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:

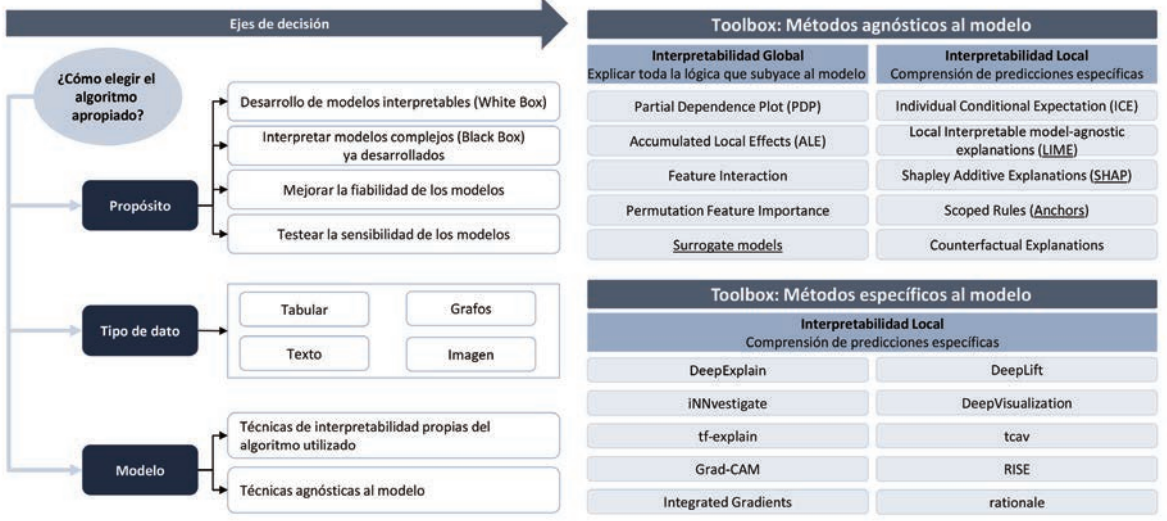
- La reproducibilidad de los resultados, el proceso de entrenamiento e implementación del modelo[47], la consistencia en sus predicciones y la explicación de la secuencia de predicciones más probables.
- Potenciales sesgos[48] en los datos de entrada.
- Imparcialidad *(fairness)*[49].
- Exactitud de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

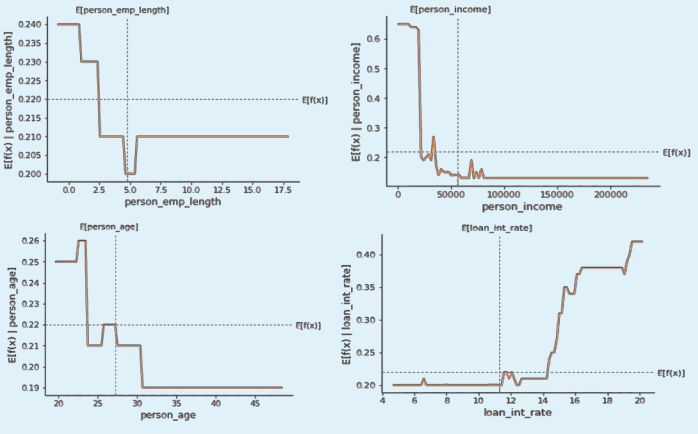
Los PDP se pueden aplicar a un caso de uso muy común en el sector bancario: la puntuación de clientes durante el proceso de concesión de préstamos para determinar su probabilidad de impago. En este ejemplo, se ha empleado una cartera anonimizada de préstamos hipotecarios con información de su actividad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

