextos policiales y de seguridad nacional.
  - Algoritmos de puntuación social que pueden utilizarse para evaluar a los individuos basándose en características personales o en su comportamiento de una manera que pueda causar daño o conducir a un trato desfavorable a un individuo.
  - Sistemas manipuladores que explotan las vulnerabilidades de determinados individuos específicos para distorsionar su comportamiento de manera que pueda causar daño físico o psicológico.

- **Sistemas de AI de alto riesgo,** enumerados en el Anexo III y que probablemente constituyan la mayoría de los sistemas de AI. Entre ellos se incluyen:
  - Identificación biométrica y categorización de personas físicas […].
  - Gestión y funcionamiento de infraestructuras críticas […] [p. ej. tráfico].
  - Educación y formación profesional […].
  - Empleo y gestión de trabajadores […].
  - Acceso a servicios esenciales […], incluida la evaluación de la solvencia, la calificación crediticia o el establecimiento del orden de prioridad de acceso a dichos servicios. (Nota: este aspecto se aplica en particular a los sistemas de AI utilizados en el sector de los servicios financieros).
  - Fuerzas de seguridad […].
  - Gestión de controles fronterizos […].
  - Administración de justicia y procesos democráticos […].

- **Sistemas de AI de bajo riesgo** (o riesgo limitado), que incluyen sistemas que no utilizan datos personales ni hacen predicciones que puedan afectar directa o indirectamente a ninguna persona, como las aplicaciones industriales de mantenimiento predictivo.

En lo relativo a la interpretabilidad de los modelos de AI clasificados como de alto riesgo, el AI Act establece[28] en sus Artículos 13 y 14:

Art. 13. Transparencia y comunicación de información a usuarios

1. Los sistemas de AI de alto riesgo se diseñarán y desarrollarán de un modo que garantice que funcionan con un **nivel de transparencia suficiente para que los usuarios interpreten y usen correctamente su información de salida.** […]

2. Los sistemas de AI de alto riesgo irán acompañados de las instrucciones de uso correspondientes en un formato digital o de otro tipo adecuado, las cuales incluirán información concisa, completa, correcta y clara que sea pertinente, accesible y comprensible para los usuarios. […]

Art. 14. Vigilancia humana

1. Los sistemas de AI de alto riesgo se diseñarán y desarrollarán de modo que puedan ser vigilados de manera efectiva por personas físicas durante el período que estén en uso, lo que incluye dotarlos de una herramienta de interfaz humano-máquina adecuada, entre otras cosas.

[…]

4. Las medidas mencionadas […] permitirán que las personas a quienes se encomiende la vigilancia humana puedan, en función de las circunstancias:

   a. **Entender por completo las capacidades y limitaciones del sistema de AI de alto riesgo** y controlar debidamente su funcionamiento, de modo que puedan detectar indicios de anomalías, problemas de funcionamiento y comportamientos inesperados y ponerles solución lo antes posible;
   b. ser conscientes de la posible tendencia a confiar automáticamente o en exceso en la información de salida generada por un sistema de AI de alto riesgo («sesgo de automatización») […];
   c. interpretar correctamente la información de salida del sistema de AI de alto riesgo […];
   d. decidir, en cualquier situación concreta, no utilizar el sistema de AI de alto riesgo o desestimar, invalidar o revertir la información de salida que este genere;
   e. intervenir en el funcionamiento del sistema de AI de alto riesgo o interrumpir el sistema […].

Como se puede observar, el AI Act impone condiciones restrictivas sobre la interpretabilidad de los modelos de AI de alto riesgo (Fig. 4), que en breve serán de obligado cumplimiento en toda la Unión. Es previsible que esto

[26] Se prevé que entre en vigor a los 20 días desde su publicación en el Diario Oficial de la Unión Europea, y que sea de plena aplicación a los 24 meses desde su entrada en vigor.
[27] Floridi et al. (2022).
[28] Comisión Europea (2021).

desencadene una cantidad significativa de iniciativas de adaptación al Reglamento, incluyendo una documentación más exhaustiva de los modelos y de sus usos, la aplicación de técnicas de interpretabilidad, el desarrollo de cuadros de mando de seguimiento y alertas sobre los modelos, o la revisión del procedimiento integral de desarrollo, validación, implementación y uso de los modelos, entre otros.

3. Directrices éticas para una inteligencia artificial fiable (Comisión Europea)

En abril de 2019, el grupo de expertos de alto nivel sobre AI de la Comisión Europea presentó las Directrices Éticas para una AI fiable[29], tras un proceso de consulta con más de 500 respuestas de la industria.

Las Directrices proponen siete requisitos clave que deben cumplir los sistemas de AI para ser considerados fiables, que en resumen son: (i) acción y supervisión humanas, (ii) solidez técnica y seguridad, (iii) gestión de la privacidad y de los datos, (iv) transparencia, (v) diversidad, no discriminación y equidad, (vi) bienestar ambiental y social, y (vii) rendición de cuentas.

En concreto, en lo relativo a la interpretabilidad de los modelos de AI, las Directrices establecen lo siguiente dentro de su requisito de transparencia:

> 53. La explicabilidad es crucial para conseguir que los usuarios confíen en los sistemas de AI y para mantener dicha confianza. Esto significa que los procesos han de ser transparentes, que es preciso comunicar abiertamente las capacidades y la finalidad de los sistemas de AI y que las decisiones deben poder explicarse —en la medida de lo posible— a las partes que se vean afectadas por ellas de manera directa o indirecta. Sin esta información, no es posible impugnar adecuadamente una decisión.
>
> No siempre resulta posible explicar por qué un modelo ha generado un resultado o una decisión particular (ni qué combinación de factores contribuyeron a ello). Esos casos, que se denominan algoritmos de «caja negra», requieren especial atención.
>
> En tales circunstancias, puede ser necesario adoptar otras medidas relacionadas con la explicabilidad (por ejemplo, la trazabilidad, la auditabilidad y la comunicación transparente sobre las prestaciones del sistema), siempre y cuando el sistema en su conjunto respete los derechos fundamentales.
>
> El grado de necesidad de explicabilidad depende en gran medida del contexto y la gravedad de las consecuencias derivadas de un resultado erróneo o inadecuado.

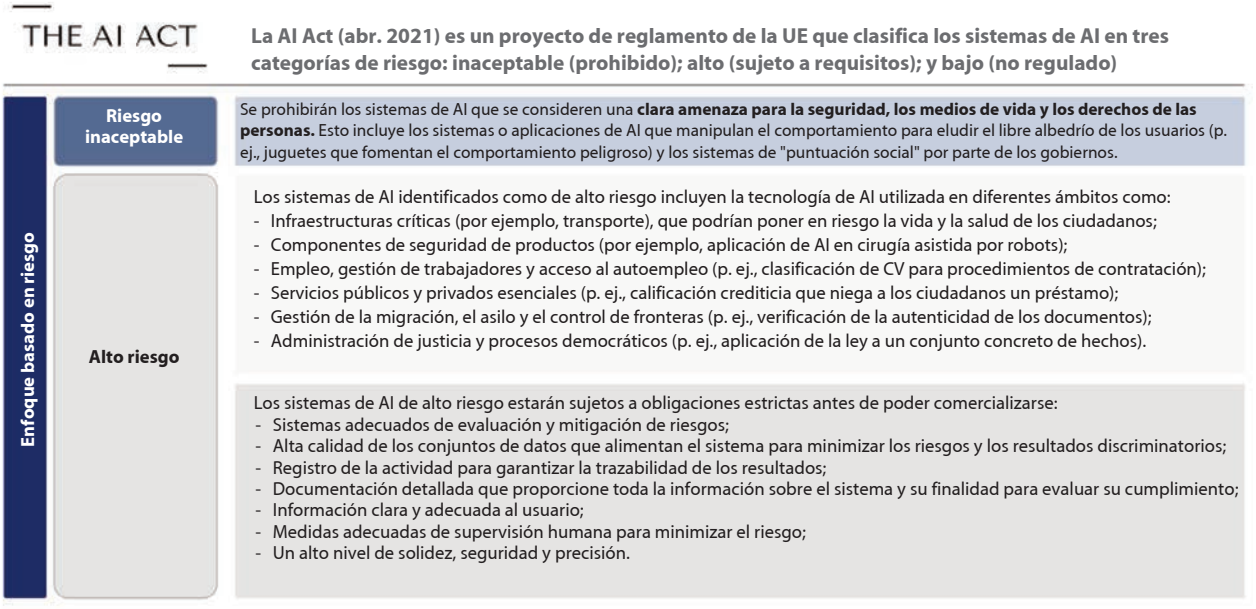
Como se puede apreciar, las Directrices apuntan en la misma dirección: el requerimiento (que se eleva al nivel de necesidad ética) de que los modelos de AI sean explicables.

Asimismo, lo que a primera vista podría parecer un requerimiento más relajado de interpretabilidad de los modelos de AI, por cuanto las Directrices reconocen que hay modelos de AI más difíciles de explicar, en realidad introduce una complejidad adicional: la necesidad de clasificar los modelos de AI según su riesgo y su potencial de ser interpretados, para aplicar un mayor o menor grado de esfuerzo en su explicación.

Por último, las Directrices están orientadas a evaluar en qué medida un modelo de AI cumple con estos siete requisitos, y para ello propone un listado de criterios de evaluación, que debe adaptarse a cada caso específico. En lo relativo a la explicabilidad, las Directrices formulan los siguientes criterios de evaluación[30], que deberían integrarse con otras herramientas de evaluación de las que ya dispongan las organizaciones:

[29] Comisión Europea (2019).
[30] Ibíd

Figura 4. Ámbitos de aplicación y requerimientos de la Artificial Intelligence Act.



- ¿Ha evaluado en qué medida son comprensibles las decisiones y, por tanto, el resultado producido por el sistema de AI?
- ¿Se ha asegurado de que se pueda elaborar una explicación comprensible para todos los usuarios que puedan desearla sobre las razones por las que un sistema adoptó una decisión determinada que diera lugar a un resultado específico?
- ¿Ha evaluado en qué medida la decisión del sistema influye en los procesos de adopción de decisiones de la organización?
- ¿Ha evaluado por qué se desplegó ese sistema en particular en esa área concreta?
- ¿Ha evaluado el modelo de negocio del sistema (por ejemplo, de qué modo crea valor para la organización)?
- ¿Ha diseñado el sistema de AI teniendo en mente desde el principio la interpretabilidad?
- ¿Ha investigado y tratado de utilizar el modelo más sencillo e interpretable posible para la aplicación en cuestión?
- ¿Ha evaluado si puede analizar sus datos relativos a la formación y los ensayos realizados? ¿Puede modificar y actualizar estos datos a lo largo del tiempo?
- ¿Ha evaluado si, tras la formación y el desarrollo del modelo, tiene alguna posibilidad de examinar su interpretabilidad o si dispone de acceso al flujo de trabajo interno del modelo?

4. Blueprint for an AI Bill of Rights (White House)

En octubre de 2022, la Casa Blanca propuso un Borrador de Declaración de Derechos sobre inteligencia artificial[31], impulsado por el presidente Joe Biden y desarrollado por la Oficina de Política Científica y Tecnológica (OSTP) de la Casa Blanca, y se acompaña de un manual *(From principles to practice)* sobre cómo implementarlo en la práctica.

El *AI Bill of Rights* establece cinco principios o derechos de los ciudadanos en lo relativo a la AI, que se resumen en[32]:

- Sistemas seguros y efectivos.
- Protección contra la discriminación de los algoritmos.
- Privacidad de los datos.
- Notificación y explicación.
- Alternativa, evaluación por un ser humano y proceso de corrección en caso de fallo de la AI *(fallback)*.

Dentro de su cuarto principio, en lo relativo a la explicabilidad de los modelos de AI, establece, entre otros, que[33]:

> Los diseñadores, desarrolladores e implantadores de sistemas automáticos deben proporcionar documentación en lenguaje sencillo y generalmente accesible que incluya descripciones claras del funcionamiento general del sistema. [...]
>
> Los sistemas automáticos deben acompañarse de explicaciones que sean técnicamente válidas, significativas y útiles para usted y para cualquier operador u otras personas que necesiten entender el sistema. [...]
>
> Los sistemas automáticos deben proporcionar notificaciones de uso demostrablemente claras, oportunas, comprensibles y accesibles, y **explicaciones sobre cómo y por qué el sistema ha tomado una decisión** o realizado una acción.

5. Principios sobre inteligencia artificial (OCDE)

Los Principios de la OCDE sobre inteligencia artificial promueven el uso de una AI digna de confianza y que respete los derechos humanos y los valores democráticos. Fueron adoptados en mayo de 2019 por los 38 países miembros de la OCDE. Fueron los primeros principios de este tipo suscritos por gobiernos, e incluyen recomendaciones concretas para la política y la estrategia públicas sobre AI.

Entre otros, establecen que "los responsables de la AI deben comprometerse con la transparencia y la divulgación responsable de los sistemas de AI. Para ello, deben proporcionar información significativa, adecuada al contexto y coherente con el estado de la técnica […] para que los afectados por un sistema de IA puedan comprender el resultado"[34]. El Observatorio de Políticas de AI de la OCDE, lanzado en febrero de 2020, tiene como objetivo ayudar a los responsables a aplicar estos Principios.

6. Discussion paper on machine learning for IRB models (EBA)

Por su relevancia en el sector bancario, es destacable el *Discussion paper on machine learning for IRB models,* de la Autoridad Bancaria Europea (EBA), publicado en noviembre de 2021 (Fig. 5).

El documento tiene como objetivo analizar la relevancia de los posibles obstáculos para la implementación de técnicas de aprendizaje automático en el ámbito del enfoque IRB de cálculo de capital en entidades financieras, incluye los desafíos y los beneficios potenciales del uso de estas técnicas, y establece ciertos principios y recomendaciones[35]. Un eje central del documento es, lógicamente, cómo hacer compatible el uso de estas técnicas con el cumplimiento de la regulación europea sobre capital (CRR[36]).

[31] White House OSTP (2022).
[32] Ibíd.
[33] Ibíd.
[34] OECD (2019).
[35] Ver un análisis detallado en Management Solutions (2021).
[36] CRR: Capital Requirements Regulation, regulación central sobre capital en entidades financieras en Europa.

