Modelos predilectos en salud basados en aprendizaje de máquina

Predictive models in health based on machine learning

Denisse Vargas

21 de octubre de 2023

Resumen

La investigación científica en medicina se enfoca en descubrir la relación causa-efecto, que impulsa la creación de intervenciones para tratar o curar enfermedades. La mayoría de los modelos estadísticos tradicionales identifican asociaciones, pero solo un pequeño número de diseños pueden demostrar de manera rigurosa una relación de causa y efecto. La medicina basada en la evidencia se basa en la formulación de hipótesis respaldadas por datos. Esto también se aplica a la creación de modelos predictivos confiables que puedan afectar la atención clínica de manera práctica. El uso creciente de registros clínicos electrónicos y la mejora en la capacidad de procesamiento de datos han dado a las técnicas de aprendizaje automático una posición fundamental en el desarrollo de análisis predictivos y la identificación de patrones previamente desconocidos. Al proporcionar información más precisa y rápida para respaldar la toma de decisiones médicas, estos enfoques computacionales están ganando terreno en la práctica clínica. El objetivo de este artículo es proporcionar una base teórica y evidencia para destacar cómo estas técnicas de aprendizaje automático modernas están produciendo resultados más efectivos y se están adoptando en el ámbito clínico con mayor frecuencia.

1. Introducción

La epidemiología moderna fue establecida por John Snow, quien en 1854 realizó un análisis crítico de la pandemia de cólera en Londres. Snow fue ingenioso al relacionar numerosas variables biodemográficas de causa y efecto para su investigación, y las representó gráficamente en un mapa (Figura 1). Esta visualización convenció a las autoridades para tomar medidas y llevar a cabo una respuesta efectiva para combatir la pandemia porque permitió comprender claramente la distribución de las muertes por cólera alrededor de la bomba de

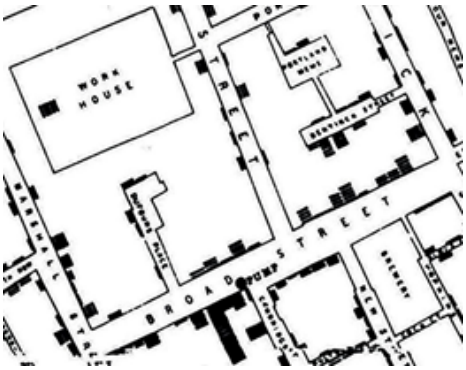

Figura 1: Figura 1. Mapa confeccionado por John Snow de las muertes por cólera ocurridas en el área de Broad Street.

agua contaminada más cercana en Broad Street. Es bien saber que la investigación científica en el ámbito de la salud, ya sea en ciencia básicas o clínicas, ha producido hallazgos y descubrimientos importantes de manera inesperada o no prevista. El descubrimiento de la penicilina es un ejemplo clásico. Esto podría parecerse, manteniendo las proporciones, al enfoque en minería de datos, que se refiere a "dejar que los datos hablen", lo que implica que, a través del análisis de datos, estos proporcionen nueva información, como agrupación o clasificación, distribución, etc., que no se había considerado o planteado previamente [4]. A partir de ese modelo, se comenzó a investigar otros datos de salud en diferentes situaciones, utilizando la misma lógica para encontrar explicaciones y/o correlaciones entre varios eventos y la salud. El registro y análisis sistematizados de datos se han vuelto cada vez más importantes. La aplicación gradual de modelos matemáticos y estadísticos permitió la creación de hipótesis y la confirmación de su validez en relación con un problema. [7] Un modelo fundamental se utiliza en la bioestadística clásica, que es esencial en la medicina basada en la evidencia. Este modelo implica crear una hipótesis y elegir una metodología de investigación adecuada que sea coherente con la evidencia disponible. Se recopilan datos para respaldar el modelo y responder a la pregunta de investigación después de establecer estos elementos. Es importante destacar que con frecuencia hay mala interpretación de los resultados, como la sobrevaloración de la significancia estadística en algunas publicaciones (Figura 2). En contraposición a la bioestadística tradicional, el cambio de paradigma impulsado por el aprendizaje automático y la ciencia de datos, especialmente el procesamiento de grandes volúmenes de información, presenta una perspectiva innovadora. Esta nueva perspectiva utiliza modelos computacionales que permiten obtener datos pertinentes. En este enfoque, el modelo de aprendizaje automático no solo es una herramienta para el investigador, sino que también ayuda a definir y identificar las características del problema que influirán en la respuesta deseada, ya sea por clasificación automática o predicción. Esto ayuda a reducir los sesgos y produce resultados mucho mejores en comparación con los

