Aprendizaje automático y salud mental: de la promesa a la aplicación clínica

Jon Garrido-Aguirre

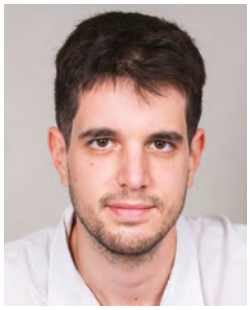
B2SLab,
Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Barcelona, España

Institut de Recerca Pediàtrica Hospital Sant Joan de Déu, Esplugues de Llobregat, Barcelona, España

Enrico Manzini

B2SLab,
Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Barcelona, España

Institut de Recerca Pediàtrica Hospital Sant Joan de Déu, Esplugues de Llobregat, Barcelona, España



Alexandre Perera-Lluna

B2SLab,
Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Barcelona, España

Institut de Recerca Pediàtrica Hospital Sant Joan de Déu, Esplugues de Llobregat, Barcelona, España

John McCarthy, científico estadounidense pionero en el campo de la inteligencia artificial en la década de 1950, describió la disciplina como "la ciencia e ingeniería de construir máquinas inteligentes". En las décadas siguientes el campo de la inteligencia artificial se ha transformado en un terreno amplio y fértil donde conviven diversas disciplinas: asistentes virtuales, software de análisis de imágenes, motores de búsqueda, sistemas de reconocimiento de voz y rostro, o sistemas de inteligencia artificial integrada, incluyendo robots, drones, vehículos autónomos o el internet de las cosas[a].

En este contexto, e independientemente de consideraciones sobre su estatus en relación con la inteligencia artificial, abierto a debate (¿es un subconjunto de la inteligencia artificial o, en cambio, solamente deberíamos considerar alguno de sus elementos como subconjunto de la inteligencia artificial?), el aprendizaje automático (*machine learning*, en inglés) ha cobrado relevancia en los últimos años, hasta el punto de que la definición de McCarthy está más cerca de lograr su pleno significado.

El aprendizaje automático

Por la coincidencia de una serie de factores –la disponibilidad del mayor volumen de datos de la historia y de ordenadores de última generación con gran capa-

[a] Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe, Brussels , 25.4.2018 COM(2018) 237 final.

Foto de Oleksandr Koval en Unsplash.

cidad de cálculo y almacenamiento, junto con avances en el campo de la teoría computacional, principalmente– hoy en día se concentran esfuerzos en el diseño de sistemas que puedan aprender de forma análoga a como lo hacemos los humanos. Los sistemas de aprendizaje automático aprenden a extraer patrones relevantes de los datos en base a la experiencia adquirida con conjuntos de datos que se les han mostrado previamente. La forma en la que extraen estos patrones permite clasificar los sistemas –a partir de ahora, modelos– de aprendizaje automático en tres grandes familias principales: modelos de aprendizaje supervisado, no supervisado y por refuerzo. En el primer caso, el modelo recibirá el conjunto de datos junto a la especificación del resultado deseado, sea este una etiqueta –por ejemplo, "sí enfermedad" o "no enfermedad"–, en problemas de clasificación, o un valor numérico –por ejemplo, la concentración de un fármaco en sangre–, en problemas de regresión, de modo que aprenderá a asignar las etiquetas o valores numéricos correspondientes en función de los datos que se le presenten. Los modelos de aprendizaje no supervisado detectan patrones en los datos sin recurrir a información *a priori* sobre los mismos. El aprendizaje semisupervisado se sitúa a medio camino entre el aprendizaje supervisado y el no supervisado; estos modelos aprovechan datos etiquetados y un volumen mayor de datos sin especificar, lo cual resulta conveniente teniendo en cuenta que realizar la anotación de los datos es a menudo lo más costoso. Finalmente, los modelos de aprendizaje por refuerzo se basan de forma fundamental en la interacción entre el modelo y su entorno mediante un sistema de acciones y recompensas: las acciones correctas no llevan asociada una respuesta específica, sino que cada acción recibe –o no– una señal de refuerzo que modifica el comportamiento del modelo con el objetivo final de maximizar las recompensas.

Los sistemas de aprendizaje automático aprenden a extraer patrones relevantes de los datos en base a la experiencia adquirida con conjuntos de datos que se les han mostrado previamente.