En lo relativo a la interpretabilidad de los modelos, el documento lo aborda bajo el epígrafe de "Preocupaciones sobre el uso de las técnicas de aprendizaje automático", y afirma[37]:

> Las principales preocupaciones derivadas del análisis de los requisitos de la CRR se refieren a la complejidad y fiabilidad de los modelos de ML, en los que los principales retos parecen ser la interpretabilidad de los resultados, la gobernanza, con especial referencia a las mayores necesidades de formación del personal, y la dificultad de evaluar la capacidad de generalización de un modelo (es decir, evitar el sobreajuste).
>
> Para comprender las relaciones subyacentes entre las variables explotadas por el modelo, los profesionales han desarrollado varias técnicas de interpretabilidad [...] [y] la elección de cuál de estas técnicas utilizar puede plantear un reto en sí misma, ya que a menudo estas técnicas solo permiten una comprensión limitada de la lógica del modelo.

Más allá de esto, el documento introduce la necesidad de encontrar un equilibrio entre complejidad e interpretabilidad del modelo, y, a diferencia de otra regulación, baja a un nivel más técnico al recomendar a las entidades financieras:

a. Analizar de forma estadística: i) la relación de cada variable de entrada con la variable de salida, *ceteris paribus*; ii) el peso global de cada variable de entrada en la determinación de la variable de salida, para detectar qué variables influyen más en la predicción del modelo. Estos análisis son especialmente pertinentes cuando no es posible determinar una representación estrecha y puntual de la relación entre la variable de salida del modelo y las variables de entrada debido a la complejidad del modelo.

b. Evaluar la relación económica de cada variable de entrada con la variable de salida para garantizar que las estimaciones del modelo son plausibles e intuitivas.

c. Presentar un documento de síntesis que explique de forma sencilla el modelo a partir de los resultados de los análisis descritos en el punto a. El documento deberá describir como mínimo:
   i. Los factores clave del modelo.
   ii. Las principales relaciones entre las variables de entrada y las predicciones del modelo.

   Los destinatarios del documento son todas las partes interesadas, incluido el personal que utiliza el modelo con fines internos.

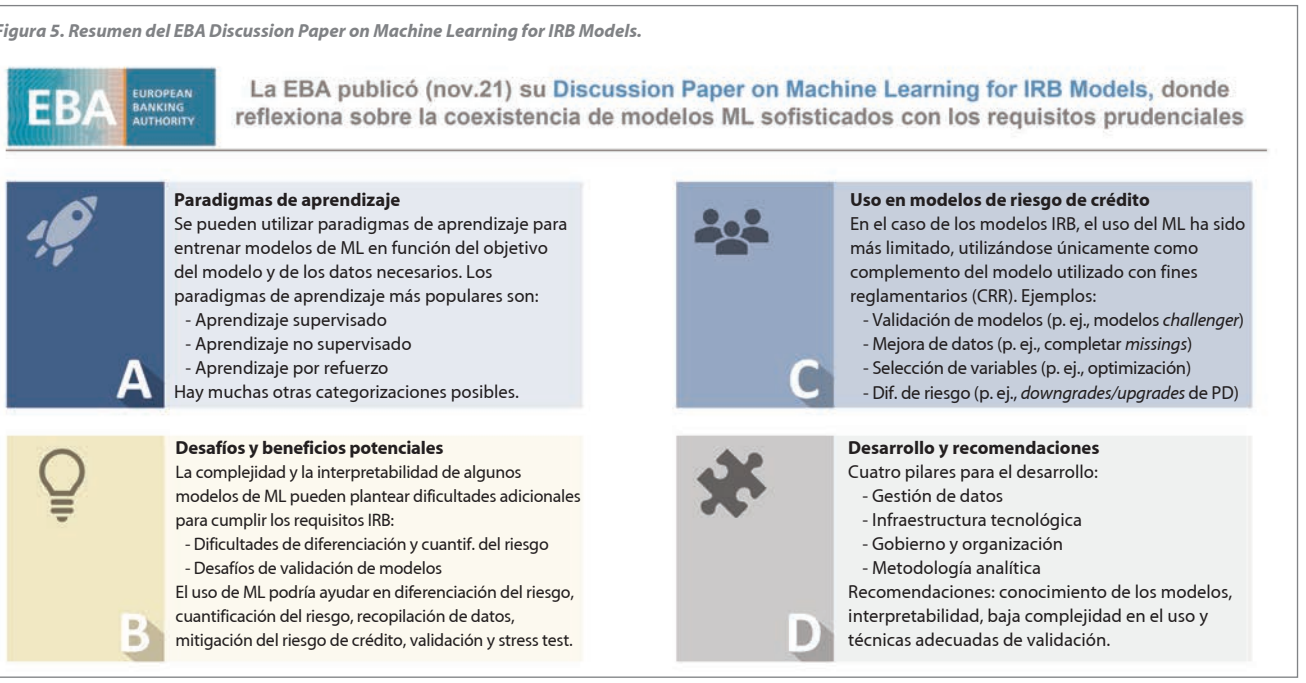
d. Garantizar la detección de posibles sesgos en el modelo (por ejemplo, un ajuste excesivo a la muestra de entrenamiento).

En la práctica, al tiempo que la industria bancaria espera la versión final del documento consultivo de la EBA, la mayor parte de entidades que utilizan técnicas de aprendizaje automático en sus modelos IRB ya están adaptando sus marcos de desarrollo, seguimiento y validación de modelos para asegurar su cumplimiento en el futuro.

Un elemento común a todas las referencias regulatorias mencionadas, como se puede apreciar, es la necesidad de proporcionar una explicación a los ciudadanos sobre el uso de la AI, y de hacerlo en dos niveles: la interpretabilidad y transparencia del modelo de AI en su conjunto, y la capacidad de explicar una decisión concreta del modelo, en caso de ser requerido.

[37] EBA (2021).

Figura 5. Resumen del EBA Discussion Paper on Machine Learning for IRB Models.



Más allá de las referencias regulatorias descritas, hay un gran número de publicaciones, principios, directrices y borradores de regulación en múltiples jurisdicciones que abordan la interpretabilidad de los modelos de AI, tanto de ámbito general como sectoriales, y tanto regionales como locales de cada país; la selección expuesta en esta sección incluye las consideradas de mayor ámbito y potencial influencia.

Impactos en organización y procesos

Un principio esencial de la XAI como disciplina es que, más allá del desarrollo de las técnicas específicas de explicabilidad o de la construcción de modelos inherentemente interpretables, esta explicabilidad e interpretabilidad se debe integrar en la organización y los procesos de la compañía.

Llevado a la práctica, este principio supone el desarrollo y la puesta en funcionamiento de un marco de XAI, que se puede estructurar en cuatro elementos:

1. Técnicas de interpretabilidad de los modelos de AI
2. Integración en los procesos de gestión del riesgo de modelo (MRM)
3. Soporte tecnológico
4. Factor humano

1. Técnicas de interpretabilidad de los modelos de AI

El núcleo de un marco de XAI lo constituyen las técnicas de interpretabilidad y explicabilidad, que de forma resumida se pueden clasificar en tres aspectos:

- Interpretabilidad del diseño del modelo: esto incluye analizar cómo se comportaría el modelo en diferentes escenarios (p. ej. ataques adversarios, escenarios extremos...), comprender cómo funcionan los submodelos y los conjuntos ("ensembles") de modelos, e integrar la interpretabilidad en el diseño del modelo aplicando restricciones durante su desarrollo.
- Interpretabilidad de los resultados del modelo: se refiere a detectar qué variables y cómo influyen en la predicción del modelo a través de la interpretabilidad local (LIME, SHAP, etc.) y global (PDP, importancia de las variables, modelos subrogados, análisis de sensibilidad); a evaluar el sentido económico de cada variable (p. ej., análisis de casos de uso de una muestra representativa de datos), y a asegurar que la documentación del modelo lo describe correctamente, incluidas las variables de entrada y su relación con los resultados.
- Otros aspectos: garantizar la detección de posibles sesgos en el modelo (p. ej., sobreentrenamiento, datos de entrada sesgados, errores en los datos) y supervisar periódicamente el modelo, especialmente cuando cambie su alcance o cuando se aplique a datos distintos de los de desarrollo.

Por su importancia, las principales técnicas de interpretabilidad y explicabilidad se desarrollan en la siguiente sección.

2. Integración en los procesos de gestión del riesgo de modelo (MRM)

La interpretabilidad de los modelos de AI es una característica que trasciende el desarrollo e impacta a lo largo de toda la cadena del ciclo de vida de los modelos, y por tanto en todo el marco de la gestión del riesgo de modelo. Un resumen no exhaustivo de la incorporación de XAI en el marco de MRM de una compañía incluye revisar los siguientes elementos:

- **Gobierno:** actualizar el marco de organización y gobierno para incorporar XAI; evaluar el impacto de la regulación aplicable a los modelos de AI; actualizar el sistema de tiering de los modelos para contemplar la falta de interpretabilidad como un mayor riesgo; actualizar el inventario y los procedimientos de inventariado de los modelos para incorporar los elementos de XAI (p. ej. atributos específicos para modelos de AI).
- **Desarrollo:** actualizar las políticas y procedimientos de desarrollo de los modelos, así como los requisitos de documentación; evaluar imparcialidad (*fairness*) y sesgos, interpretabilidad de inputs, diseño y resultados, datos, riesgo de proveedores, métricas de capacidad predictiva, límites al uso de los modelos de AI, etc.; realizar análisis de sensibilidad de los modelos de AI para identificar vulnerabilidades; incluir en el marco de desarrollo tests específicos para XAI.
- **Seguimiento:** actualizar el marco de seguimiento de los modelos y completarlo con tests específicos de XAI; revisar los umbrales y las acciones derivadas de su incumplimiento; desarrollar sistemas de alerta temprana para detectar cambios en los modelos de AI; revisar el cumplimiento del apetito al riesgo de modelo; valorar la necesidad de desarrollar un módulo de seguimiento ad hoc para modelos de aprendizaje dinámico (i.e. que se recalibran automáticamente sin intervención humana).

- **Validación:** actualizar el marco de validación interna para detectar posibles riesgos asociados a los modelos de AI e incorporar tests de XAI; establecer un marco de validación cruzada para garantizar la calidad de los modelos de AI; evaluar el impacto de los cambios en el entorno de producción en los modelos de AI.

- **Implementación:** actualizar el proceso de implementación del modelo para incorporar tests propios de las características de XAI; actualizar, en su caso, la plataforma tecnológica para permitir la puesta en producción de los modelos de AI.

- **Uso:** actualizar los procedimientos de uso de los modelos de AI para determinar su adecuación al contexto en que se van a emplear; revisar y completar la formación a usuarios respecto a los modelos de AI; actualizar los protocolos para detectar posibles situaciones de mal uso o explotación de los modelos.

- **Auditoría:** implementar un marco de auditoría de los modelos de AI para asegurar su adecuada implementación y uso; establecer tests de XAI para la auditoría de los modelos de AI; evaluar la adecuación de los sistemas de control interno para garantizar la calidad de los modelos de AI; analizar los registros de auditoría para detectar posibles riesgos asociados a los modelos de AI.

Por tanto, el uso de modelos de AI conlleva una revisión completa de las políticas y procedimientos a lo largo de todo el ciclo de vida del modelo para incorporar, como mínimo, los elementos propios de XAI.

3. Soporte tecnológico

La implementación de un marco de XAI tiende a comenzar por herramientas departamentales, y tan pronto como alcanza un mínimo nivel de madurez, requiere de soluciones tecnológicas profesionales para dar soporte a los aspectos propios de la interpretabilidad de los modelos de AI.

Estas soluciones pueden clasificarse en dos grupos:

- Interpretabilidad: desarrollo de sistemas que implementen las técnicas de interpretabilidad de forma estandarizada y homogénea. Deben permitir realizar la interpretación de los modelos de forma automática, fácilmente configurable y con una alta calidad, incorporando las técnicas más comunes y proporcionando flexibilidad para añadir nuevas técnicas conforme se desarrollen[38].

- Gobierno de modelos: desarrollo o actualización de los sistemas de gobierno de modelos para dar soporte a los aspectos de XAI en MRM (inventario, *tiering*, documentación, etc.), asegurando así que los modelos disponibles cumplan los requisitos de calidad, seguridad y explicabilidad requeridos[39].

Más allá de esto, es recomendable una aproximación holística que abarque todos los aspectos del marco de XAI. Esto incluye el uso de herramientas de análisis de datos, el desarrollo de APIs para la integración de los sistemas de interpretabilidad y gobierno de modelos antes descritos, la creación de mecanismos de seguridad y auditoría, y la definición de protocolos para garantizar el cumplimiento de los estándares de calidad y explicabilidad.

4. Factor humano

Un cuarto elemento en la integración de XAI en la organización y procesos es la consideración del factor humano. Esto incluye, entre otros:

- **Captación y retención del talento:** desarrollo de programas para la captación y retención del talento especializado en XAI, para asegurar la presencia de profesionales con los conocimientos técnicos y la experiencia necesaria para aplicar XAI en la compañía, lo que es especialmente relevante en un mercado laboral con escasez de este perfil profesional.

- **Formación:** desarrollo de programas de formación para equipos de desarrollo de modelos de AI, equipos de gobierno de modelos y usuarios de los modelos de AI, con el fin de asegurar que todos los involucrados comprendan los principios básicos de XAI y cómo aplicarlos en el contexto específico de la compañía.

- **Cultura:** desarrollo de una cultura en la compañía que potencie el uso y la explotación de la explicabilidad y la interpretabilidad de los modelos de AI. Esto puede incluir la adopción de metodologías ágiles para el desarrollo de modelos de AI, la creación de una cultura de colaboración entre los equipos de desarrollo de modelos y de gobierno de modelos, y la consideración de la explicabilidad como un factor crítico en la aprobación de los modelos de AI.

- **Gestión del cambio:** desarrollo de programas de gestión del cambio para asegurar la adecuada adopción de XAI por parte de los equipos de la compañía que trabajan con modelos de AI. Esto incluye motivar a los equipos de desarrollo, el análisis de los costes y beneficios de la explicabilidad, la definición de protocolos de comunicación con terceros, etc.

En conclusión, la explicabilidad y la interpretabilidad de los modelos de AI son aspectos clave que deben integrarse en la organización y los procesos de la compañía mediante un marco apropiado y completo de XAI, lo que resulta esencial para garantizar el uso de estos modelos conforme a la regulación y las buenas prácticas.

[38] A este respecto, Management Solutions dispone de ModelCraft™, un sistema propietario de AutoML y modelización por componentes, que incorpora un módulo completo de interpretabilidad. Ver Management Solutions (2023).