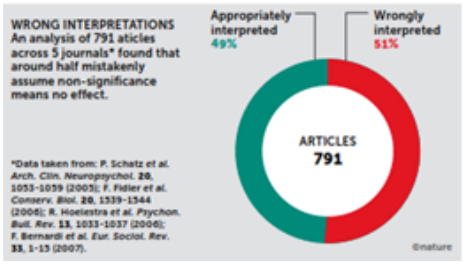

Figura 2: Inadecuada interpretación de la significancia estadística en artículos publicados.

métodos convencionales. Estas técnicas son útiles en medicina y el ámbito de la salud [5]. Los registros clínicos completos se utilizan como datos de entrada para el aprendizaje automático. Estos datos no necesariamente deben ser estructurados o tabulares; pueden presentarse en una variedad de formatos, incluidos texto libre, imágenes, audio, videos y series de tiempo. Los modelos que se derivan de este proceso pueden utilizarse más tarde para ayudar a los profesionales de la salud a realizar diagnósticos precisos para nuevos pacientes. Esto significa que el estudio y diagnóstico de un paciente pueden ser más rápidos, precisos y confiables si se utilizan técnicas de aprendizaje automático en el análisis de datos clínicos. Los registros clínicos completos se utilizan como datos de entrada para el aprendizaje automático. Estos datos no necesariamente deben ser estructurados o tabulares; pueden presentarse en una variedad de formatos, incluidos texto libre, imágenes, audio, videos y series de tiempo. Los modelos que se derivan de este proceso pueden utilizarse más tarde para ayudar a los profesionales de la salud a realizar diagnósticos precisos para nuevos pacientes. Esto significa que el estudio y diagnóstico de un paciente pueden ser más rápidos, precisos y confiables si se utilizan técnicas de aprendizaje automático en el análisis de datos clínicos. (Nota de aclaración: siempre que sea posible, se ha intentado traducir o interpretar los términos en inglés. Sin embargo, debido a que muchas publicaciones en el campo de la informática y la ciencia de datos se realizan en inglés y se utilizan términos técnicos y librerías de programación en ese idioma, en este artículo se presentan los términos en inglés en cursiva y entre paréntesis cuando sea necesario para una interpretación precisa de su significado original o de su expresión en español) [1].

2. APRENDIZAJE DE MÁQUINA VERSUS ESTADÍSTICA

Con base en los conceptos generales de ciencia de datos abordados en un artículo anterior de esta revista y en el libro Üna Mirada a la Era de los Datos.[es]crito por investigadores de la Universidad de Chile, esta revisión se centra en analizar los modelos más utilizados para predecir los resultados fina-

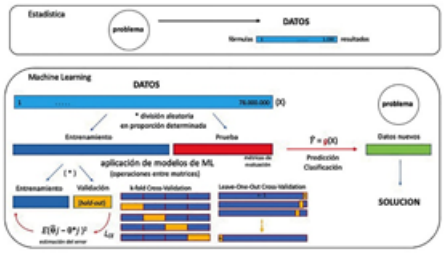

Figura 3: Figura 3. Esquema del paradigma de los datos en estadística y en las técnicas de aprendizaje de máquina.

les en el campo de la medicina. Dentro de estas ideas, es importante resaltar la idea de aprendizaje automático de Mitchell, que se define como: Ün programa computacional se considera que aprende de una experiencia E en relación con una tarea T y una medida de rendimiento P, si su desempeño en la tarea T, evaluado por la medida P, mejora con la experiencia E" [3].

