CLAVES PARA UNA INTELIGENCIA ARTIFICIAL RESPONSABLE EN SALUD

Atributos como el de transparencia, explicabilidad, equidad y no discriminación deben gobernar la Inteligencia Artificial en todos los ámbitos, pero si hablamos de salud, sector sanitario e investigación clínica, entonces estos conceptos se hacen imprescindibles.

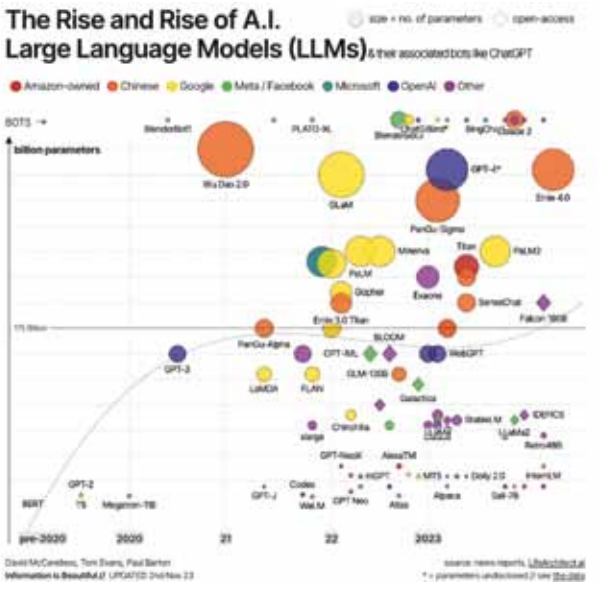
JULIO BONIS,
Director de Inteligencia Artificial & Big Data de Sermes CRO.

La inteligencia artificial está revolucionando el ámbito de la salud, por las grandes posibilidades que ofrece en la ayuda al diagnóstico, la planificación de tratamientos y la optimización de los procesos asistenciales. Sin embargo, el uso responsable de estas tecnologías requiere poner el foco en su transparencia, explicabilidad, equidad y no discriminación.

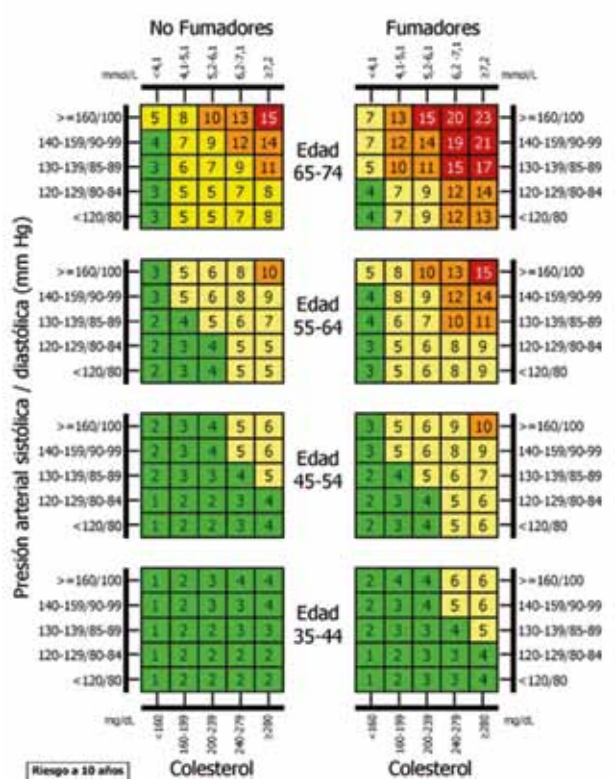
Poner en marcha un sistema de IA en salud implica trabajar con tres componentes principales: los datos de entrada, los modelos entrenados y el despliegue de los mismos en el mundo real. El proceso comienza con la recolección y el procesamiento de datos, que se utilizan para entrenar modelos predictivos o generativos mediante el uso de distintos algoritmos de aprendizaje automático. Una vez el modelo se despliega el trabajo no termina ahí puesto que se requiere una monitorización constante para gestionar riesgos y asegurar su correcto funcionamiento a lo largo del tiempo; y al aplicarse el modelo a poblaciones distintas a las que se utilizaron para entrenar los modelos originalmente.