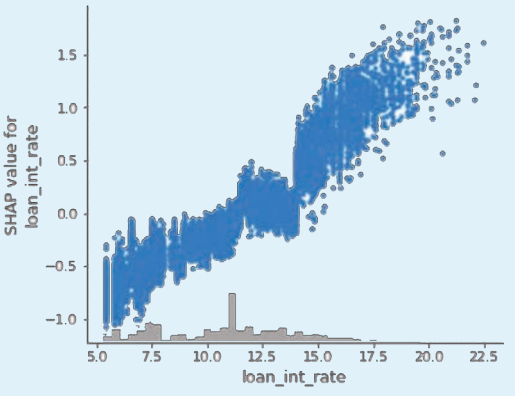
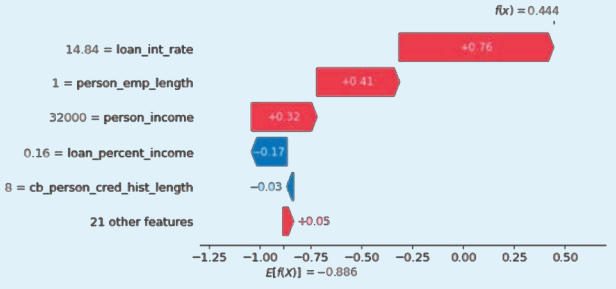
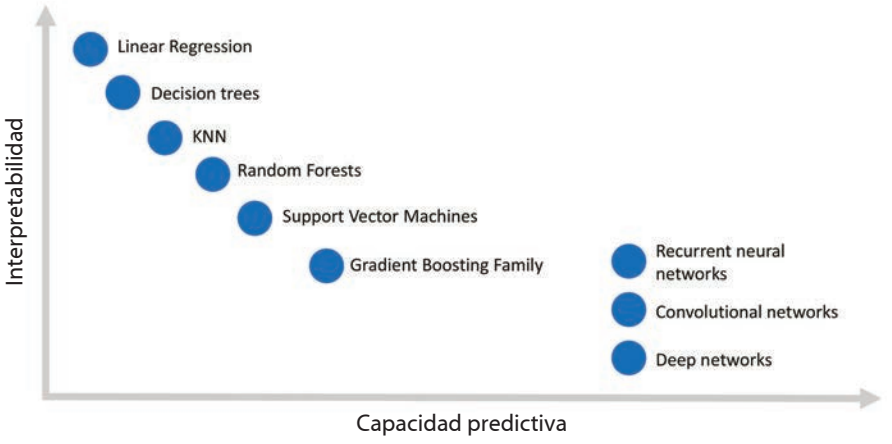


Figura 9. Valores de Shapley que influyen en la predicción de un cliente con préstamo denegado[2].



[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.

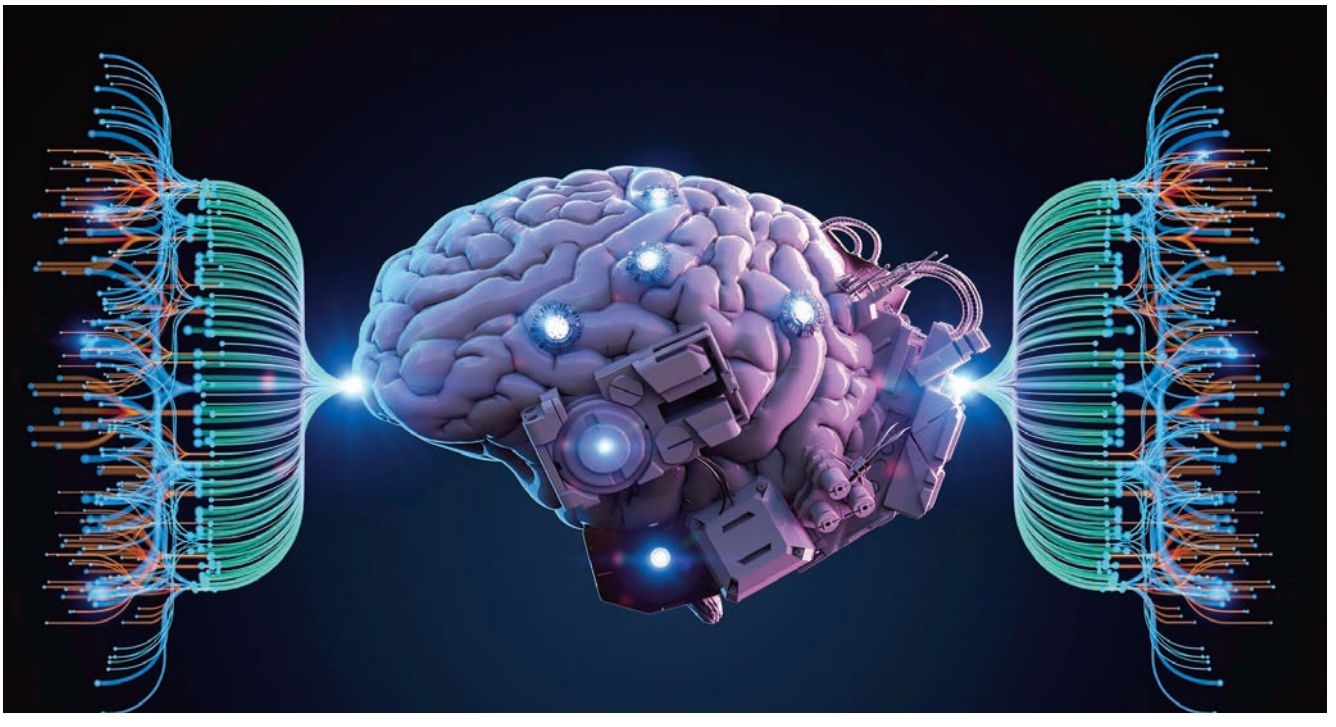
2. Otros factores de impacto

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.