La principal distinción metodológica entre el aprendizaje de máquina y la estadística radica en que el aprendizaje de máquina se enfoca en la creación de conocimiento a partir de datos y su almacenamiento en modelos matemáticos o algoritmos. La transferencia de aprendizaje es el proceso mediante el cual se produce conocimiento que puede usarse para identificar patrones o predecir resultados (figura 3). Es importante destacar que la estadística y la ciencia de datos y el aprendizaje automático no están en conflicto. De hecho, estos modelos computacionales emplean una amplia gama de técnicas estadísticas, especialmente para evaluar y comparar su rendimiento dinámico e iterativo. El resultado de tareas como la clasificación se presenta como una distribución de probabilidades para diferentes posibilidades. La Figura 4 muestra que la estadística es una disciplina clave en el campo de la ciencia de datos. En el pasado, se han desarrollado técnicas de aprendizaje automático y estadística en paralelo. En 1980, se publicó un libro de cuatro estadísticos llamado Çlasificación y Árboles de regresión", cuyas técnicas estadísticas han sido ampliamente utilizadas por investigadores en ciencias de la computación para mejorar el rendimiento de clasificación y para lograr una organización computacional eficiente de procedimientos.

3. APLICACIONES DE MODELOS PREDICTIVOS CON APRENDIZAJE AUTOMÁTICO

Las redes neuronales, como las redes completamente conectadas, convolucionales y recurrentes, así como las máquinas de vectores de apoyo, árboles de decisión, bosques aleatorios, regresiones lineales, modelos bayesianos (naive

Figura 4: Figura 4. Diagrama del concepto de la Ciencia de Datos y las disciplinas que la reúnen.

bayes) y vecinos más cercanos, son los principales modelos predictivos de aprendizaje automático utilizados en medicina. Las redes convolucionales, que fueron propuestas por Yann LeCun en 1989, son algoritmos de aprendizaje automático que se basan en el funcionamiento del cortex visual de los animales. Estas redes "verïmágenes dividiéndolas en campos receptivos o segmentos de la imagen que pasan a través de capas convolucionales, lo que les permite extraer diferentes características de la imagen inicial para su posterior clasificación. Esta técnica se ha utilizado en medicina para procesar una variedad de imágenes con el fin de predecir diagnósticos, desde, biopsias hasta radiografías, escáneres, resonancia nuclear magnética y fotos de diversas patologías [3].

Los modelos de predicción que utilizan redes neuronales se han desarrollado en el campo de las neurociencias. Por ejemplo, Liu del Hospital Universitario de Taiwán desarrolló un modelo llamado EANN AAN13 que utiliza redes neuronales para predecir la muerte cerebral. Rughani comparó los modelos predictivos de redes neuronales utilizados en neurocirugía con las predicciones de los cirujanos y los modelos de regresión lineal, y los resultados mostraron una mejora significativa. Para evaluar la gravedad de las lesiones cerebrales traumáticas, Güler creó un modelo predictivo con redes neuronales que tenía una precisión del 91 %15. Además, para predecir resultados basados en la escala de Glasgow, Low utilizó árboles de decisión y regresión lineal con una precisión de hasta el 80 %. Estos métodos muestran la utilidad de las redes neuronales y otros modelos predictivos en el campo de las neurociencias para mejorar la precisión en la predicción de los resultados médicos [5]. La predicción del comportamiento de los pacientes con respecto a la asistencia a sus citas médicas es un tema importante en la gestión clínica. A nivel global, la tasa de ausentismo oscila entre el 10 % y el 20 %, llegando incluso al 30 % en algunos hospitales de China. Se han utilizado técnicas basadas en redes neuronales profundas (también conocidas

como aprendizaje profundo o aprendizaje representacional) para abordar este problema. Estas estrategias han demostrado ser capaces de disminuir la tasa de pacientes que no acuden a sus citas en un rango del 6 % al 14 %. Además, ayudan a mejorar la atención efectiva de los pacientes, lo que hace que los recursos y el tiempo de atención en los centros de salud sean más eficientes. Estos modelos predictivos se basan en la automatización de la comunicación con los pacientes a través de mensajes de texto para evitar retrasos y ausencias proporcionando información relevante de manera oportuna. Además, pueden automatizar la reprogramación de citas en los espacios que el paciente o el sistema han liberado. Se han llevado a cabo importantes investigaciones en esta área en Chile, con la colaboración del Centro de Modelamiento Matemático de la Universidad de Chile. Una empresa chilena también ha implementado soluciones basadas en redes neuronales para mejorar el acceso a la atención médica en hospitales públicos y privados y optimizar la gestión de citas. Estas innovaciones están mejorando la gestión clínica y la atención de los pacientes [2].