[39] Management Solutions dispone de Gamma™, un sistema propietario de gobierno de modelos que cubre todos los aspectos mencionados. Ver Management Solutions (2022).

Técnicas de interpretabilidad: estado del arte

"Con mucha diferencia, el mayor peligro de la inteligencia artificial es que las personas concluyen demasiado pronto que la entienden".
Eliezer Yudkowsky[40]

Concepto

La comunidad científica[41,42] propone numerosas definiciones de "interpretabilidad" y "explicabilidad" de un modelo, y tiende a hacer una cierta distinción entre ellas, aunque en la práctica estos conceptos se suelen usar indistintamente. Con carácter general, la interpretabilidad estaría ligada a la capacidad para explicar a un ser humano los resultados de un modelo (su relación causa-efecto), mientras que la explicabilidad está asociada con la comprensión de la lógica interna del algoritmo, cómo se diseña y entrena, y los pasos que se siguen en la toma de decisiones para llegar a un resultado concreto.

Algunas definiciones académicas a este respecto son:

- La interpretabilidad es la capacidad de explicar o presentar en términos comprensibles para un ser humano[43].
- La interpretabilidad es el grado en que un ser humano puede comprender la causa de una decisión[44].
- La explicabilidad de un resultado de un modelo es la descripción de cómo se ha producido el output arrojado por el modelo[45].
- La explicabilidad es la medida en que la mecánica interna de un sistema de aprendizaje automático se puede explicar en términos humanos[46].

La necesidad de la explicabilidad e interpretabilidad de modelos ha favorecido la aparición de técnicas cada vez más sofisticadas de interpretabilidad local y global de los resultados de los modelos, y la situación actual es una cierta estandarización y convergencia en el uso de ciertas técnicas (p. ej., PDP, LIME o SHAP).

Al mismo tiempo, estas técnicas no resuelven por completo el problema de la interpretabilidad y, bajo determinadas circunstancias, pueden arrojar resultados contradictorios o sesgados, lo que convive con otros factores que pueden impactar en la interpretabilidad del modelo, como son:

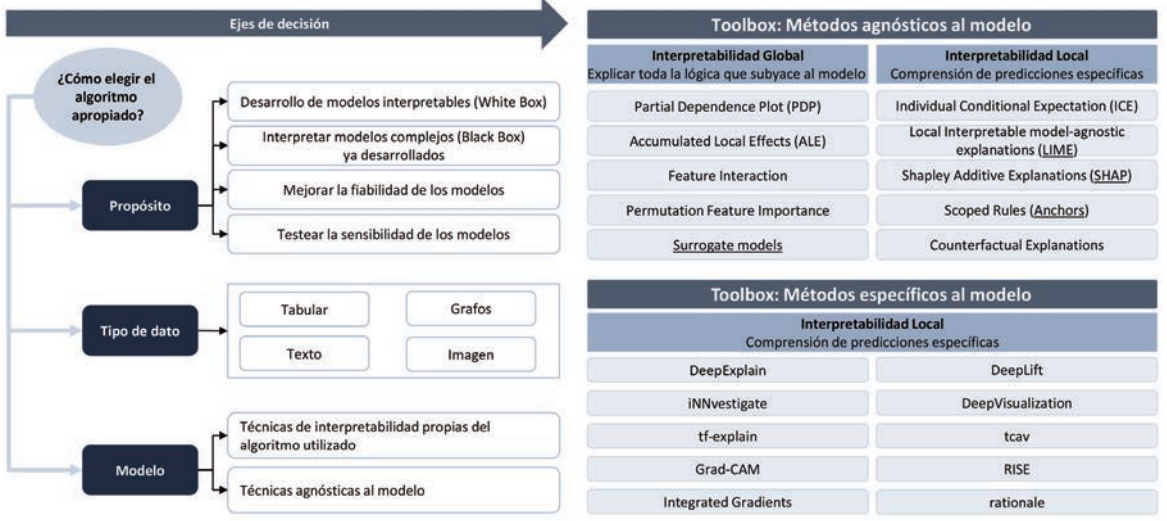
- La reproducibilidad de los resultados, el proceso de entrenamiento e implementación del modelo[47], la consistencia en sus predicciones y la explicación de la secuencia de predicciones más probables.
- Potenciales sesgos[48] en los datos de entrada.
- Imparcialidad *(fairness)*[49].
- Exactitud de la explicación[50].
- Solidez conceptual del modelo[51].

Para superar varias de estas dificultades, algunos investigadores[52] están desarrollando enfoques alternativos para la mejora de la interpretabilidad de los modelos de AI, fundamentalmente centradas en el desarrollo de modelos inherentemente interpretables *(white boxes)*.

En esta sección se describen las principales técnicas de interpretabilidad, consideradas estándares en la industria, y se recoge también el estado del arte sobre el desarrollo de *white boxes*.

[40] Eliezer Shlomo Yudkowsky (n. 1979), investigador y escritor estadounidense especializado en teoría de la decisión e inteligencia artificial, conocido por popularizar la idea de la inteligencia artificial amigable y abogar por la Singularidad.
[41] Gall, R. (2018). Redactor en Thoughtworks y The New Stack.
[42] Broniatowsky, D. (2021). Profesor asociado del Departamento de Gestión de Ingeniería e Ingeniería de Sistema, Universidad George Washington.
[43] Doshi-Velez, F., et al. (2017). Profesor de Informática en la Escuela Paulson de Ingeniería y Ciencias Aplicadas, Universidad de Harvard.
[44] Miller, T. (2019). Profesor en la Escuela de Computación y Sistemas de Información, Universidad de Melbourne.
[45] Broniatowsky D. (2021).
[46] Gall, R. (2018).
[47] Leventi-Peetz, A.-M., et al. (2022). Científico de la Oficina Federal para la Seguridad de la Información Alemana.
[48] Zhou, N., et al. (2021). Analista financiero senior en Wells Fargo.
[49] Ibíd.
[50] Jonathon Phillips et al. (2021). Profesor de Informática e Ingeniería, Instituto Nacional de Normas y Tecnología (NIST).
[51] Sudjianto, A., et al. (2021).
[52] Íbid.

Figura 6. Visión general de las técnicas de interpretabilidad



Técnicas de interpretabilidad más comunes

Las técnicas de interpretabilidad más comúnmente usadas se pueden agrupar según su enfoque[53]: interpretabilidad *post-hoc* y modelos inherentemente interpretables. Asimismo, existen estrategias complementarias que permiten mejorar el entendimiento del modelo.

Interpretabilidad post-hoc

Las técnicas de interpretabilidad *post-hoc*, o interpretabilidad de modelos *black box,* se centran en la explicación de la salida de modelos ya entrenados, a partir de la información que proporcionan los pesos asignados a cada variable de entrada y los resultados de los modelos. Estas técnicas son útiles para la comprensión de los resultados de los modelos, aunque no proporcionan información sobre el proceso de entrenamiento ni explican la lógica interna del algoritmo.

Se suelen dividir en técnicas de interpretabilidad global y local, en referencia a si la técnica explica todo el modelo en conjunto o únicamente los resultados en un subconjunto de observaciones o datos.

Las técnicas de interpretabilidad *post-hoc* más comunes son las siguientes (para un inventario más exhaustivo, véase Fig. 6):

- PDP *(Partial Dependence Plots,* curvas de influencia de la variable). Esta técnica permite visualizar la influencia de cada variable individual en la salida del modelo, excluyendo el resto de las variables.

- LIME *(Local Interpretable Model-agnostic Explanations)*. Esta técnica permite la explicación de resultados a nivel local, es decir, la explicación de los resultados de una instancia concreta en particular, a partir de la información de otros casos similares.

- SHAP *(SHapley Additive exPlanations).* Esta técnica permite la explicación local y global de los resultados de un modelo, es decir, la explicación de la influencia de cada variable en observaciones del modelo, y la importancia de cada variable en los resultados globales del modelo.

- *Anchors*. Consiste en la búsqueda de reglas de decisión que expliquen el resultado.

Modelos inherentemente interpretables

La interpretabilidad inherente, o interpretabilidad mediante modelos *white box*, se centra en el desarrollo de modelos que son interpretables por diseño o que se pueden convertir en interpretables por construcción, mediante una serie de condiciones dependientes del tipo de modelo (p. ej., redes neuronales[54], en concreto de tipo ReLu[55], y modelos basados en árboles[56], entre otros).

Estos modelos permiten una explicación de la lógica interna del algoritmo y de la secuencia de pasos que se dan para llegar a un resultado concreto, y permiten por tanto una mayor comprensión de los resultados, aunque su aplicabilidad en problemas complejos puede ser más limitada, dependiendo del tipo de algoritmo empleado.

Estrategias complementarias

También se puede citar el uso de estrategias que contribuyen a la interpretabilidad de los modelos, como son la simplificación del modelo para facilitar su interpretación, el uso de variables con sentido de negocio, el análisis de datos para identificar sesgos o falta de imparcialidad *(fairness)* en los *inputs* que dificulten la explicabilidad, o el análisis de la reproducibilidad del desarrollo del modelo o su implementación, entre otras.

[53] iDanae (2022).
[54] Yang, Z., et al. (2019). Departmento de Estadística y Ciencias Actuariales, Universidad de Hong Kong.
[55] Sudjianto. A., et al. (2011).
[56] Sudjianto. A., et al. (2021).

Interpretabilidad post-hoc

1. PDP

Los gráficos PDP[57] *(Partial Dependence Plots)* muestran cómo varía la predicción de un modelo de AI en función de una o dos variables independientes en la predicción, es decir, el efecto marginal de los predictores. Así, permiten evaluar el tipo de relación entre variables independientes y dependientes.

Sintéticamente:

- Los PDP muestran gráficamente en una curva la variación promedio de la predicción.
- Esta variación promedio se obtiene variando un predictor para todas las observaciones del *dataset*, y luego obteniendo el impacto medio en la predicción.
- Una variante de los PDP son los gráficos ICE[58] *(Individual Conditional Expectation)*, que análogamente muestran cómo varía una predicción para cada observación concreta, si se varía uno de los predictores del modelo, y se mantiene constante el resto.

2. LIME

LIME[59] *(Local Interpretable Model-agnostic Explanations)* es un método local que comprueba cómo varían las predicciones de un modelo cuando se perturban los datos introducidos. Para ello, LIME aplica los siguientes pasos:

- Generar datos sintéticos alrededor de la instancia de datos de entrada: LIME toma como punto de partida una única predicción y los datos de entrada que la generaron, y genera nuevos datos de entrada perturbando esta observación, obteniendo las correspondientes predicciones por el modelo de AI.
- Entrenar un modelo simple sobre los datos sintéticos: el dataset resultante compuesto por los datos de entrada perturbados y las predicciones generadas por el modelo se usa para entrenar un modelo que sí es interpretable (p. ej., modelos lineales, árboles de decisión).
- Explicar las predicciones del modelo simple en función de los datos originales: la importancia de cada variable en la predicción se obtiene, por ejemplo, en función de sus coeficientes en la regresión y su signo correspondiente.
- Calcular la explicabilidad: el porcentaje de explicabilidad por LIME es equivalente al coeficiente de ajuste del modelo lineal (p. ej., R^2). Por tanto, el modelo interpretable arroja una buena aproximación de las predicciones de manera local.

[57] Friedman, J. H. (2001). Profesor en el Departamento de Estadística, Universidad de Stanford.
[58] Goldstein, A., et al. (2015). Profesor en el Departamento de Estadística, The Wharton School, Universidad de Pensilvania.
[59] Ribeiro, M. T., et al. (2016). Investigador de Microsoft Research en el grupo de Sistemas Adaptativos e Interacción y Profesor Adjunto de la Universidad de Washington.

Caso de uso: admisión de préstamos en el sector bancario. Uso de PDP.

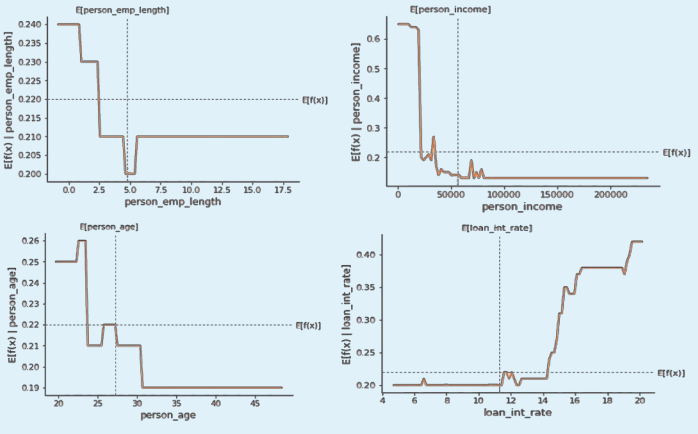
Los PDP se pueden aplicar a un caso de uso muy común en el sector bancario: la puntuación de clientes durante el proceso de concesión de préstamos para determinar su probabilidad de impago. En este ejemplo, se ha empleado una cartera anonimizada de préstamos hipotecarios con información de su actividad en los primeros tres años.

Se ha empleado un XGBoost, que es un modelo no aditivo de árboles, una característica que puede dificultar su explicación. Las variables empleadas por el modelo durante el entrenamiento incluyen el monto del préstamo, su finalidad, el régimen de propiedad del prestatario, los años de empleo en su trabajo actual y la tasa de interés, entre otras.

En este contexto, un área de negocio puede solicitar entender por qué el modelo asigna a un determinado cliente una probabilidad de impago determinada.

Un gráfico PDP muestra la explicación que se obtendría a nivel global de las variables que más han participado en el resultado, y que permitirían ver el impacto que distintos rangos de esa variable tienen en la predicción del modelo (Fig. 7).

Figura 7. PDP para las variables "años empleado" (en años), "salario" (EUR anual), "edad" (años) y "tasa de interés" (tanto por uno). El eje X representa la propia variable en estudio, y el eje Y representa el impacto que distintos rangos de cada variable tiene en la predicción del modelo.



Formalmente, una explicación usando modelos subrogados locales con LIME se puede definir como:

$$Explanation(X) = \arg\min_{g \in G} L(f, g, \pi_X) + \Omega(g)$$

donde:

f es un modelo *black box* (p. ej., un *random forest*), g es el modelo que explica f (p. ej., una regresión lineal).

L es la función de pérdidas que se trata de minimizar en el modelo (p. ej., error cuadrático medio), y que LIME minimiza.

Ω es la complejidad del modelo (p. ej., número de variables seleccionadas) decidida por el usuario.