Desde su concepción, y a lo largo de la historia de la disciplina, la inteligencia artificial, en general, y el campo del aprendizaje automático, en particular, han estado influidos por la neurociencia[1]. Desde los tiempos de los pioneros en el campo del aprendizaje automático hasta la época actual, los avances más importantes en este campo –las redes neuronales convolucionales, por poner un ejemplo cercano– se han inspirado en estudios sobre la estructura y funcionamiento del cerebro, y se ha recurrido a la neurociencia como justificación y validación de modelos ya establecidos. Así, en la década de 1940 nace la disciplina moderna del aprendizaje automático con estudios sobre la forma en que el cerebro realiza computaciones, a través del diseño de redes neuronales artificia-

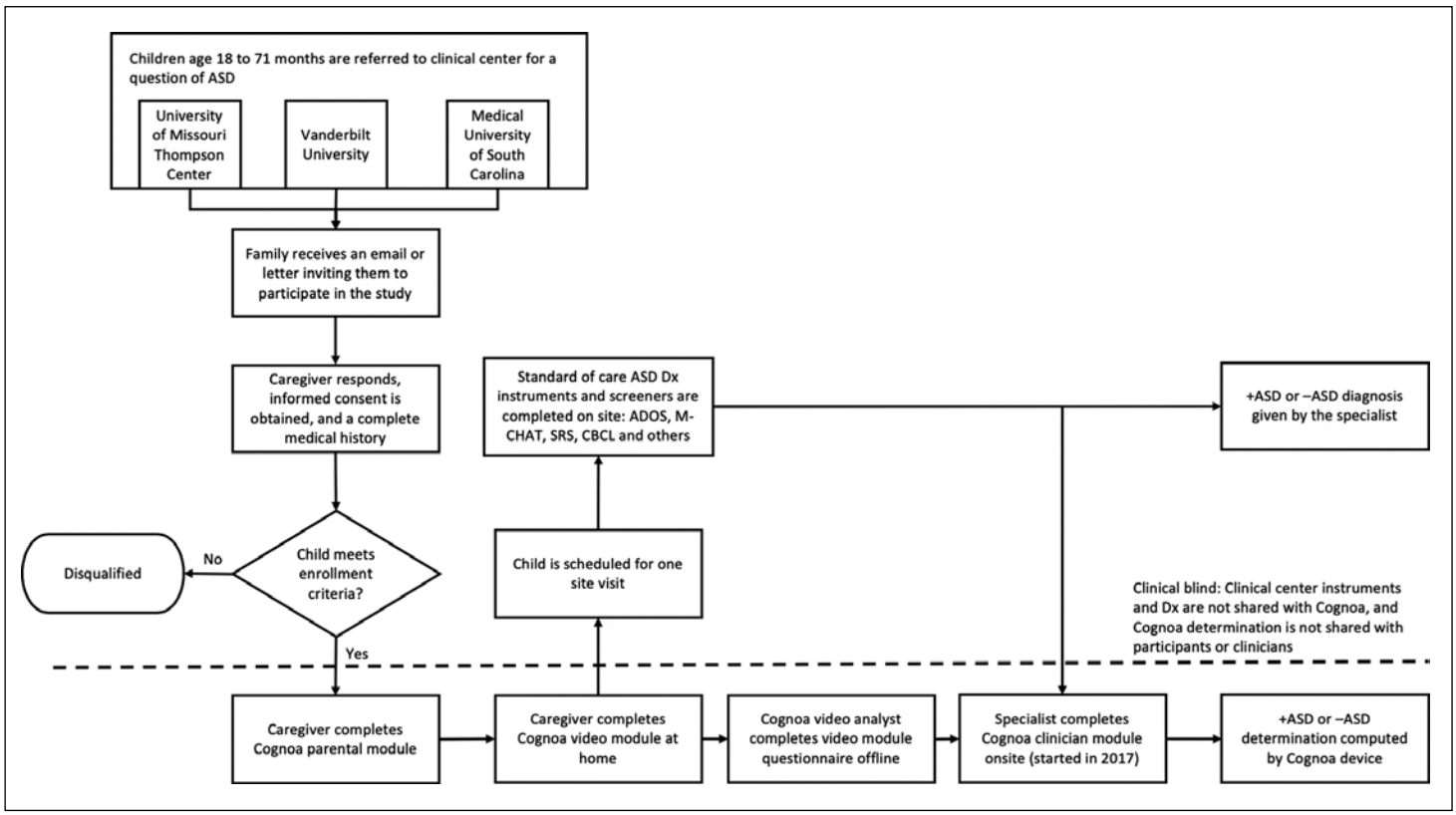
Foto de Testalize.me en Unsplash.