El estudio Framingham: un ejemplo pionero

Un ejemplo pionero de un modelo predictivo en salud basado en el análisis de datos es el estudio de Framingham *(Framingham Heart Study)*. Iniciado en 1948, este estudio longitudinal recopiló sistemáticamente durante años datos de población sana en la ciudad de Framingham para identificar



qué factores estaban asociados y por tanto permitían de algún modo predecir el desarrollo de enfermedades cardiovasculares. En 1997, se publicó un modelo predictivo basado en estos datos, utilizando variables como edad, colesterol, presión arterial, diabetes y tabaquismo. Utilizando un tipo de algoritmo conocido como regresión logística se construyó un modelo predictivo que asignaba a cada una de esas variables un peso (o importancia) a la hora de predecir quién sufriría



años después un ictus o un infarto agudo de miocardio.

En cuanto al despliegue de este modelo en el mundo real, los investigadores del estudio de Framingham elaboraron unas "infografías" para que los profesionales sanitarios, que en aquellos tiempos aún no tenían acceso a un ordenador en la consulta, pudieran aplicar este modelo predictivo. Estas infografías impresas en un papel permitían de forma sencilla sabiendo la edad del paciente, si era fumador, diabético y sus niveles de colesterol, predecir la probabilidad de sufrir un evento cardiovascular en los próximos 10 años y de este modo ayudaban al clínico a tomar decisiones sobre si iniciar un tratamiento preventivo.

El modelo de Framingham tenía un grado de explicabilidad bastante alto pues simplemente mirando la infografía y los pocos parámetros que se utilizaban para hacer la predicción, el clínico podía entender por qué un paciente particular había sido clasificado como de riesgo alto o bajo.

Pronto se descubrió que el modelo Framingham, aunque eficaz en su contexto original para población estadounidense, tendía a sobre estimar el riesgo en otras poblaciones como la española. Esto llevó a la creación del score REGICOR, adaptado específicamente a nuestra población *(REGICOR – Regicor)*. REGICOR se desarrolló a partir de datos recogidos de una forma sistemática en Girona y otros lugares de España para reflejar mejor la idiosincrasia de su población. Este caso resalta la importancia de entrenar y validar modelos de IA en contextos demográficos y geográficos específicos para garantizar su precisión.

La transparencia en términos de explicabilidad de los modelos de IA es crucial para que los profesionales de salud puedan confiar en las predicciones generadas por dichos modelos. En este sentido la elección del algoritmo condiciona la explicabilidad de los modelos de "inteligencia artificial". En un extremo tenemos algoritmos clásicos como los árboles de decisión que permiten comprender claramente por qué se clasifica a un paciente de una determinada manera. En el otro extremo los recientes "grandes modelos de lenguaje" que se basan en intrincadas redes neuronales profundas con billones de parámetros son extremadamente difíciles de explicar siendo la explicabilidad de estos modelos de IA generativa un campo activo de investigación.

Equidad y no discriminación

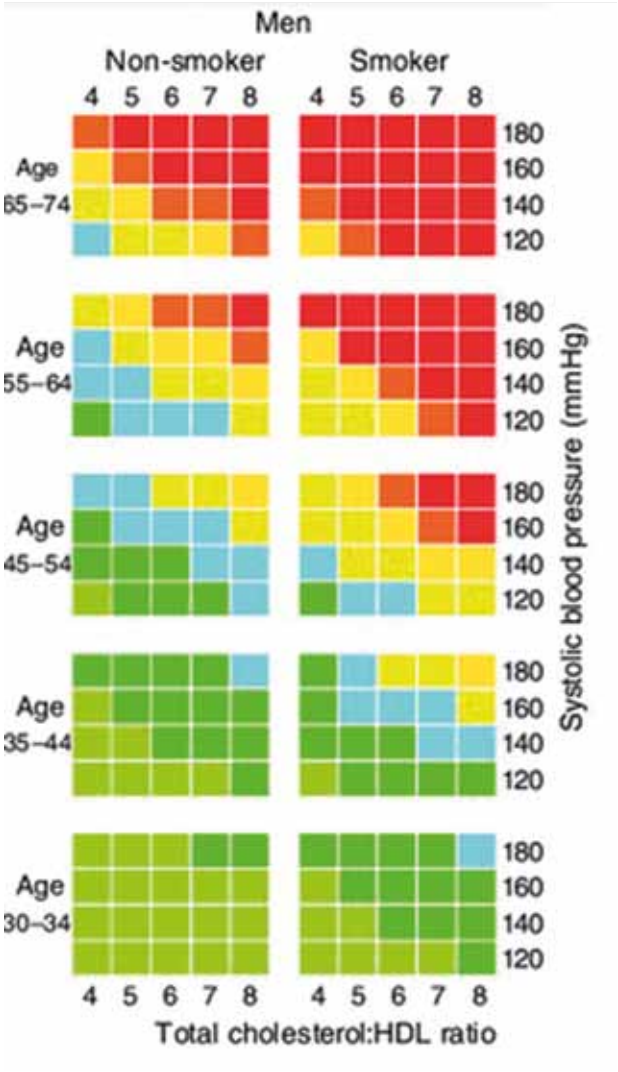
Uno de los desafíos más complejos en la implementación de la IA en salud es evitar los sesgos y asegurar la equidad. Los datos utilizados para entrenar modelos pueden contener sesgos históricos o reflejar desigualdades sociales, lo que puede llevar a decisiones discriminatorias si no se manejan adecuadamente.