G es el conjunto de posibles explicaciones del modelo f.

arg min representa el valor $g \in G$ que minimiza la función $L(f,g,\pi_X) + \Omega(g)$.

π_X representa la amplitud de las perturbaciones usadas para generar nuevas observaciones decidida por el usuario.

3. SHAP

SHAP[60] *(SHapley Additive exPlanations)* es un método de explicación de modelos basado en el Teorema de Valor de Shapley[61], que fue propuesto en 1952 para distribuir el valor de un juego entre los jugadores. SHAP se utiliza para explicar la importancia de cada variable (medida como el cambio promedio en la predicción del modelo cuando varía el valor de la variable) en una predicción concreta.

En concreto, SHAP utiliza una combinación de líneas de base, funciones de importancia local y el Teorema de Valor de Shapley para calcular la importancia de cada variable en una predicción individual.

En este método:

- Se calculan los valores de Shapley, donde las variables independientes se interpretan como jugadores que colaboran para recibir el *payout*.
- Los valores de Shapley se corresponden con la contribución de cada variable a la predicción del modelo.
- El *payout* es la predicción concreta realizada por el modelo menos el valor promedio de todas las predicciones.
- Los jugadores se "reparten" este *payout* en función de su contribución, y este reparto viene calculado por los valores de Shapley y refleja la importancia de cada variable.

Este método también permite obtener interpretaciones a nivel global obteniendo el promedio de las contribuciones de cada variable para cada predicción del modelo.

Formalmente, los valores de Shapley se pueden definir como la contribución de cada variable al resultado del modelo, pesada en función de todas las posibles combinaciones de variables empleadas:

$$\phi_j(val) = \sum_{S \subseteq \{1,\dots,p\}/\{j\}} \frac{|S|!(p-|S|-1)!}{p!} (val(S \cup \{j\}) - val(S))$$

donde val se corresponde con la predicción del modelo para variables incluidas en el conjunto S, respecto a la predicción para variables no incluidas en S:

$$val = \int f(x_1 \dots x_p) dP_{x \notin S} - E_X(f(X))$$

donde:

x es el vector de variables usadas en el modelo.

S es un subconjunto de x.

p es es el número de variables usadas en el modelo.

$dP_{(x \notin S)}$ representa el conjunto de variables no incluidas en S respecto a las que se realiza la integración.

E es el valor esperado de la predicción de X con el modelo f.

Usando estos valores, SHAP se puede utilizar para obtener una explicación local al modelo como:

$$Expl(x) = E_X(f(X)) + \sum \phi_j x_j$$

Por último, SHAP también es capaz de calcular explicaciones locales a través de la agregación de valores Shapley en un conjunto de datos.

4. Anchor

Anchors[62] es un método que explica predicciones individuales (i.e., locales) de modelos de clasificación *black box*, mediante la búsqueda de reglas de decisión llamadas "anchors" que expliquen el resultado.

- Al igual que en LIME, se toma como punto de partida una única predicción y los datos de entrada que la generaron, y se generan nuevos datos de entrada perturbando esta observación, obteniendo las correspondientes predicciones por el modelo de AI.

[60] Lundberg, S. M., et al. (2017). Investigador en la Escuela Paul G. Allen de Informática, Universidad de Washington.
[61] Shapley, L. (1953). Profesor de la Universidad de California en Los Ángeles, perteneciente a los departamentos de Matemáticas y Economía.
[62] Ribeiro, M. T.; Singh, S.; Guestrin, C. (2018). Investigador de Microsoft Research en el grupo de Sistemas Adaptativos e Interacción y Profesor Adjunto de la Universidad de Washington.

- La explicación local de la predicción se obtiene buscando reglas de tipo *if-else* que sean capaces de explicar el resultado del modelo. Se considera que una regla explica la predicción si cambios en otras variables independientes no consideradas en la regla no la modifican.

Formalmente, un anchor A se define como:

$$\text{Prec}(A) = \mathbb{E}_{\mathcal{D}(z|A)}[1_{f(x)=f(z)}] \geq \tau, \qquad A(x) = 1$$

donde:

f es un modelo *black box*.

D es una distribución arbitraria según la cual se perturba X.

X es una observación del *dataset* a explicar, y Z es una muestra de D.

Prec es la precisión en la explicación y τ es la precisión requerida.

Una manera de encontrar un anchor dada cualquier distribución D es buscar que la precisión supere un umbral con una cierta probabilidad $(1 - \delta)$, de manera que:

$$P(Prec(A) \geq \tau) \geq 1 - \delta$$

Desarrollo de modelos inherentemente interpretables *(white box)*

Los modelos inherentemente interpretables *(white box)* se basan en el diseño de algoritmos que, por diseño, son interpretables y permiten la explicación de los resultados tanto a nivel global como local.

Los modelos *white box* generalmente se agrupan según el tipo de algoritmo empleado:

- Modelos lineales, como las regresiones lineales o logísticas.
- Modelos basados en árboles, como los árboles de decisión o los árboles aleatorios.
- Modelos basados en reglas, como los sistemas basados en reglas *(rule-based systems)*.
- Redes neuronales profundas, con funciones de activación como ReLU o el uso de capas intermedias, sujetas a ciertas restricciones que las hacen inherentemente interpretables[63].

[63]Yang, Z., et al. (2019). Investigador en el Departamento de Estadística y Ciencias Actuariales, Universidad de Hong Kong.

Caso de uso: Admisión de préstamos en el sector bancario. Uso de SHAP

Si se aplica SHAP sobre el mismo caso planteado para la creación de PDP, se obtiene información local adicional sobre una decisión del modelo para un determinado cliente.

En este caso, emplear SHAP sobre una muestra de observaciones resulta en valores de Shapley completamente distintos y con signo variable dependiendo de las características del cliente que ha pedido el préstamo. Incluso para clientes que reciben la misma tasa de interés, se observa que la influencia de esta variable varía debido a la mayor o menor importancia de las otras variables del modelo.

No obstante, se observa una tendencia con sentido de negocio: a mayor tasa de interés, mayor es la contribución de esta variable en el modelo a una probabilidad de impago mayor. Por ello, la media de los valores de Shapley de cada variable usada como interpretación global del modelo puede llevar a errores en la explicación si se interpreta como una generalización (Fig. 8).

Los valores de Shapley dan una explicación de casos particulares como el siguiente, donde se observa que la probabilidad de impago[1] de un cliente está determinada por las condiciones solicitadas de la hipoteca, historial crediticio y condiciones laborales (p. ej., salario) (Fig. 9).

Figura 8. Valores de Shapley para la variable "tasa de interés" en toda la muestra frente a esa variable. La gráfica de barras grises muestra la distribución de la variable.

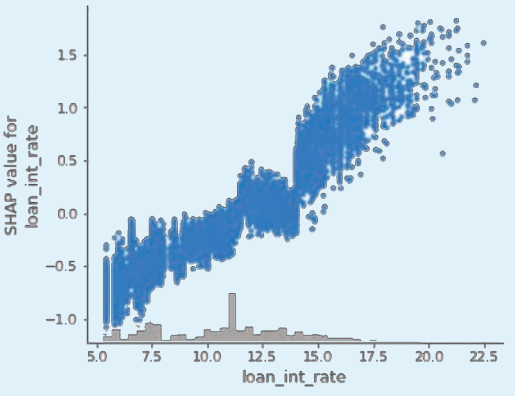
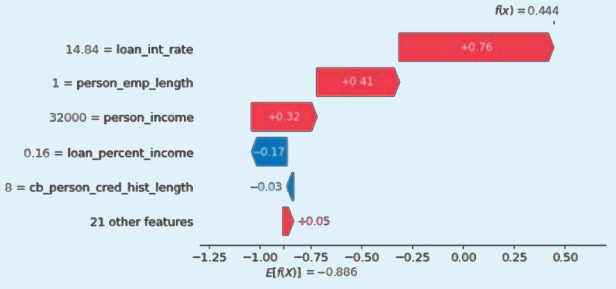
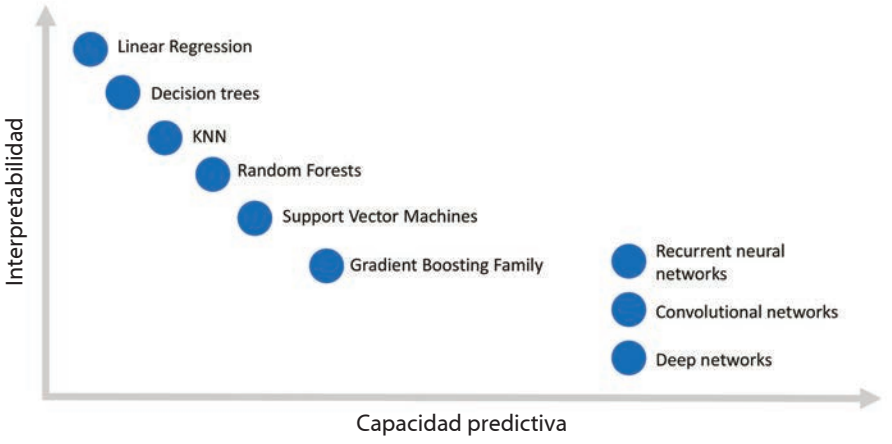


Figura 9. Valores de Shapley que influyen en la predicción de un cliente con préstamo denegado[2].



[1]Escala del gráfico mostrada en log-odds (0 corresponde a una probabilidad 50%).
[2]Gráfico en escala log-odds.

Figura 10. Balance entre interpretabilidad y capacidad predictiva por familias de modelos (incluyendo white y black boxes).



El desarrollo de estos modelos se suele basar en limitaciones sobre los parámetros a optimizar, que permiten que el modelo sea interpretable, a diferencia de las *black box,* aunque a cambio sean menos precisos (Fig. 10). Estas limitaciones incluyen usar solo variables con sentido de negocio, o restringir:

- El número de variables seleccionadas por el modelo para su predicción.
- El número de variables explicadas por el modelo.
- El grado de complejidad de las reglas de decisión.
- El número de pasos en la predicción.
- La profundidad de los árboles de decisión.
- La longitud y profundidad de las redes neuronales.

Gracias al desarrollo de modelos inherentemente interpretables, se pueden obtener resultados más precisos, ya que permiten una mayor comprensión de la información, lo que a su vez permite una mejor toma de decisiones. Esto es especialmente necesario en aquellos sectores en los que la interpretabilidad es un factor crítico para las decisiones finales.

A continuación, se detallan dos aspectos relevantes para la construcción de modelos inherentemente interpretables: el concepto y desarrollo de aprendizaje supervisado y no supervisado interpretable, y la aplicación de otros factores en el ámbito de la interpretabilidad.

1. Aprendizaje supervisado y no supervisado interpretable

Pese a que las líneas de investigación actuales están avanzando hacia el desarrollo de modelos inherentemente interpretables, no existe un formalismo matemático que describa por completo la construcción de estos modelos bajo cualesquiera condiciones iniciales y algoritmos empleados.

El estado del arte es la construcción de estos modelos bajo condiciones iniciales que los convierten en más fácilmente interpretables o equivalentes a otros modelos interpretables. Una de las maneras de definir esta condición de interpretabilidad en el entrenamiento del modelo es modificar la función de pérdida[64] a minimizar durante su entrenamiento, incluyendo una penalización por baja interpretabilidad, que depende de una condición impuesta de interpretabilidad en el modelo f:

$$Min\left(\frac{1}{n}\sum Loss(f,\ z_i) + C \cdot InterpretabilityPenalty(f)\right)$$

Por ejemplo, la *sparsity* es una de las condiciones empleadas en el desarrollo de modelos para calificar un modelo como más explicable respecto al resto. Esta condición se puede añadir a la función de pérdidas como:

$$Min\left(\frac{1}{n}\sum Loss(f,\ z_i) + \varphi(f)\right)$$

tal que $\varphi(f)$ sea una función de regularización que penalice la pérdida siendo proporcional a la *sparsity* del modelo (p. ej., si la *sparsity* es reducida, ese término de la función de pérdida también lo será).

[64] Rudin, C., et al. (2022). Catedrática de Informática, ECE, Estadística y Bioestadística y Bioinformática en la Universidad de Duke.

Algunos autores[65] han formalizado la creación de modelos inherentemente interpretables para ciertas familias como: modelos basados en árboles de decisión (p. ej., SIMTree o *single-index model tree*, que genera un modelo de árboles de un solo índice para cada nodo terminal), o la simplificación de redes con función de activación ReLu, que se demuestran equivalentes a un conjunto de modelos lineales locales.

Otros autores[66] se han centrado en definir las características que deberían cumplir los modelos inherentemente interpretables, con objeto de optimizarlas durante el proceso de modelización, tales como:

- Aditividad de las variables de entrada, de manera que sus efectos se agreguen en el modelo de manera sencilla.
- *Sparsity*, y la optimización de modelos para cumplir esta condición.
- Linealidad de las variables de entrada frente al output del modelo.
- Monotonía, de manera que para el mayor número de rangos posibles la relación entre la variable de entrada y el resultado a predecir sea monótona.
- Desacoplamiento de conceptos en el entrenamiento de redes neuronales, que se refiere a mantener en la medida de lo posible la información sobre un concepto determinado en caminos determinados de la red (i.e., frente a información de un mismo concepto que atraviesa un mayor número de neuronas y caminos dispersos en la red).
- Reducción de la dimensionalidad como herramienta visual para facilitar las explicaciones *post-hoc* a humanos.