les capaces de implementar funciones lógicas[2]. En la siguiente década continuó el estudio de las redes de neuronas artificiales y se presentó un diseño capaz de aprender a partir de datos de forma supervisada, denominado Perceptrón[3]. En las siguientes décadas el estudio de las redes neuronales quedó aparcado en favor de otras técnicas de aprendizaje, pero en años recientes se ha recuperado el interés en los modelos inspirados en la neurociencia a partir de los resultados obtenidos en el estudio de las redes neuronales profundas[4]. Por analogía inversa, desde un punto de vista fundamental, los últimos avances en aprendizaje automático se postulan como sistemas muy útiles para ayudar en el avance de la neurociencia básica. En este sentido, hay autores que consideran que estudiar los elementos básicos de las redes neuronales profundas –arquitecturas, reglas de aprendizaje, funciones objetivo– puede servir para inspirar avances hacia un nuevo marco teórico en el campo de la neurociencia, dada la dificultad para diseñar modelos a gran escala del cerebro humano[5]. Asimismo, hay trabajos que postulan que el aprendizaje por refuerzo basado en redes neuronales profundas provee de un marco de estudio sobre la forma en que la recompensa da forma a la representación aprendida y cómo lo aprendido va a su vez moldeando los procesos de aprendizaje y la toma de decisiones, dos objetivos de interés para la neurociencia[6].

Aprendizaje automático en el ámbito de la salud mental

Del mismo modo y al igual que ocurre en todas las disciplinas médicas, los modelos de aprendizaje automático están comenzando a considerarse herramientas útiles en el ámbito de la salud mental. Siguiendo el paradigma de la medicina de precisión, el aprendizaje automático aplicado a la salud mental se centra en dos ejes principales[7]: 1) la predicción de la respuesta terapéutica y posibles efectos secundarios de los tratamientos, y 2) el apoyo al diagnóstico diferencial y la detección del riesgo de desarrollar una enfermedad mental.

Debido, principalmente, a la ausencia de biomarcadores objetivos y a una comprensión incompleta de los mecanismos subyacentes a las enfermedades mentales, los resultados de las líneas de tratamiento existentes, en ocasiones, no van asociadas a un diagnóstico, de modo que las prescripciones terapéuticas pueden resultar inadecuadas o inefectivas y sus resultados, inciertos. Esto obliga a visitas clínicas repetitivas y aboca a una estrategia terapéutica de prueba y error, aumentando la probabilidad de discapacidad a largo plazo. Así, la aplicación de técnicas de aprendizaje automático en este ámbito tiene como principal objetivo prede-



Pasos detallados de enfoque multimodular de IA para agilizar el diagnóstico de autismo en niños pequeños.
Imagen de Abbas, H., Garberson, F., Liu-Mayo, S. et al.

cir el tratamiento más adecuado para cada paciente desde los primeros síntomas. Para ello se hace uso tanto de fuentes de datos tradicionales (genética, electrofisiología, neuroimagen, pruebas cognitivas) como de fuentes de datos de nueva disponibilidad (minado de historias clínicas, actividad y uso de teléfonos móviles y otros dispositivos portátiles, actividad en redes sociales). La mayoría de estudios en este ámbito se centran en predecir resultados de tratamientos farmacológicos, muchos de ellos en el tratamiento de la depresión, debido a la prevalencia de la enfermedad y la disponibilidad de datos. En este sentido, el estudio STAR*D (Sequenced Treatment Alternatives to Relieve Depression) es una fuente recurrente de datos[b]. En 2013, utilizando el STAR*D, un estudio prospectivo identificó variables clínicas que influyen en la resistencia al tratamiento con antidepresivos, utilizando modelos de aprendizaje con el objetivo de discriminar entre individuos en remisión tras uno o dos tratamientos farmacológicos frente a aquellos que no logran la remisión[8], con resultados prometedores. Utilizando la misma base de datos y una base de datos independiente (RIS-INT-93) como validación externa del modelo, un estudio posterior[9] identificó que las variables más relevantes para predecir la resistencia al tratamiento son la respuesta inicial al tratamiento y la gravedad de los síntomas a las dos semanas, y demostró, en ausencia de biomarcadores y de forma sólida, la posibilidad de predecir la resistencia al tratamiento antes de iniciar una segunda ronda de tratamientos, con resultados consistentes utilizando diferentes algoritmos de aprendizaje. Más allá de estos ejemplos, en general los estudios desarrollados en el campo terapéutico demuestran que las técnicas de aprendizaje automático pueden ser útiles en el desarrollo de modelos predictivos aplicados a distintos tipos de terapias, si bien la mayoría de resultados son preliminares y no han sido validados en estudios independientes[10].