Fuera de los ámbitos más técnicos, el término sesgo se suele utilizar de una forma un tanto ambigua por lo que es fundamental diferenciar entre factores de confusión y sesgos.

Los factores de confusión, como la relación aparente entre el uso de anticonceptivos orales y las enfermedades de transmisión sexual, pueden ser mitigados mediante técnicas estadísticas durante el entrenamiento del modelo. Por otro lado, los sesgos inherentes en los datos, como los derivados de cómo se registran ciertos datos clínicos, requieren intervenciones en la recopilación y manejo de los datos mismos.

Numerosas evidencias recientes muestran cómo ciertos sesgos pueden afectar los resultados de la IA en salud. Por ejemplo, un estudio reciente sobre determinantes sociales de la salud *(Factors associated with social determinants of health mentions in PubMed clinical case reports from 1975 to 2022: A natural language processing analysis (accscience.com)* encontró que los casos clínicos publicados en la literatura científica tendían a incluir detalles como la orientación sexual o el ser inmigrante en el caso de ciertas patologías como las infecciones de transmisión sexual. Dado que esos casos clínicos son utilizados durante el entrenamiento de grandes modelos de lenguaje existe el riesgo de que estos modelos reproduzcan dichos sesgos y tiendan a sugerir diagnósticos de enfermedad de transmisión sexual en personas con determinada orientación sexual u origen étnico.

Esta asociación en los datos entre ser inmigrante y tener un diagnóstico de enfermedad infecciosa no implica necesariamente una relación causal directa, sino que refleja un sesgo en el registro de datos clínicos, donde los médicos son más propensos a preguntar y registrar el país de origen o la



orientación sexual en ciertos contextos clínicos.

Naturalmente, estos sesgos deben ser identificados y corregidos para asegurar decisiones justas y precisas.

La complejidad de los modelos modernos

Con la introducción de arquitecturas de redes neuronales profundas como es el caso de los Transformers que se utilizan para entrenar grandes modelos de lenguaje, la explicabilidad se vuelve aún más difícil. Estos modelos, que pueden tener billones de parámetros, son capaces de capturar con gran detalles ciertos patrones en los datos pero resulta muy difícil comprender qué patrones concretos están utilizando al generar sus respuestas. Este nivel de complejidad plantea nuevos retos en términos de transparencia y control.

Por ejemplo, se ha observado que determinados modelos avanzados pueden cambiar significativamente sus predicciones basadas en pequeños cambios en los datos de entrada.

> Numerosas evidencias recientes muestran cómo ciertos sesgos pueden afectar los resultados de la IA en salud. Por ejemplo, un estudio reciente sobre determinantes sociales de la salud encontró que los casos clínicos publicados en la literatura científica tendían a incluir detalles como la orientación sexual o el ser inmigrante en el caso de ciertas patologías como las infecciones de transmisión sexual.

Se ha demostrado empíricamente cómo el simple cambio en la secuencia en la que los síntomas de un caso clínico son presentados a un gran modelo de lenguaje puede impactar en el diagnóstico que el modelo va a sugerir. También cómo la inclusión u omisión de un antecedente de enfermedad mental grave puede hacer que el modelo de lenguaje emita un diagnóstico erróneo al ser consultado por un caso clínico de origen orgánico y no relacionado con el trastorno mental del paciente. En el ámbito clínico, esto puede llevar a decisiones potencialmente perjudiciales para pacientes con enfermedades mentales o con presentaciones clínicas complejas.

Modelos comprensibles, ajustables y representativos

Es por todo esto que a la hora de implementar una IA responsable en salud, sea crucial abordar la transparencia, la explicabilidad, la equidad y la no discriminación desde el inicio. Esto implica desarrollar modelos que sean comprensibles y ajustables, capacitar a los profesionales de salud en el uso y la interpretación de estas herramientas, y asegurar que los datos utilizados para entrenar estos modelos sean representativos y libres de sesgos injustos. Solo a través de un enfoque riguroso y ético podemos maximizar los beneficios de la IA en salud mientras minimizamos sus riesgos.