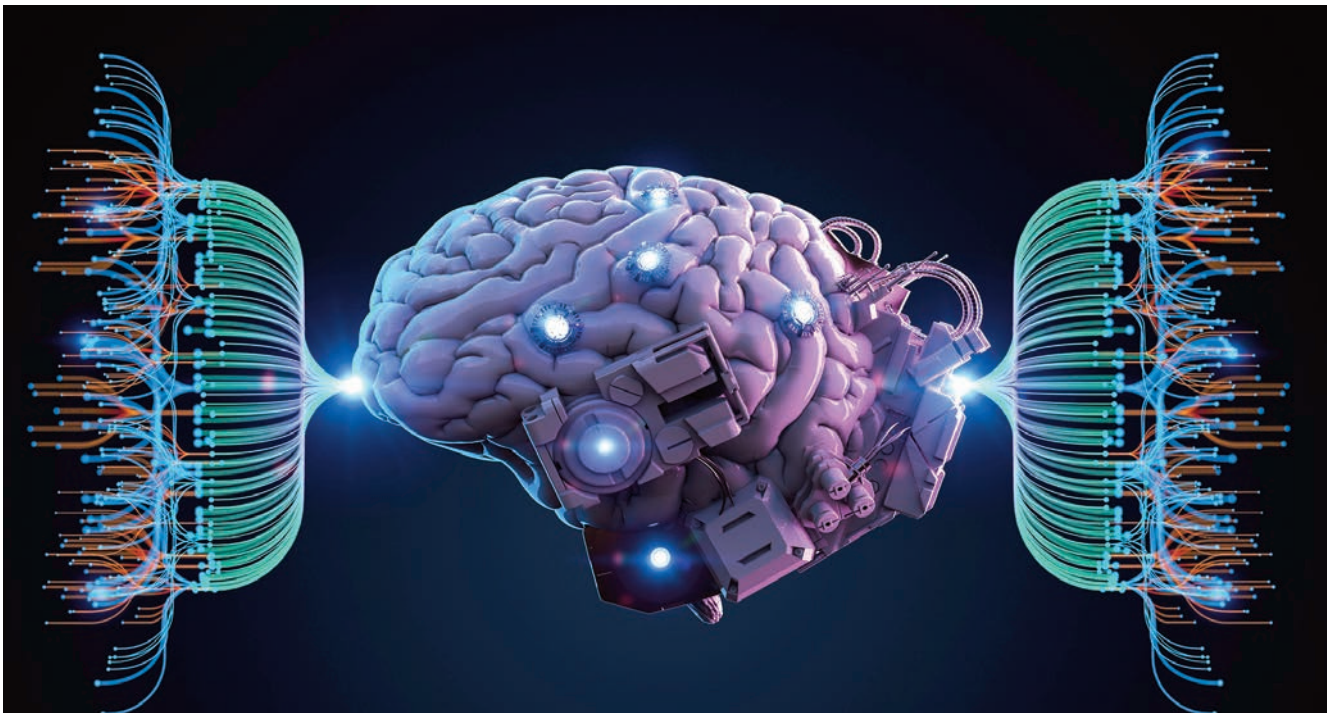
2. Otros factores de impacto

En combinación con los desafíos mostrados en esta sección, existen elementos clave adicionales que pueden considerarse para mejorar la interpretabilidad del modelo, tales como la imparcialidad del modelo *(fairness)*, la ausencia de sesgos en los datos de entrada, potenciales componentes expertos, o un rendimiento adecuado y un marco de control de los modelos que evite errores en su interpretación.

Por su relevancia, como se ha indicado anteriormente[67], estos elementos también han sido destacados en el AI Act como requisitos imprescindibles para sistemas de AI de riesgo elevado.

En la actualidad, existen múltiples técnicas y métodos para evaluar el rendimiento de los modelos, y prevenir problemas de sobreentrenamiento. Existen también varias maneras de evaluar el error producido por el modelo y el equilibrio entre el error por sesgo *(bias)* y por varianza. No obstante, debido a limitaciones en el uso de datos personales introducidas por la normativa de protección de datos, por el momento una de las mayores complejidades se encuentra en detectar y corregir potenciales imparcialidades (p. ej., por raza, género, religión, orientación política o sexual, creencias o posición social) en los modelos de AI, especialmente cuando las variables no se han almacenado y por tanto no están disponibles para el análisis.

[65] Sudjianto. A., et al. (2021).
[66] Rudin, C., et al. (2022).
[67] Véase la sección sobre regulación.



A este respecto, en el ámbito académico se han propuesto varias técnicas de identificación de variables de entrada imparciales, tales como:

- Análisis de interpretabilidad a través de redes bayesianas causales[68] como cuantificación del grado de imparcialidad del modelo.
- Definición[69] de métricas de imparcialidad, tales como la paridad demográfica, paridad del ratio predictivo, falsos positivos y falsos negativos iguales en segmentos susceptibles a sesgo.

Entre estas métricas destaca la imparcialidad o equidad contrafactual *(counterfactual fairness),* que proporciona una medida de cuán parecidos son los resultados de un modelo frente a individuos (observaciones) con las mismas características, pero con atributos sensibles a sesgos o parcialidad ligeramente distintos.

Ventajas y desventajas de las técnicas más comunes de interpretabilidad

Por regla general, no existe una técnica de interpretabilidad que permita dar una explicación única, global e intuitiva ante cualquier escenario. Las técnicas de interpretabilidad se suelen combinar bajo varios casos de uso y escenarios para verificar que dan explicaciones reproducibles y aplicables a distintos grupos de observaciones.

A la hora de seleccionar cuáles de estas técnicas emplear, es conveniente tener en cuenta las ventajas o desventajas de su aplicación (Fig. 11).

Últimas tendencias y retos

A pesar de los avances en la interpretabilidad de los modelos, todavía se plantean retos y desafíos en la explicación de los resultados (Fig. 12).

En primer lugar, la interpretabilidad de los modelos se ve aún limitada por una serie de factores como la reproducibilidad de los resultados[70], el proceso de entrenamiento e implementación del modelo, la consistencia de sus predicciones, la explicación de la secuencia de predicciones más probables, los sesgos en los datos de entrada, la imparcialidad *(fairness)* y la exactitud de la explicación.

En segundo lugar, las técnicas de XAI actualmente disponibles solo permiten explicaciones locales (i.e., para una única observación o dato) o globales (i.e., para el conjunto de datos). Esto genera la necesidad de desarrollar técnicas que permitan explicaciones en entornos intermedios, es decir, explicar resultados para grupos o subconjuntos de datos[71].

[68] Oneto, L.,Chiappa, S., (2020).
[69] Zhou, N., et al. (2021). Analista financiero senior en Wells Fargo.
[70] Leventi-Peetz, A.-M., et al. (2022).
[71] Si bien SHAP es capaz de obtener explicaciones sobre subconjuntos a través de medias ponderadas de valores de Shapley, es posible que estas explicaciones varíen en función de la granularidad del subconjunto datos.

Figura 11. Comparativa de las técnicas de interpretabilidad más comunes

Técnica	Ventajas	Desventajas
1 PDP (Partial Dependence Plot)	✓ Fácil de aplicar e intuitiva implementación. ✓ El cálculo de los gráficos de dependencia parcial tiene una interpretación causal.	X Por diseño, no permite ver el impacto de más de 2 variables intuitivamente en el gráfico. X No explica cómo varia la explicación según una única variable independiente si varían el resto de variables independientes.
2 LIME (Local interpretable model-agnostic explanations)	✓ Dada una predicción, este método evalúa el impacto de ligeras modificaciones en los inputs. ✓ Se utiliza un modelo subrogado local para evaluar las diferencias entre las predicciones originales y las modificadas, así como las variables más importantes que contribuyen a la predicción. ✓ El método es agnóstico respecto al modelo de predicción utilizado.	X Se asume linealidad local. X Puede arrojar explicaciones contrarias en distintos subconjuntos de datos, por lo que es necesario verificar las explicaciones en rangos representativos del *dataset*. X No da una explicación global del modelo.
3 SHAP (SHapley Additive exPlanations)	✓ Calcula la contribución de cada variable a una predicción específica. ✓ No asume linealidad local. ✓ Puede cubrir la importancia global de las características para todo el conjunto de datos. ✓ Agnóstico respecto al modelo de predicción utilizado. ✓ Muy costoso computacionalmente y asume que las variables del modelo son independientes.	X Puede arrojar explicaciones contrarias en distintos subconjuntos de datos, por lo que es necesario verificar las explicaciones en rangos representativos del *dataset*. X No da una explicación global del modelo.
4 Anchors	✓ Agnóstico al tipo de modelo y fácil de interpretar. ✓ Recoge comportamientos no lineales de modelos complejos.	X Gran número de hiperparámetros (forma de perturbación, precision...). X Requiere discretizar variables continuas en muchos casos, pudiendo llevar a errores en la interpretación.
5 *Construcción de Modelos "White Box"*	✓ Reduce el esfuerzo en interpretación de modelos tras el entrenamiento, y durante su ciclo de vida. ✓ No lleva a contradicciones en la interpretación del modelo y facilita su uso. ✓ No requiere del empleo de modelos o técnicas adicionales *post-hoc*.	X Incrementa el esfuerzo durante la construcción del modelo. X No existen técnicas aplicables para todo tipo de modelos, por el momento.

Adicionalmente, sin un análisis en profundidad, los resultados arrojados por distintas técnicas de interpretabilidad a distintos niveles pueden parecer contradictorios en un inicio (p. ej., si se trata de comparar resultados globales "promedio" con resultados locales en un entorno).

En tercer lugar, todavía son necesarias mejoras en el desarrollo de modelos *white box,* ya que, a pesar de los avances en los últimos años, estos modelos aún no son capaces de competir en precisión con los modelos *black box* en problemas complejos.

Por último, la necesidad de explicar los modelos más complejos (p. ej., ciertos tipos de redes neuronales profundas) sigue siendo un reto aún no resuelto.

A este respecto, se están desarrollando nuevas técnicas para mejorar la interpretabilidad de los modelos, como son el uso de la información de las capas intermedias de las redes neuronales profundas, la agregación de métricas de interpretabilidad para medir la explicabilidad de los modelos, el desarrollo de modelos adversarios para cuantificar el grado de explicabilidad, la limitación de los parámetros a optimizar para aumentar su interpretabilidad, o el uso de técnicas de visualización para facilitar la comprensión de los resultados.

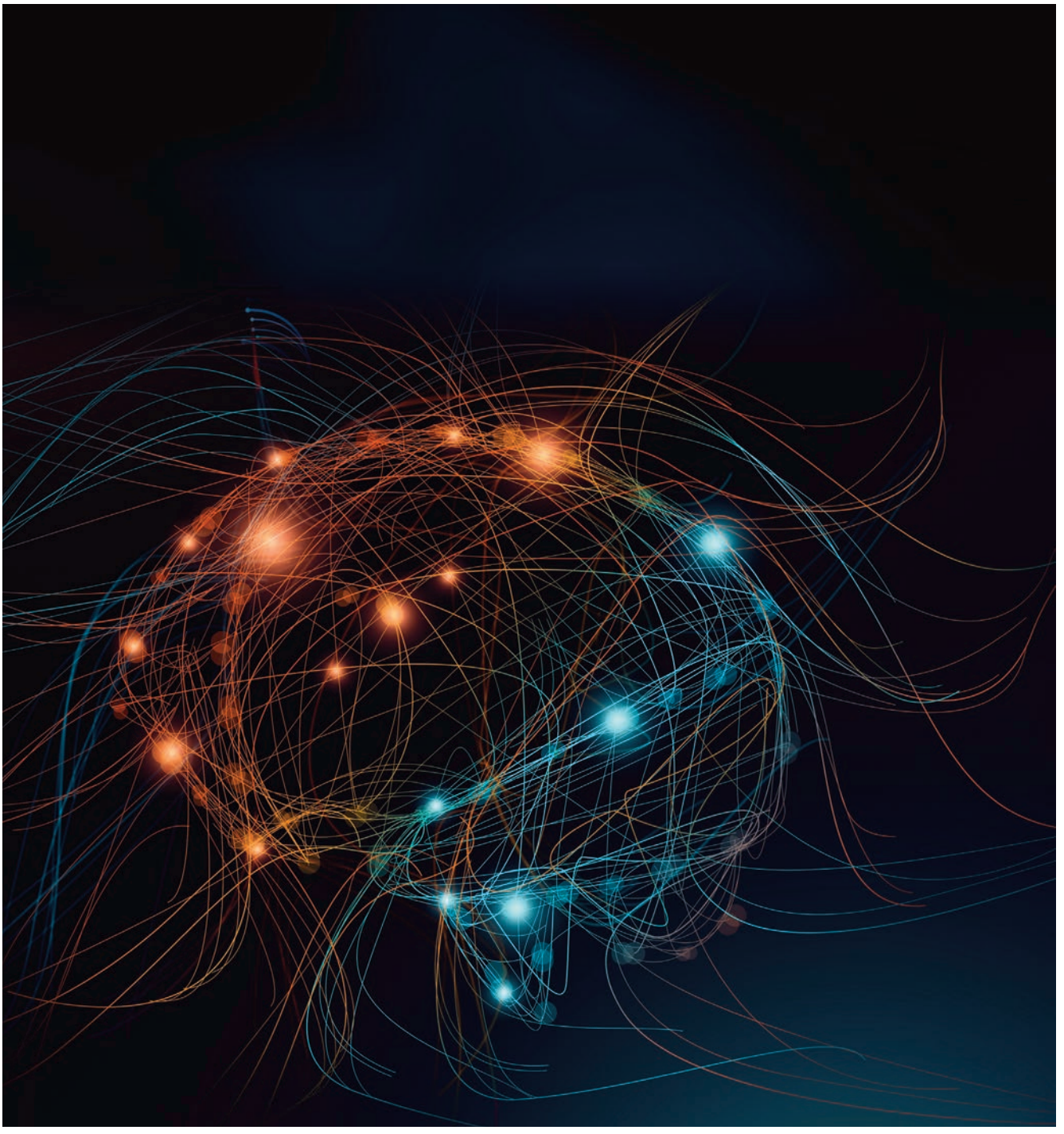
Figura 12. Retos comunes en la interpretabilidad de modelos de AI.

	Retos	Soluciones
Interpretabilidad del modelo	**El proceso de modelización debe ser reproducible,** pese a la varianza inherente de los modelos de AI.	Control de aleatoriedad en variables y algoritmos (e.g. uso de semillas).
	Las explicaciones de interpretabilidad **pueden no ser únicas, intuitivas,** explicar completamente la predicción, o ser coherentes para datos similares.	Comparación de diversas técnicas de interpretabilidad aplicadas sobre muestras representativas, verificando la consistencia en datos similares. Creación de reportes de interpretabilidad entendibles por áreas menos técnicas.
	La secuencia de predicciones más probables del modelo puede no ser coherente, y las predicciones pueden ser sensibles a **perturbaciones en datos de entrada**.	Aplicación de técnicas de interpretabilidad sobre la secuencia de predicciones más probables del modelo, e incluir análisis de sensibilidad específicos al revisarlo.
	No existe un **formalismo matemático para describir las propiedades de muchos modelos de ML** (e.g. en modelos de aprendizaje profundo).	Documentación completa sobre modelos, y su fundamento matemático, y reflejo del aumento del riesgo en el tiering.
Riesgo de modelo	No hay **medidas globalmente aceptadas de evaluación** para los modelos de ML, ni de interpretabilidad.	Aplicación de un marco de pruebas diverso y completo a lo largo de todo el ciclo de vida del modelo.
	Los usuarios del modelo deben **poder acceder a los módulos de interpretabilidad para entender las predicciones,** incluso después del desarrollo.	Implementación de módulos de XAI que den acceso a resultados de interpretabilidad a distintos usuarios.
Datos, documentación e implementación	Existe **varianza en la implementación en ciertos algoritmos de AI.**	Documentación adecuada y controles en la implementación.
	El nivel de **documentación debe reflejar la complejidad del algoritmo.**	Incluir en detalle los resultados de los test de interpretabilidad, garantizando su réplica por un tercero.