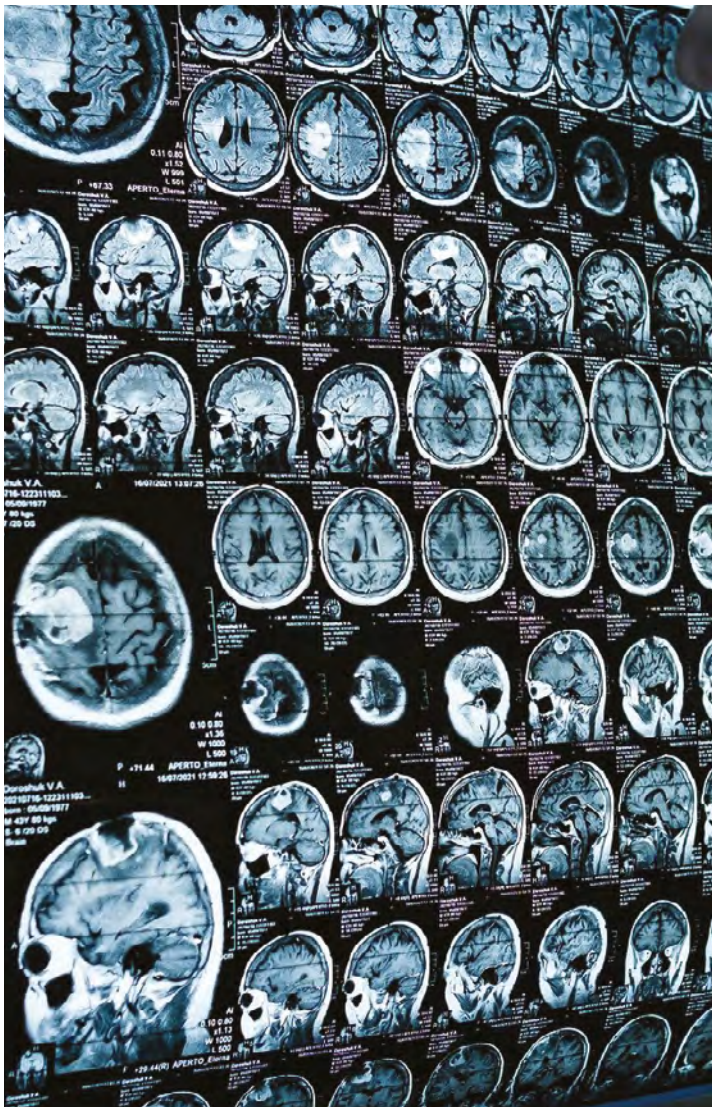
Las técnicas de aprendizaje automático pueden ser útiles en el desarrollo de modelos predictivos aplicados a distintos tipos de terapias.

[b] NIMH » Sequenced Treatment Alternatives to Relieve Depression (STAR*D) Study, https://www.nimh.nih.gov/funding/clinical-research/practical/stard.