4. PROYECCIONES FUTURAS

Es importante tener en cuenta que incluso un modelo de predicción clínicamente preciso no siempre conducirá han una mejora en la atención médica, a pesar de los avances en la investigación de ciencias de la computación y ciencias de datos aplicadas a la salud. La alta precisión en la predicción de resultados de salud no proporciona automáticamente instrucciones sobre qué hacer para cambiar esos resultados. Además, incluso si se conocen con precisión los resultados de salud previstos, no se puede dar por sentado que sea posible cambiarlos. La investigación en este campo debe ir más allá de la predicción e incluir consideraciones adicionales sobre cómo aplicar el conocimiento en mejoras reales en la atención médica y en la salud de los pacientes. Los modelos predictivos basados en aprendizaje automático ayudan a los profesionales de la salud en su experiencia clínica. Estos modelos se integran en sistemas expertos que codifican y combinan el conocimiento médico y reducen la subjetividad en la toma de decisiones médicas. Al hacerlo, reducen la subjetividad y los sesgos en el proceso de toma de decisiones. Además, estos métodos predictivos pueden generar nuevo conocimiento médico cuando se aplican a diferentes áreas de la medicina o áreas geográficas. Ajustar un modelo de aprendizaje automático con datos locales puede resultar más eficiente que usar una fórmula basada en datos de poblaciones de otros países, lo que resalta la versatilidad y el potencial de estas herramientas en el campo de la medicina [2].

En la actualidad, la realización de estudios de predicciones de resultados clínicos es de gran importancia porque pueden ayudar al equipo médico a tomar decisiones más precisas utilizando técnicas de aprendizaje automático asequibles y fáciles de implementar. La evolución constante de estos modelos y la posibilidad de introducción de nuevos en el futuro prometen proporcionar predicciones cada vez más precisas y dinámicas en la práctica clínica diaria. La librería PyHealth se está desarrollando continuamente para adoptar mejores

prácticas de desarrollo, pruebas e integración interactiva, con un enfoque en la solidez y la escalabilidad. Esto permitirá utilizar más algoritmos y otras librerías de Python comunes en el análisis de datos clínicos. En síntesis, PyHealth está evolucionando para brindar soluciones más efectivas y versátiles en el ámbito de la medicina y la salud [6].

5. CONCLUSIONES

La predicción en medicina no es algo nuevo y ha estado presente en la práctica clínica durante mucho tiempo, con herramientas como puntajes de riesgo para guiar tratamientos o estratificar pacientes en unidades de cuidados intensivos. Sin embargo, en la actualidad es posible desarrollar rápidamente o incluso en tiempo real modelos de predicción para una amplia gama de tareas clínicas gracias a técnicas modernas de aprendizaje automático y el acceso han una variedad de fuentes de datos clínicos. La inclusión de datos no estructurados, como texto libre en registros clínicos o imágenes directas, que antes era difícil de abordar con métodos tradicionales, se beneficia de estos modelos. En medicina, la rapidez en el diagnóstico es crucial, y esta estrategia basada en datos tiene un gran potencial en la detección temprana de alertas critica, diagnósticos de imágenes altamente precisos, optimización de la gestión de citas clínicas y más. En resumen, el uso de aprendizaje automático en datos clínicos está mejorando la atención médica y la toma de decisiones en tiempo real.

Referencias

[1] H.O. Alanazi, A.H. Abdullah, and K.N. Qureshi. A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. *J Med Syst*, 41(4):69, 2017.

[2] V. Amrhein, S. Greenland, and B. McShane. Scientists rise up against statistical significance. *Nature*, 567(7748):305–307, 2019.

[3] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Routledge, 2017.

[4] A. Carreno, I. Inza, and J.A. Lozano. Eventos raros, anomalías y novedades vistas desde el paraguas de la clasificación supervisada. In *IX Simposio de Teoría y Aplicaciones de la Minería de Datos*, pages 925–930, 2018.

[5] A.A.H. de Hond, A.M. Leeuwenberg, L. Hooft, I.M.J. Kant, S.W.J. Nijman, and H.J.A. van Os. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ Digit Med*, 5(1):2, 2022.

[6] J. Dunstan, A. Maass, and F. Tobar. *Una Mirada a la Era de los Datos*. Ed. Universitaria, 2022.

[7] J. Mora. Proyecciones de la ciencia de datos en la cirugía cardíaca. *Rev Med Clin Condes*, 33(3):294–306, 2022.