Caso práctico de interpretabilidad

"Los necios ignoran la complejidad. Los pragmáticos la sufren. Algunos pueden evitarla. Los genios la eliminan".

Alan Perlis[72]



Planteamiento

En esta sección se presenta un caso práctico de interpretabilidad en inteligencia artificial con el objetivo de ilustrar cómo se aplican las técnicas de XAI descritas en la sección anterior.

El caso de estudio seleccionado aborda el problema de la retención de empleados en una organización, centrándose en comprender y explicar las causas que llevan a los empleados a abandonar su puesto de trabajo. La identificación de estos factores puede permitir a las organizaciones tomar medidas preventivas y desarrollar estrategias para mejorar la satisfacción laboral y la retención del talento.

En este caso práctico, se utilizará un conjunto de datos ficticios generados por IBM y publicados en Kaggle[73]. Este conjunto de datos contiene información sobre los empleados de una organización, incluyendo características demográficas, datos sobre su puesto de trabajo, y si han abandonado la empresa o no.

En el ejercicio que se plantea, la compañía presenta un nivel de abandono de empleados del 16%, un 6% por encima del promedio histórico, y está preocupada por conocer las causas para elaborar un plan de remediación.

Las principales variables presentes en el conjunto de datos incluyen:

- Nivel de educación (desde "secundaria" hasta "doctorado").
- Satisfacción con el ambiente laboral (desde "bajo" hasta "muy alto").
- Involucración en el trabajo (desde "bajo" hasta "muy alto").
- Satisfacción con el trabajo (desde "baja" hasta "muy alta").
- Calificación del rendimiento (desde "bajo" hasta "sobresaliente").
- Satisfacción con las relaciones laborales (desde "baja" hasta "muy alta").
- Equilibrio entre la vida personal y profesional (desde "malo" hasta "óptimo").
- Años desde la última promoción en el trabajo (variable numérica).
- Salario mensual (variable numérica).
- Años en el puesto de trabajo actual (variable numérica).
- Distancia al puesto de trabajo (variable numérica).
- Número de empresas en las que se ha trabajado (variable numérica).
- Rol en el puesto de trabajo actual (variable categórica, incluye *Manager*, *Director*, *Research Scientist*, entre otros).

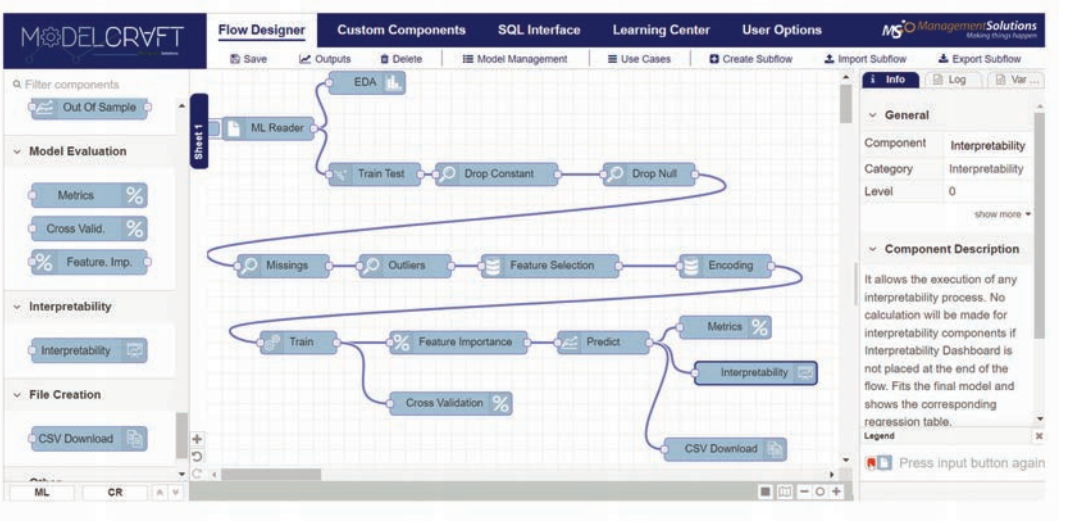
El enfoque del caso práctico será entrenar y validar diferentes modelos de inteligencia artificial para predecir el abandono de los empleados, utilizando las técnicas de XAI para analizar y comprender el comportamiento y las decisiones de los modelos seleccionados.

Para simplificar y agilizar el proceso, se ha empleado el sistema de modelización por componentes ModelCraft™, que contiene múltiples técnicas relevantes de AI y XAI. Este sistema permitirá realizar el estudio de forma eficiente y sin necesidad de escribir código.

[72] Alan Jay Perlis (1922-1990), informático estadounidense, doctor en Informática por el MIT y profesor de la Universidad de Purdue, la Universidad Carnegie Mellon y la Universidad de California en Berkeley, conocido por sus trabajos pioneros en lenguajes de programación y por ser el primer ganador del Turing Award.

[73] Kaggle (2017). IBM HR Analytics Employee Attrition & Performance.

Figura 13. Flujo de modelización en ModelCraft™.



A lo largo del desarrollo del caso práctico, se aplicarán las técnicas de interpretabilidad SHAP, LIME y PDP para analizar los modelos seleccionados y comprender qué variables influyen en la decisión de los empleados de abandonar su puesto de trabajo. Además, se explorará cómo estas variables interactúan entre ellas y cómo afectan a diferentes segmentos de la población de empleados.

Al finalizar el caso práctico, se evaluará la efectividad y las limitaciones de las técnicas de interpretabilidad utilizadas. También se discutirá cómo la combinación de modelos de inteligencia artificial y módulos de interpretabilidad puede mejorar la capacidad predictiva y la comprensión de los modelos, facilitando así la toma de decisiones basadas en datos en el ámbito empresarial.

Proceso de modelización

El proceso de modelización se realiza en tres fases: ingeniería de datos, modelización y análisis de interpretabilidad del modelo.

1. Ingeniería de datos

La ingeniería de datos es la fase inicial en la que se prepara y procesa el conjunto de datos para su uso en la creación de modelos de inteligencia artificial. En este caso, se realizan las siguientes acciones:

- Definición del ámbito de análisis: en este caso, se toma como población a todos los empleados que han sido baja en los últimos dos años.
- Limpieza de datos: se verifica la calidad de los datos y se eliminan o corrigen registros con información faltante o inconsistente.
- Transformación de variables: se convierten las variables categóricas en numéricas mediante técnicas como el *one-hot encoding* o el *ordinal encoding*. Además, se normalizan o estandarizan las variables numéricas cuando es necesario.
- Selección de variables: se identifican las variables más relevantes para predecir el abandono de los empleados utilizando técnicas de selección de variables como la correlación de Pearson, la importancia de las características en modelos basados en árboles o la eliminación recursiva de características.
- Construcción de variables: se generan nuevas variables a partir de las ya existentes para analizar si son mejores predictores del abandono de empleados, tales como la "satisfacción total", que se ha construido como suma de las puntuaciones de las variables "Satisfacción con el ambiente", "Satisfacción con el trabajo", "Calificación del rendimiento", "Equilibrio entre la vida personal y la laboral", "Involucración en el trabajo" y "Satisfacción con las relaciones laborales".
- División del conjunto de datos: se divide el conjunto de datos en dos subconjuntos: entrenamiento y prueba. El subconjunto de entrenamiento se utiliza para ajustar y optimizar los modelos de inteligencia artificial, mientras que el subconjunto de prueba se emplea para evaluar el rendimiento y la capacidad predictiva de los modelos.

2. Desarrollo del modelo

En esta fase, se entrenan y validan diferentes modelos de inteligencia artificial utilizando el subconjunto de entrenamiento. En concreto, se ajustan y comparan varios de los algoritmos de aprendizaje automático más comunes, como regresión logística, árboles de decisión, máquinas de vectores de soporte, redes neuronales y *random forest,* para seleccionar el modelo con el mejor rendimiento.

Para evitar el sobreentramiento y optimizar los hiperparámetros de los modelos, se emplean técnicas de validación cruzada y búsqueda en cuadrícula o aleatoria. Asimismo, se ha tenido especialmente en cuenta la complejidad del modelo durante el entrenamiento a la hora de seleccionar un algoritmo determinado, para facilitar su interpretación.

Para ello, se ha generado un flujo de desarrollo de modelos en ModelCraft™ (Fig. 13).

Para seleccionar el modelo con la mejor capacidad predictiva, se evalúa su rendimiento en el subconjunto de prueba utilizando métricas como la precisión, la sensibilidad, la especificidad y el área bajo la curva ROC (AUC-ROC). Estas métricas permiten evaluar la efectividad del modelo seleccionado en términos de su capacidad para predecir correctamente el abandono de empleados en datos no vistos previamente.

Considerado todo lo anterior, el algoritmo random forest arroja resultados de rendimiento superiores, aunque plantea un desafío de interpretabilidad a la hora de comprender sus predicciones. Este modelo ha considerado 300 árboles de decisión y ha arrojado una precisión del 75% y una sensibilidad del 84%. Por tanto, se trata de predicciones muy fiables y en pocas ocasiones se obtienen falsos negativos. Esto es relevante para este caso de estudio, en el que la compañía previsiblemente querría reducir lo máximo posible este tipo de error.

3. Análisis de interpretabilidad

En esta última fase, se aplican técnicas de interpretabilidad para analizar y comprender el comportamiento y las decisiones del modelo seleccionado. En concreto, los objetivos del análisis son:

- Entender qué variables son más importantes en la toma de decisiones para la compañía a nivel global, para lo que se ha empleado la comparación por importancia de cada variable.
- Entender cómo impactan cambios en las variables más importantes para distintos rangos de población.
- Entender los resultados del modelo en casos particulares donde se observa una determinada probabilidad de abandono.

En este caso práctico, se utilizan las técnicas de SHAP, LIME y PDP para explicar cómo el modelo toma decisiones y cómo las variables de entrada influyen en las predicciones.

SHAP permite obtener resultados de interpretabilidad global, que dan una interpretación de la importancia de cada variable, y LIME permite realizar un análisis intuitivo de interpretabilidad local que permita explicar el resultado del modelo para cada empleado partiendo de modelos lineales más sencillos. Como complemento, los gráficos PDP permiten visualizar cómo cambios en cada variable impactan en la predicción del modelo.

Con ello, se ha obtenido la siguiente distribución de la importancia de cada variable (Fig. 14).

En este caso, se observa que la variable con más importancia en la predicción de abandono (15,65%) es la "satisfacción total", un indicador sintético definido como una media ponderada de seis elementos (ambiente de trabajo, adecuación de funciones y áreas al puesto, rating interno, conciliación familiar, relación con compañeros y supervisores, y cargo y responsabilidad del empleado).

Este resultado es intuitivo y demuestra que la variable "satisfacción total" está bien diseñada. Sin embargo, las siguientes tres variables por importancia (antigüedad en la empresa, salario y distancia desde casa al puesto de trabajo) han demostrado tener una elevada influencia en el abandono de los empleados, que colectivamente duplica la del indicador "satisfacción total".

Figura 14. Interpretabilidad global del modelo random forest mediante SHAP, donde se utilizan los valores de Shapley para la obtención de la importancia de las variables.

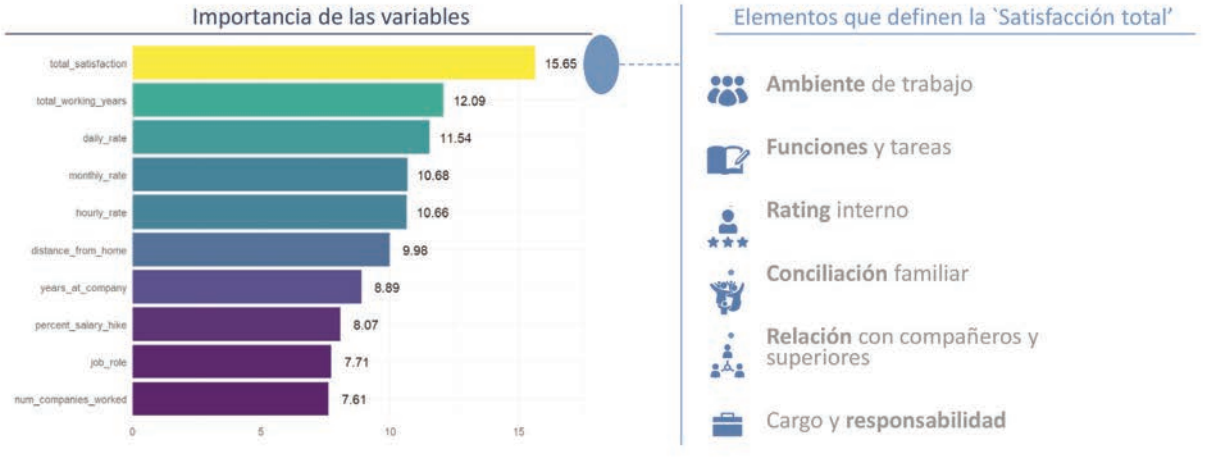
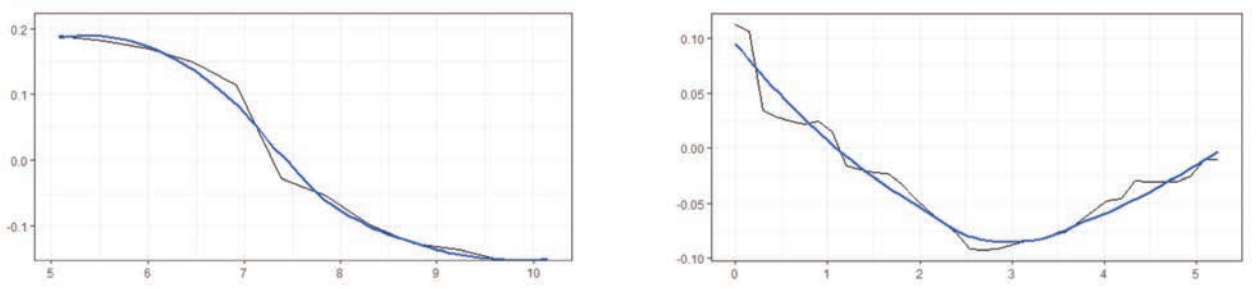


Figura 15. Gráficos PDP para las variables "satisfacción total" y "antigüedad en la empresa".