Aplicación clínica

Actualmente, el diagnóstico de las enfermedades mentales se basa en sistemas de clasificación basados en signos y síntomas (DSM-5, CIE-10) que no son necesariamente un reflejo de la evidencia neurobiológica. En este contexto, el sistema RDoC (Research Domain Criteria) desarrolla un sistema de clasificación de las enfermedades mentales que busca integrar características genéticas, pruebas cognitivas, pruebas de neuroimagen y neurofisiología y datos obtenidos en cuestionarios y autoinformes. Este nuevo paradigma diagnóstico, que recurre de manera intensiva a fuentes de datos diversas, enlaza directamente con el campo del aprendizaje automático y su aplicación en el diagnóstico y detección. Este ámbito abunda en estudios en los que se aplican técnicas de aprendizaje automático supervisado, muchos de los cuales se centran en desarrollar modelos aplicados a datos procedentes de pruebas de neuroimagen, aunque la búsqueda de soluciones más equitativas requeriría redoblar esfuerzos en estudiar el uso de datos de fuentes más accesibles, por ejemplo, cuestionarios. En este sentido, merece la pena señalar que la aplicación de estas técnicas para la detección temprana de enfermedades mentales está dando sus frutos desde un punto de vista traslacional[11]: recientemente, la Food and Drug Administration (FDA) ha aprobado la comercialización de un sistema de detección precoz de autismo basado en tres modelos de aprendizaje supervisado asociados a cuestionarios[c], dos de los cuales son, en primer lugar, cuestionarios dirigidos a padres a través de una aplicación móvil y, en segundo lugar, cuestionarios clínicos completados por médicos en el momento de la evaluación clínica, a través de un portal online. El tercer módulo está basado en vídeos semiestructurados capturados a través del móvil que son evaluados por expertos humanos mediante un cuestionario. Se aplica un modelo de aprendizaje supervisado para cada tipo de cuestionario y se agregan los resultados en un único valor que determina el diagnóstico (p. ej., "TEA" o "no TEA"). Este ejemplo demuestra dos cosas: en primer lugar, el potencial, aún no explotado, que tiene la utilización de fuentes de datos accesibles para el desarrollo de aplicaciones de aprendizaje automático con utilidad clínica y, en segundo lugar, la validez de las soluciones que incorporan elementos de uso masivo, como los ordenadores portátiles –incluyendo teléfonos móviles y tabletas–, como sistemas de captura de datos para su uso diagnóstico.

[c] Cognoa - Leading the way for pediatric behavioral health, https://cognoa.com/.



A pesar de los resultados prometedores y los –contados– ejemplos exitosos de traslación entre la investigación y la aplicación clínica, el uso de modelos de aprendizaje automático en el ámbito de la salud mental está aún en un estado embrionario, lo que refleja el carácter preliminar de muchos de los estudios publicados hasta la fecha. Para desarrollar modelos de forma que se maximice la probabilidad de pasar del campo conceptual al campo de la práctica clínica es necesario que estos trabajos cumplan con una serie de criterios metodológicos que determinarán la calidad de las soluciones propuestas y, finalmente, su adopción en la labor asistencial del día a día[12]. En primer lugar, y de forma fundamental, los resultados de un modelo deben ser generalizables a datos independientes de los datos a partir de los cuales ha aprendido, sin perjuicio del desempeño del modelo en la tarea asignada. Para ello deben cumplirse varias condiciones, pero, a falta de grandes conjuntos de datos, que es lo habitual en el ámbito de la salud mental, se debe garantizar una muestra que represente la diversidad poblacional en términos de género, raza y características socioeconómicas, para evitar modelos

