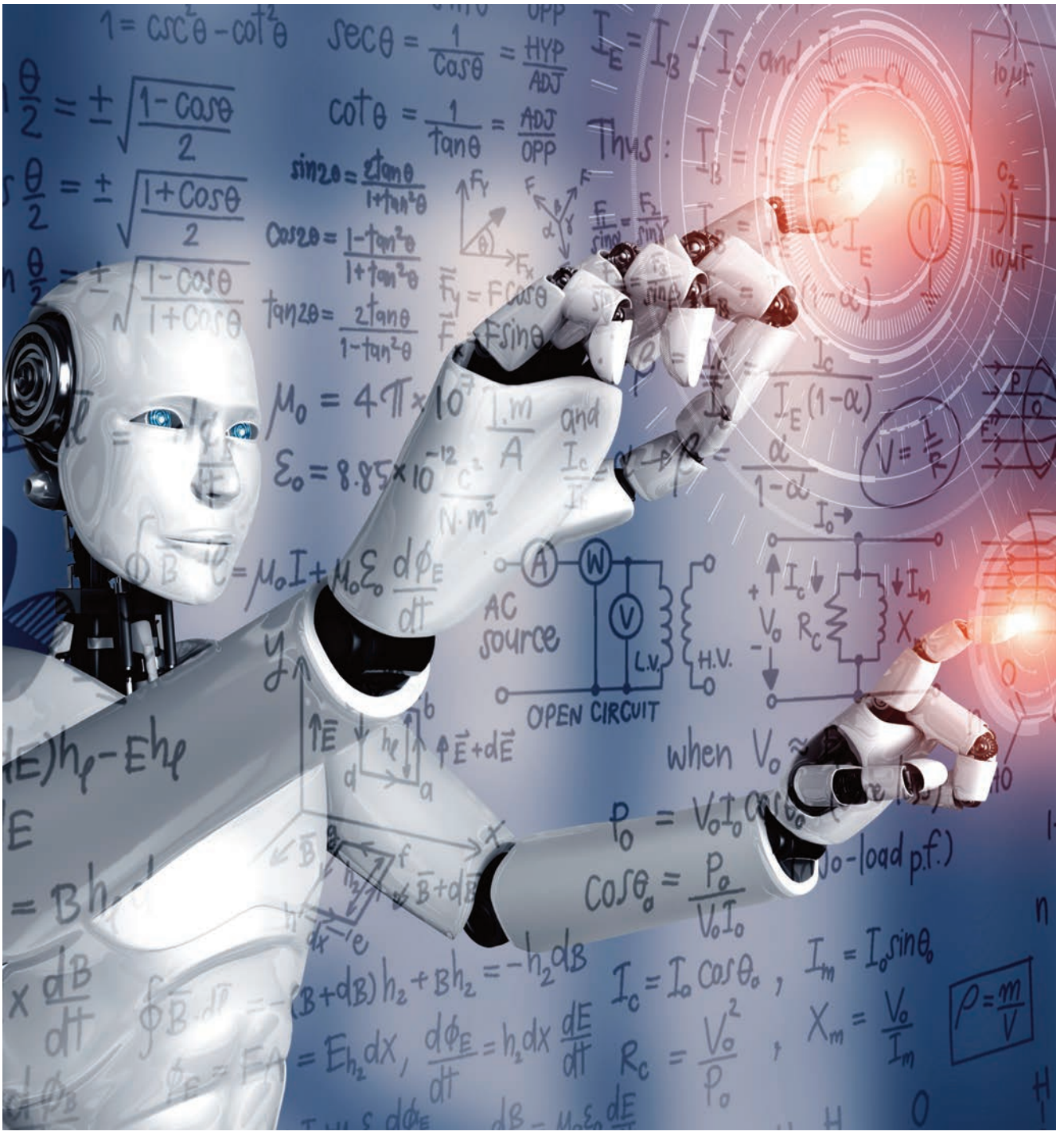
Resumen ejecutivo

"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).