Para entender cómo influye cada variable individualmente, se han estudiado los PDP (fig. 15).

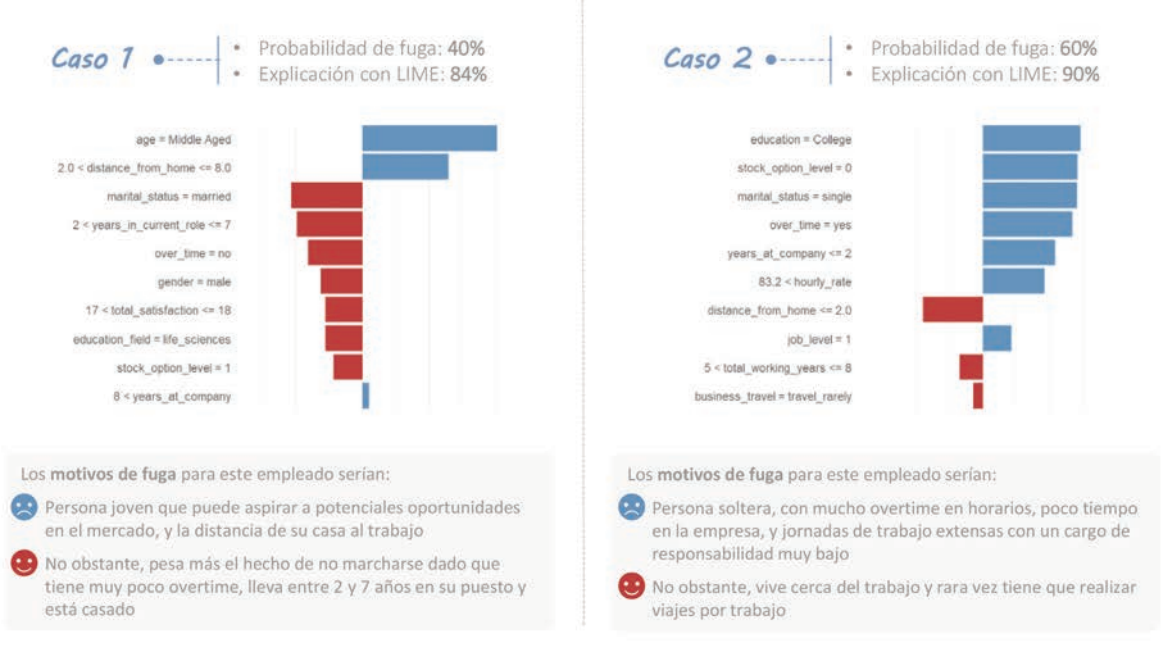
En la antigüedad en la empresa, se observa que a los tres años la tendencia se invierte: los empleados de antigüedad intermedia son, en promedio, los menos propensos a abandonar la compañía. Para la satisfacción total se observa una tendencia intuitiva: una mayor satisfacción reportada en las encuestas internas resulta en una menor tasa de abandono.

Para complementar el análisis anterior, se ha empleado LIME para analizar caso a caso los valores de las variables que influyen en la probabilidad de abandono de determinados empleados. La fig. 16 muestra dos empleados con distintas probabilidades de abandono obtenidas mediante el modelo. LIME muestra una métrica de explicabilidad que representa cómo de bueno es el ajuste lineal que ha obtenido mediante el modelo subrogado local para explicar estas predicciones.

Es destacable que las causas de abandono más relevantes en estos dos casos no corresponden necesariamente a las variables más influyentes a nivel global. Si bien se puede observar que la satisfacción total contribuye a explicar la probabilidad de abandono del empleado en el caso 1, no parece tener un impacto significativo en el caso 2, donde la probabilidad de abandono es mayor.

Esto refleja las dificultades existentes a la hora de interpretar este modelo, generalizable a modelos similares: pese a que la satisfacción total puede explicar de manera notable la probabilidad de abandono en promedio, esta conclusión es una generalización; se dan casos individuales y de colectivos en los que el abandono se explica en mayor medida por otras variables.

Figura 16. Interpretabilidad local del modelo random forest mediante LIME



Conclusiones del caso práctico

Del caso práctico de interpretabilidad en inteligencia artificial presentado se pueden extraen diversas conclusiones y lecciones aprendidas que pueden ser de utilidad en futuras aplicaciones de modelos de AI y XAI:

- **Aplicación del modelo:** la correcta aplicación e interpretación del modelo en este caso puede permitir anticipar y prevenir el abandono de empleados. Entre los usos que se pueden dar al modelo, destaca la capacidad de crear distintos perfiles con propensión al abandono e identificar las características de estos empleados con antelación para tomar las medidas adecuadas, lo que a largo plazo puede contribuir a reducir el nivel de rotación de la empresa.

- **Elección del modelo:** el proceso de modelización ha demostrado la importancia de comparar y validar diferentes algoritmos de aprendizaje automático para seleccionar el modelo con la mejor capacidad predictiva. En este caso, el modelo de *random forest* resultó ser el más adecuado para predecir el abandono de empleados.

- **Importancia de la interpretabilidad:** la aplicación de técnicas de interpretabilidad, como SHAP, LIME y PDP, ha proporcionado una comprensión más profunda de cómo el modelo toma decisiones y cómo las variables de entrada influyen en las predicciones. Esta información es crucial para validar la aplicabilidad del modelo en el contexto real y para garantizar que las predicciones se basen en características relevantes y significativas.

- **Variables influyentes:** el análisis de interpretabilidad ha permitido identificar las variables más relevantes para predecir el abandono de empleados. Estas variables pueden ser útiles para desarrollar estrategias de retención y mejorar la satisfacción laboral. Además, la comprensión de cómo estas variables interactúan entre ellas y cómo afectan a diferentes segmentos de la población de empleados puede enriquecer el análisis y facilitar la toma de decisiones basada en datos.

- **Implementación práctica:** el caso práctico demuestra la viabilidad y utilidad de aplicar técnicas de AI y XAI en un escenario realista, utilizando datos ficticios pero representativos de una situación empresarial. Este enfoque puede adaptarse a otros contextos y problemas empresariales, aprovechando las ventajas de la inteligencia artificial y la interpretabilidad para mejorar la toma de decisiones y obtener resultados más eficientes y efectivos.

- **Limitaciones:** al mismo tiempo, este caso de uso ha puesto de manifiesto las limitaciones y dificultades en la aplicación de las técnicas de interpretabilidad *post-hoc*. Es importante reconocer que los métodos de interpretabilidad no son infalibles y que, en ocasiones, pueden presentar resultados aproximados o parciales. Por lo tanto, es fundamental aplicar un enfoque crítico y riguroso al interpretar y validar los resultados de las técnicas de interpretabilidad.

- **Combinación de modelos de AI y módulos de interpretabilidad:** este caso práctico muestra cómo la integración de modelos de AI y módulos de interpretabilidad puede mejorar la capacidad predictiva y la comprensión de los modelos. Esto facilita la adopción de soluciones basadas en AI en la toma de decisiones empresariales.

- **Continuidad en el análisis de interpretabilidad:** por último, cabe destacar que el análisis de interpretabilidad no debe ser un ejercicio aislado aplicado durante el desarrollo de los modelos, sino que debe realizarse de manera continuada, reproducible y fiable a lo largo de toda la vida del modelo.

En conclusión, este caso práctico de interpretabilidad en inteligencia artificial ha proporcionado una experiencia valiosa en la aplicación de técnicas de AI y XAI en un contexto empresarial, y ha mostrado el potencial de la AI y la interpretabilidad para mejorar la toma de decisiones, al tiempo que ha revelado las limitaciones y dificultades asociadas con estas técnicas y la necesidad de un enfoque crítico y riguroso al interpretar y validar los resultados de la AI.

Conclusiones

"Con la programación adecuada, un ordenador puede convertirse en un teatro, un instrumento musical, un libro de consulta, un contrincante de ajedrez. Ninguna otra entidad en el mundo salvo el ser humano tiene una naturaleza tan adaptable y universal".

Daniel Hillis[74]

En este estudio se ha presentado la inteligencia artificial explicable (XAI), sus fundamentos, contexto y técnicas para mejorar la interpretabilidad de los modelos. Se han discutido los principales desafíos que enfrentan los modelos de inteligencia artificial en términos de interpretabilidad y cómo la tecnología puede ayudar a abordarlos, incluyendo un caso práctico desarrollado con ModelCraft™ para demostrar cómo se pueden emplear estas técnicas para entender y explicar los modelos de AI.

La disciplina de la AI, y dentro de ella la XAI, ha crecido en importancia a nivel mundial en los últimos años, cuando el desarrollo de tecnologías de AI de alto rendimiento se ha convertido en una prioridad para muchos sectores, desde la salud hasta la seguridad, pasando por los servicios financieros o la energía, entre otros. La interpretabilidad surge como una necesidad para entender y mejorar los modelos de AI, lo que reviste especial complejidad en el caso de determinadas técnicas.

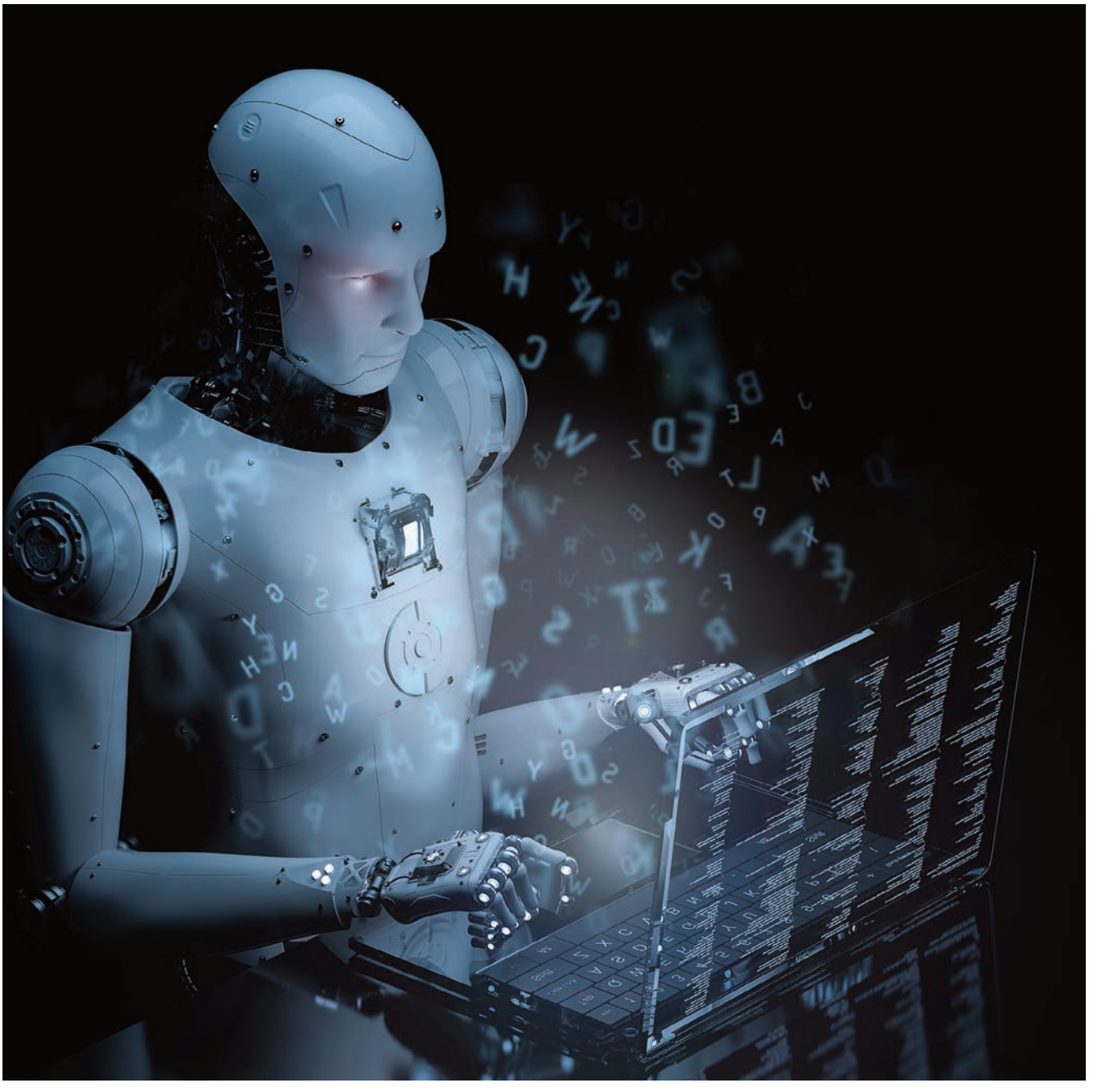
Como se ha visto, los modelos de AI pueden enfrentar dificultades para explicar los resultados o la lógica detrás de sus decisiones. Esto se debe a que estos modelos utilizan técnicas de aprendizaje profundo y algoritmos complejos para aprender a partir de datos, los cuales a menudo son difíciles de interpretar, y esto plantea desafíos a la hora de evaluar los modelos de AI y la confiabilidad de sus resultados.

Por todo ello, el marco regulatorio sobre AI está evolucionando con rapidez, y se espera que las organizaciones se adapten a los nuevos requerimientos de transparencia, explicabilidad e imparcialidad en el uso de modelos de AI. Esto implica la necesidad de un enfoque integral que integre la interpretabilidad y explicabilidad en la organización y los procesos de cada compañía, abarcando técnicas de interpretabilidad, gestión del riesgo de modelo, colaboración interdisciplinaria y capacitación en XAI para los profesionales involucrados en el desarrollo y la aplicación de la AI, entre otros.

En conclusión, la interpretabilidad de los modelos de inteligencia artificial es un área de investigación emergente, y es esperable que continúe desarrollándose y creciendo en importancia a medida que los modelos de AI se vuelvan más complejos, la regulación siga proliferando, y su uso se extienda a más ámbitos de alta sensibilidad.