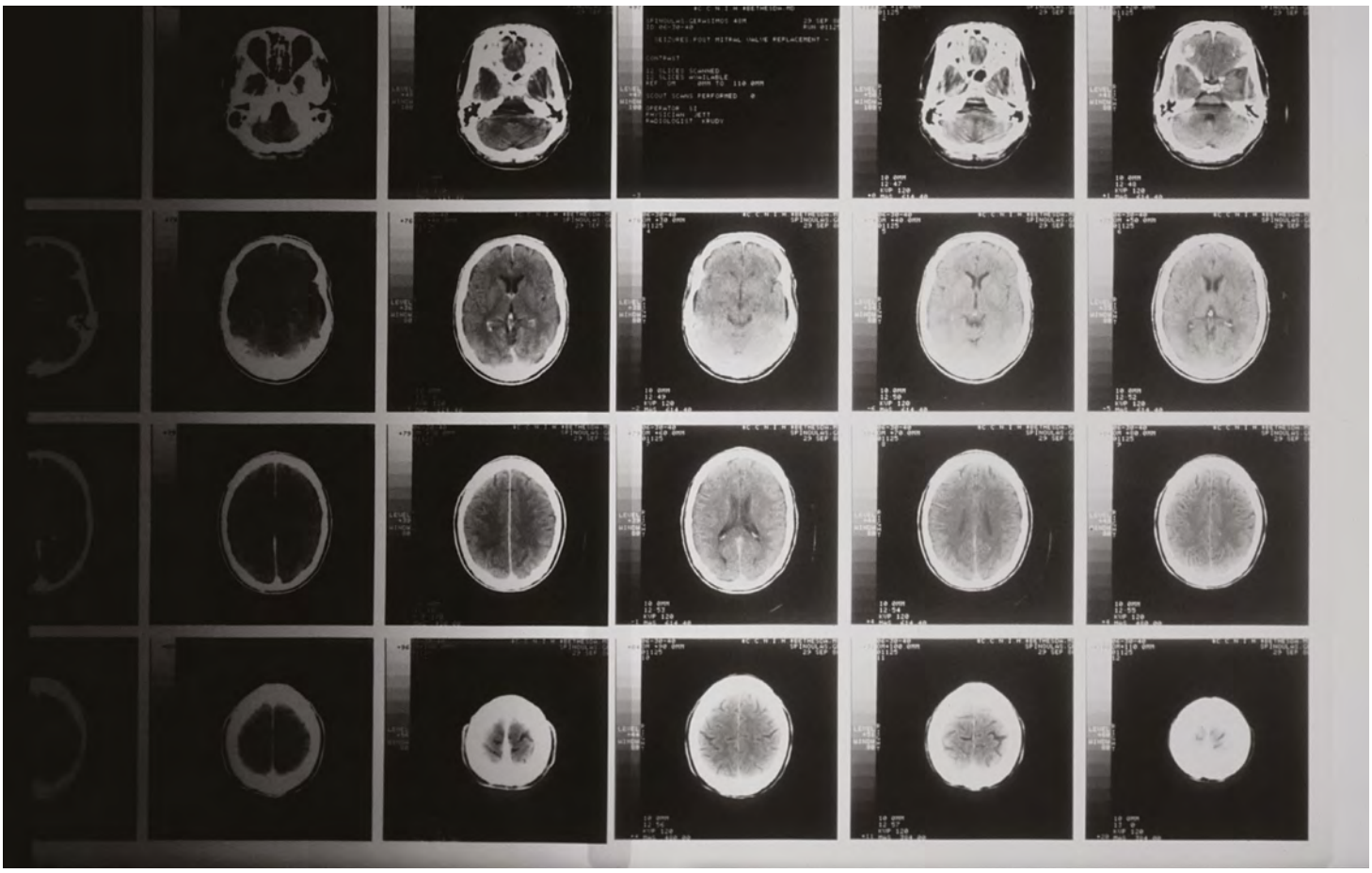


Photo by National Cancer Institute on Unsplash.

sesgados que perpetúen las desigualdades existentes en la sociedad. Además, los modelos deben de entrenarse de forma adecuada, garantizando que los datos con los que se ajusta un modelo son independientes de los datos con los que se evalúa su desempeño y constituyen la validación final del mismo, lo que debería de hacerse con datos procedentes de una muestra diversa extraída, idealmente, de un contexto cultural o sociodemográfico diferente. El criterio de validación externa, de hecho, es ya una condición necesaria para publicar los resultados de modelos de aprendizaje automático en algunas revistas científicas que no aceptan resultados de modelos que incumplan dicho requisito, puesto que es lo que determina la capacidad de generalización de un modelo y, en último término, su aplicabilidad clínica. En segundo lugar, para garantizar la reproducibilidad de los modelos publicados en la literatura y favorecer la responsabilidad y el rigor en la práctica científica, máxime tratándose de una tecnología con potencial impacto directo sobre la vida de las personas, es fundamental proporcionar una descripción detallada de los modelos utilizados, justificando las decisiones de diseño y reconociendo las asunciones realizadas en el proceso, proporcionando a su vez una caracterización completa de los datos, incluyendo una motivación de su encaje con el objetivo final del modelo, así como una guía de uso y aplicación del modelo final. Con este propósito existen ya propuestas de modelos de documentación que funcionarían a modo de informe o ficha técnica tanto de modelos[13] como de datos[14], en un intento de mejorar los estándares de transparencia y reproducibilidad en los estudios sobre modelos de aprendizaje automático. Finalmente, es deseable que los modelos desarrollados sean explicables e interpretables. Cabe señalar que la explicabilidad puede significar distintas cosas según del actor de quien se trate. Así, un investigador en técnicas de aprendizaje automático puede buscar una explicación estudiando las fronteras que traza el modelo para descubrir los patrones en los datos, o caracterizando estos últimos para descubrir cuáles de sus características resultan más relevantes, mientras que un médico quizá la encuentre en la manera en que los resultados del modelo encajan con su concepción del problema. En cualquier caso, resulta de gran importancia conocer por qué un modelo se comporta como se comporta y comprender la manera en la que construye su sistema de discriminación o reconocimiento de patrones en función del espacio de entrada –los datos– y cómo esto se refleja en el desempeño del modelo y las conclusiones que se pueden extraer de los resultados obtenidos. En general, las representaciones internas de los modelos más complejos (redes neuronales profundas, modelos de aprendizaje en conjunto) son más difíciles de caracterizar,

aunque se están investigando técnicas para extraer la información de estas "cajas negras" computacionales y así armonizar su gran desempeño en distintos tipos de problemas con las exigencias de explicabilidad e interpretabilidad.

La aplicación de técnicas de aprendizaje automático en este ámbito tiene como principal objetivo prede cir el tratamiento más adecuado para cada paciente desde los primeros síntomas.

Conclusión

En definitiva, tras una historia fructífera de intercambios entre la disciplina que estudia el funcionamiento de la mente humana y la inteligencia artificial, que aún perdura, ha llegado el momento de que esta última, particularmente el aprendizaje automático, proporcione herramientas que complementen a la experiencia de los profesionales trabajando en el ámbito de la salud mental para mejorar la calidad de vida de todos los pacientes.

Referencias bibliográficas:

1. Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. Neuron, 95(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011.

2. McCulloch, W.S. & Pitts, W. (1943) A Logical Calculus of the Ideas Immanent in Nervous Activity. The Bulletin of Mathematical Biophysics, 5, 115-133. https://doi.org/10.1007/BF02478259.

3. Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386-408. https://doi.org/10.1037/h0042519.

4. Schmidhuber, J. (2015). Deep learning in neural networks: an overview. Neural Networks, 61, 85-117. https://doi.org/10.1016/j.neunet.2014.09.003.

5. Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., Poirazi, P., … Kording, K. P. (2019). A deep learning framework for neuroscience. Nature Neuroscience, 22(11), 1761-1770. https://doi.org/10.1038/s41593-019-0520-2.

6. Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep Reinforcement Learning and Its Neuroscientific Implications. Neuron, 107(4), 603-616. https://doi.org/10.1016/j.neuron.2020.06.014.

7. Shatte, A., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. Psychological Medicine, 49(9), 1426-1448. https://doi.org/10.1017/S0033291719000151.

8. Perlis, R. H. (2013). A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. Biological Psychiatry, 74(1), 7-14. https://doi.org/10.1016/j.biopsych.2012.12.007.

9. Nie, Z., Vairavan, S., Narayan, V. A., Ye, J., & Li, Q. S. (2018). Predictive modeling of treatment resistant depression using data from STAR*D and an independent clinical study. PloS One, 13(6), e0197268. https://doi.org/10.1371/journal.pone.0197268.

10. Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., Dwyer, D., & Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry, 20(2), 154–170. https://doi.org/10.1002/wps.20882.

11. Abbas, H., Garberson, F., Liu-Mayo, S., Glover, E. & Wall, D. P (2020). Multi-modular AI Approach to Streamline Autism Diagnosis in Young Children. Scientific Reports, 10(1), :5014. https://doi.org/10.1038/s41598-020-61213-w.

12. Chandler, C., Foltz, P. W. & Elvevåg, B (2020). Using Machine Learning in Psychiatry: The Need to Establish a Framework That Nurtures Trustworthiness. Schizophrenia Bulletin, 46(1), 11-14. https://doi.org/10.1093/schbul/sbz105.

13. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D. & Gebru, T (2019). Model Cards for Model Reporting. En: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 220-229. https://doi.org/10.1145/3287560.3287596.

14. Gebru, T., Morgenstern, J.H., Vecchione, B., Vaughan, J.W., Wallach, H.M., Daumé, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64, 86-92.

Contacta con nosotros para cualquier pregunta:
brains@clustersalutmental.com
Para contactar directamente con el autor:
Alexandre Perera - alexandre.perera@upc.edu