[74] Daniel Hillis (n. 1956), inventor, empresario y científico estadounidense, pionero de la computación paralela y su uso en el campo de la inteligencia artificial, con más de 300 patentes publicadas.

Glosario



Aprendizaje automático (machine learning): subcampo de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las máquinas aprender y mejorar su rendimiento en una tarea específica a través de la experiencia.

Caja blanca (white box): sistema o modelo de AI cuyo funcionamiento interno es sencillo de entender y explicar.

Caja negra (black box): sistema o modelo de AI cuyo funcionamiento interno es desconocido o difícil de entender.

Derecho a una explicación: concepto legal que sostiene que los individuos tienen derecho a saber cómo se toman las decisiones automatizadas que les afectan y a recibir una explicación comprensible de cómo funcionan los algoritmos involucrados.

Explicabilidad: capacidad de un sistema de AI para proporcionar justificaciones claras y comprensibles de sus predicciones o decisiones a los usuarios y partes interesadas. Implica ofrecer información detallada y contextualizada sobre cómo y por qué un modelo de AI llega a una conclusión particular, facilitando la confianza y la adopción de la tecnología.

GPT-4: cuarta generación del modelo Generative Pre-trained Transformer, desarrollado por la OpenAI Foundation, que se utiliza para tareas de procesamiento del lenguaje natural y generación de texto.

Inteligencia artificial (AI): campo de estudio que busca desarrollar sistemas capaces de realizar tareas que normalmente requieren inteligencia humana, como el aprendizaje, el razonamiento, la percepción y la toma de decisiones.

Inteligencia artificial explicable (XAI): enfoque de AI que busca hacer que los modelos de inteligencia artificial sean más comprensibles y transparentes para los humanos.

Interpretabilidad: facilidad con la que los humanos pueden comprender el proceso de toma de decisiones de un modelo de AI, así como las relaciones entre las características de entrada y las predicciones o decisiones. Un modelo interpretable permite a los usuarios discernir cómo se llega a una predicción o decisión específica.

LIME (Local Interpretable Model-agnostic Explanations): técnica de explicabilidad que ayuda a entender las predicciones individuales de un modelo de AI mediante la creación de aproximaciones locales interpretables.

Modelo subrogado: modelo interpretable que se entrena para imitar las predicciones de un modelo de AI complejo y menos interpretable, como una red neuronal profunda. El objetivo de un modelo subrogado es proporcionar una explicación simplificada y comprensible de cómo el modelo original toma decisiones.

Open AI Foundation: organización de investigación y desarrollo de la inteligencia artificial, actualmente participada por Microsoft, cuyo objetivo declarado es garantizar que la AI beneficie a toda la humanidad.

Partial Dependence Plot (PDP): técnica de visualización que muestra el efecto promedio de una característica en las predicciones de un modelo de AI, manteniendo constantes todas las demás características. Ayuda a comprender la relación entre las características y las predicciones, y a detectar posibles interacciones y no linealidades.

Prueba de esquemas de Winograd: prueba de comprensión del lenguaje natural que evalúa la capacidad de una IA para resolver ambigüedades en el lenguaje a través del uso de conocimiento y razonamiento común.

Reglamento General de Protección de Datos (GDPR): legislación de la Unión Europea que establece reglas para la recopilación, el almacenamiento y el procesamiento de datos personales de los ciudadanos de la UE.

Sesgos en AI: prejuicios sistemáticos presentes en los datos de entrenamiento o en el diseño de un algoritmo de AI que pueden llevar a decisiones injustas o discriminatorias.

SHAP (SHapley Additive exPlanations): técnica de explicabilidad que utiliza valores de Shapley, provenientes de la teoría de juegos cooperativos, para atribuir la importancia de cada variable en la predicción de un modelo de AI.

Sparsity: propiedad de un modelo por la que este solo considera el subconjunto de variables que son realmente relevantes para la estimación.

Test de Turing: prueba propuesta por Alan Turing en 1950 que evalúa la capacidad de una máquina de imitar la inteligencia humana al punto de ser indistinguible de un humano en una conversación.

Transformer: arquitectura de red neuronal introducida por Google Brain en 2017 que se utiliza principalmente en tareas de procesamiento del lenguaje natural (NLP). Los transformers son conocidos por su capacidad para manejar secuencias largas de datos y por su eficiencia en el entrenamiento. Se basan en mecanismos de atención, que permiten a la red ponderar la importancia relativa de las palabras o elementos en una secuencia a lo largo del tiempo. Los transformers han impulsado el desarrollo de modelos de lenguaje de vanguardia, como GPT y BERT, y han revolucionado el campo de NLP.

Transparencia: apertura y accesibilidad de un sistema de AI en términos de su diseño, estructura y procesos internos. Un sistema transparente permite a los usuarios y partes interesadas examinar y comprender sus componentes, algoritmos y decisiones.

Red neuronal profunda: tipo de algoritmo de aprendizaje automático que consta de múltiples capas de neuronas artificiales y es capaz de aprender representaciones jerárquicas de datos.

Referencias

Broniatowski, D. (2021). Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. https://www.nist.gov/publications/psychological-foundations-explainability-and-interpretability-artificial-intelligence

Comisión Europea (2021). Artificial Intelligence Act / Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión. https://artificialintelligenceact.eu/

Comisión Europea (2019). Dirección General de Redes de Comunicación, Contenido y Tecnologías, Directrices éticas para una IA fiable, Oficina de Publicaciones, 2019, https://data.europa.eu/doi/10.2759/14078

C. Rudin, C. Chen, Zhi Chen, H. Huang, L. Semenova, C. Zhong. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. http://essay.utwente.nl/91965/

Doshi-Velez, F., et al. (2017). Towards a rigorous science of interpretable machine learning. https://arxiv.org/abs/1702.08608

Devis (2011). https://cs.nyu.edu/~davise/papers/WinogradSchemas/WSCollection.html

Dimensions (2022). https://app.dimensions.ai/discover/publication

EBA (2021). Discussion paper on machine learning for IRB models. https://www.eba.europa.eu/regulation-and-policy/model-validation/discussion-paper-machine-learning-irb-models

European Parliamentary Research Service (2020). The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)641530

Floridi et al. (2022). capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091

Friedman, J. H. (2001). "Greedy function approximation: A gradient boosting machine." Annals of statistics (2001): 1189-1232. https://www.jstor.org/stable/2699986

Gall, R. (2018). Machine Learning explainability vs interpretability: two concepts that could restore trust in AI, KDnuggets. https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html

GDPR (2018), Recital 71. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Goldstein, A.; Kapelner, A.; Bleich, J; Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. https://arxiv.org/abs/1309.6392

Harnad, D. (2003). Can a machine be conscious? How? https://web-archive.southampton.ac.uk/cogprints.org/5330/

IBM (2022). Explainable AI (XAI). https://www.ibm.com/watson/explainable-ai

iDanae (2022). ML Applied to Credit Risk: building explainable models. Quarterly Newsletter 3Q22. iDanae Chair. https://blogs.upm.es/catedra-idanae/wp-content/uploads/sites/698/2022/10/Idanae-3Q22.pdf

Jonathon Phillips, P.; Hahn, H.; Fontana, P; Yates, A.; Greene, K. K.; Broniatowski, D. A.; Przybocki, M. A. (2021). Four Principles of Explainable Artificial Intelligence. NIST. https://www.nist.gov/publications/four-principles-explainable-artificial-intelligence

Kaggle (2017). IBM HR Analytics Employee Attrition & Performance. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

LeCun, Y.; Bengio, Y.; Hinton, G. (2015). Deep learning. Nature. https://pubmed.ncbi.nlm.nih.gov/26017442/

Leventi-Peetz, A.-M., et al. (2022). Deep Learning Reproducibility and Explainable AI (XAI). https://arxiv.org/abs/2202.11452

Levesque, H. (2014). On our best behaviour. Written version of the Research Excellence Lecture presented in Beijing at the IJCAI-13 conference. Artificial Intelligence, vol. 212, pages 27-35. https://doi.org/10.1016/j.artint.2014.03.007

Lundberg, S. M.; Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. https://dl.acm.org/doi/10.5555/3295222.3295230

Management Solutions (2023). ModelCraft. Modelización por componentes. https://www.managementsolutions.com/es/microsites/soluciones-propietarias/modelcraft

Management Solutions (2022). Gamma. Sistema de gobierno de modelos. https://www.managementsolutions.com/es/microsites/soluciones-propietarias/gamma

Management Solutions (2021). Nota técnica sobre el EBA Discussion paper on machine learning for IRB models. https://www.managementsolutions.com/es/publicaciones-y-eventos/apuntes-normativos/notas-tecnicas-normativas/documento-de-debate-sobre-machine-learning-en-el-enfoque-irb

Management Solutions (2020). Auto machine learning, towards model automation. https://www.managementsolutions.com/en/publications-and-events/industry-reports/white-papers/auto-machine-learning-towards-model-automation

Management Solutions (2018). Machine learning, a key component in business model transformation. https://www.managementsolutions.com/en/publications-and-events/industry-reports/white-papers/machine-learning-a-key-component-in-business-model-transformation

Management Solutions (2015). Data science and the transformation of the financial industry. https://www.managementsolutions.com/en/publications-and-events/industry-reports/white-papers/data-science

Marcinkevics, R. (2020). Interpretability and Explainability: A Machine Learning Zoo Mini-tour. ETH Zürich, Department of Computer Science, Institute for Machine Learning. https://arxiv.org/abs/2012.01805

McCarthy, J. (2004). What is artificial intelligence? Stanford University. http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell.2019,267, 1–38. https://www.sciencedirect.com/science/article/pii/S0004370218305988

OECD (2019). Principles for Artificial Intelligence. https://www.oecd.org/digital/artificial-intelligence/

Oneto, L., Chiappa, S., (2020). Fairness in Machine Learning. 2012.15816.pdf (arxiv.org)

Ribeiro, M. T.; Singh, S.; Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. https://arxiv.org/abs/1602.04938

Ribeiro, M. T.; Singh, S.; Guestrin, C. (2018). "Anchors: high-precision model-agnostic explanations". AAAI Conference on Artificial Intelligence (AAAI). https://ojs.aaai.org/index.php/AAAI/article/view/11491

Roscher, R.; Bohn, B.; Duarte, M.; Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. https://ieeexplore.ieee.org/document/9007737

Shapley, L. (1953). A Value for n-Person Games. In: Kuhn, H. and Tucker, A., Eds., Contributions to the Theory of Games II, Princeton University Press, Princeton, 307-317. https://doi.org/10.1515/9781400881970-018

Sudjianto, A.; Knauth, W.; Singh, R.; Yang, Z.; Zhang, A. (2011). Unwrapping The Black Box of Deep ReLU Networks: Interpretability, Diagnostics, and Simplification. Cornell University. https://arxiv.org/abs/2011.04041

Sudjianto, A.; Zhang, A. (2021). Designing Inherently Interpretable Machine Learning Models. https://arxiv.org/abs/2111.01743

Turing, A. (1950). Computing Machinery and Intelligence. Mind 49: 433-460.

Vilone G., Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion, vol. 76: 89-106. https://www.sciencedirect.com/science/article/pii/S1566253521001093

White House OSTP (2022). Blueprint for an AI Bill of Rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Yang, Z.; Zhang, A.; Sudjianto, A. (2019). Enhancing Explainability of Neural Networks through Architecture Constraints. https://arxiv.org/abs/1901.03838

Zhou, N.; Zhang, Z.; Nair, V. N.; Singhal, H.; Chen, J.; Sudjianto; A. (2021). Bias, Fairness, and Accountability with AI and ML Algorithms. https://arxiv.org/abs/2105.06558

Nota: esta publicación se ha elaborado con la ayuda de varias herramientas de inteligencia artificial (AI). Estas herramientas se utilizaron para diversas tareas, como la búsqueda de información, la recopilación y organización de datos y la generación de resúmenes. En todo caso, el material final de esta publicación fue redactado por una persona y no por una AI.

Nuestro objetivo es superar las expectativas de nuestros clientes convirtiéndonos en socios de confianza

Management Solutions es una firma internacional de servicios de consultoría centrada en el asesoramiento de negocio, finanzas, riesgos, organización y procesos, tanto en sus componentes funcionales como en la implantación de sus tecnologías relacionadas.

Con un equipo multidisciplinar (funcionales, matemáticos, técnicos, etc.) de más de 3.300 profesionales, Management Solutions desarrolla su actividad a través de 44 oficinas (19 en Europa, 21 en América, 2 en Asia, 1 en África y 1 en Oceanía).

Para dar cobertura a las necesidades de sus clientes, Management Solutions tiene estructuradas sus prácticas por industrias (Entidades Financieras, Energía, Telecomunicaciones y Otras industrias) y por líneas de actividad que agrupan una amplia gama de competencias: Estrategia, Gestión Comercial y Marketing, Gestión y Control de Riesgos, Información de Gestión y Financiera, Transformación: Organización y Procesos, y Nuevas Tecnologías.

El área de I+D da servicio a los profesionales de Management Solutions y a sus clientes en aspectos cuantitativos necesarios para acometer los proyectos con rigor y excelencia, a través de la aplicación de las mejores prácticas y de la prospección continua de las últimas tendencias en metodologías de medición en el ámbito de la sostenibilidad (ambiental y social).

Javier Calvo Martín
Socio de Management Solutions
javier.calvo.martin@managementsolutions.com

Manuel Ángel Guzmán Caba
Socio de Management Solutions
manuel.guzman@managementsolutions.com

Segismundo Jiménez Láinez
Gerente de Management Solutions
segismundo.jimenez@msspain.com

Luz Ferrero Peña
Supervisora de Management Solutions
luz.ferrero@msgermany.com.de

Management Solutions, servicios profesionales de consultoría

Management Solutions es una firma internacional de consultoría centrada en el asesoramiento de negocio, finanzas, riesgos, organización, tecnología y procesos,

Para más información visita **www.managementsolutions.com**

Síguenos en:

www.managementsolutions.com

Madrid Barcelona Bilbao Coruña Málaga London Frankfurt Düsseldorf Paris Amsterdam Copenhagen Oslo Warszawa Wroclaw Zürich Milano Roma Bologna Lisboa Beijing Istanbul Johannesburgo Sydney Toronto New York New Jersey Boston Pittsburgh Atlanta Birmingham Houston San Juan de Puerto Rico San José Ciudad de México Monterrey Querétaro Medellín Bogotá Quito São Paulo Río de Janeiro Lima Santiago de Chile Buenos Aires