Aprendizaje automático aplicado en área de la salud. Parte 1

Nicolas H. Quiroz, María Lourdes Posadas-Martínez, Emiliano Rossi, Diego H. Giunta y Marcelo R. Risk

RESUMEN

Este será el primero de dos artículos donde se tratarán los pasos necesarios para desarrollar un proyecto de aplicación de técnicas de *Machine Learning* en Salud, que introduce nociones sobre la recolección y análisis de datos, la selección y entrenamiento de modelos de aprendizaje automático de tipo supervisado y los métodos de validación interna para cada modelo.

Palabras clave: aprendizaje automático, cuidado de la salud, historia clínica electrónica, desarrollo, investigación, colaboración multidisciplinaria.

MACHINE LEARNING IN HEALTHCARE. PART 1

ABSTRACT

This will be the first of two articles where the steps needed to apply machine learning methods in healthcare will be discussed. It will introduce fundamental notions about data collection, selection and training of supervised ML models as well as the methods of internal validation. In a second article, we will discuss about the performance evaluation to select the most appropriate model and its external validation.

Key words: machine learning, healthcare, electronic health record, development, research, transdisciplinary collaboration.

Rev. Hosp. Ital. B.Aires 2021; 41(4): 206-209.

INTRODUCCIÓN

En las últimas décadas, la evolución de los recursos tecnológicos ayudó a la detección temprana de algunas enfermedades. La gran mayoría de estos avances fueron consecuencia de la globalización de la información, comenzando por las ciencias básicas hasta alcanzar las tecnologías traslacionales. Con la mayor capacidad de procesamiento y almacenamiento digital hubo un elevado crecimiento en la recolección de datos biomédicos[1]. La gran cantidad de información sobre un mismo paciente generada en distintos puntos de atención hace que su análisis manual sea poco eficiente[1-3] y que se reduzca el tiempo que los profesionales de la salud pueden dedicar al paciente en la consulta[4]. Además, aumenta la posibilidad de cometer errores en el informe médico[5]. Recordemos que lo importante no es la recolección de datos *per se* sino extraer conocimiento de ellos y poder aplicarlo[6].

En este contexto surgió una nueva disciplina basada en la estadística y la informática denominada Aprendizaje Automático en Informática Médica (*Machine Learning in Healthcare Informatics*)[7]. La Informática Médica es el campo científico multidisciplinario que estudia y se centra en el uso efectivo de datos, información y conocimientos biomédicos para la indagación científica, la solución de problemas y la toma de decisiones a fin de mejorar la salud humana[8,9]. Los datos pueden ser palabras, números, diagramas, imágenes y videos (o cualquier combinación de estos) cuyo objetivo es capturar el estado de salud de un paciente. A partir de los datos, contextualizando y dándoles significado, generamos información. A diferencia del ser humano, la computadora puede procesar datos de miles de pacientes en pocos segundos, dando como resultado sistemas informáticos capaces de orientar al médico en la toma de mejores decisiones diagnóstico-terapéuticas. No cabe duda de que la historia clínica es el eje de la comunicación entre los distintos profesionales que atienden al paciente y se ha demostrado el beneficio de contar con su registro electrónico[10,11]. Estas historias clínicas electrónicas (HCE) se organizan en bases de datos explotables mediante estrategias de *data mining* e inteligencia artificial capaces de analizar/clasificar información de manera automática. Utilizando esta información se ha logrado identificar, generar alertas y hasta predecir enfermedades de manera automática con elevada eficacia[12]. Incluso en grupos reducidos de pacientes con enfermedades poco frecuentes, que no llegan a tener un diagnóstico certero debido a la escasez de información, estas técnicas podrían utilizarse para detectarlos en estadios tempranos permitiendo que reciban tratamientos que reduzcan la morbimortalidad.

Recibido 15/11/21 Aceptado 6/12/21

Instituto de Medicina Traslacional e Ingeniería Biomédica (IMTIB) (N.H.Q, M.L..M., D.H.G., M.R.R.), CONICET - Instituto Universitario del Hospital Italiano de Buenos Aires - Hospital Italiano de Buenos Aires. Área de Investigación en Medicina Interna (M.L.P.S., D.H.G.), Servicio de Clínica Médica. Departamento de Investigación (M.L.P.M., E.R.). Servicio de Cardiología (E.R.). Hospital Italiano de Buenos Aires. Buenos Aires, Argentina
Correspondencia: nicolas.quiroz@hospitalitaliano.org.ar

Este será el primero de dos artículos donde se tratarán los pasos necesarios para desarrollar un proyecto de aplicación de técnicas de *Machine Learning* en Salud, que introduce nociones sobre la recolección y análisis de datos, la selección y entrenamiento de modelos de aprendizaje automático de tipo supervisado y los métodos de validación interna para cada modelo. En un segundo artículo trataremos la evaluación del rendimiento para la selección del modelo más adecuado y su validación externa.

PROCEDIMIENTO DE RECOLECCIÓN Y ANÁLISIS DE DATOS

El primer paso consiste en la recolección de las HCE de los pacientes, por lo general provenientes del centro de datos (*data center*) de la institución de salud. En esta etapa se deberán identificar las bases médicas con las que se realizará el trabajo. Es importante identificar el procedimiento adecuado a cada Institución de salud, que puede precisar pedidos a mesas de ayuda o el acceso autorizado directo a las bases de datos. Por lo general, puede ocurrir que estas bases de datos contengan características no deseables como por ejemplo datos duplicados, observaciones irrelevantes, datos faltantes, entre otras. Es importante aplicar en este momento estrategias de curado de datos (*data cleaning*) para asegurar que los datos mantengan correspondencia con los valores reales, sean interpretables por los médicos expertos para realizar controles de calidad, estén bien organizados y sean representativos de la población. La premisa fundamental en este punto es que los modelos de *Machine Learning* aprenden de los datos, aplicándose la idea de GIGO (*"garbage in, garbage out*", traducible a "entra basura, sale basura"). Muchas veces también es posible aplicar en esta instancia estrategias automáticas como son las técnicas de extracción de conocimiento para la detección de valores atípicos (*outliers*) o erróneos. Todo este proceso se resume en la figura 1.

SELECCIÓN Y ENTRENAMIENTO DE MODELOS DE APRENDIZAJE AUTOMÁTICO SUPERVISADO

El aprendizaje automático supervisado es la utilización de algoritmos que aprenden a partir de datos previamente etiquetados. La clasificación supervisada es una de las tareas más frecuentemente realizadas por los sistemas de aprendizaje automático. Esta sección describe algunos algoritmos de aprendizaje automático supervisado de uso médico como: regresión logística, clasificador *naïve*

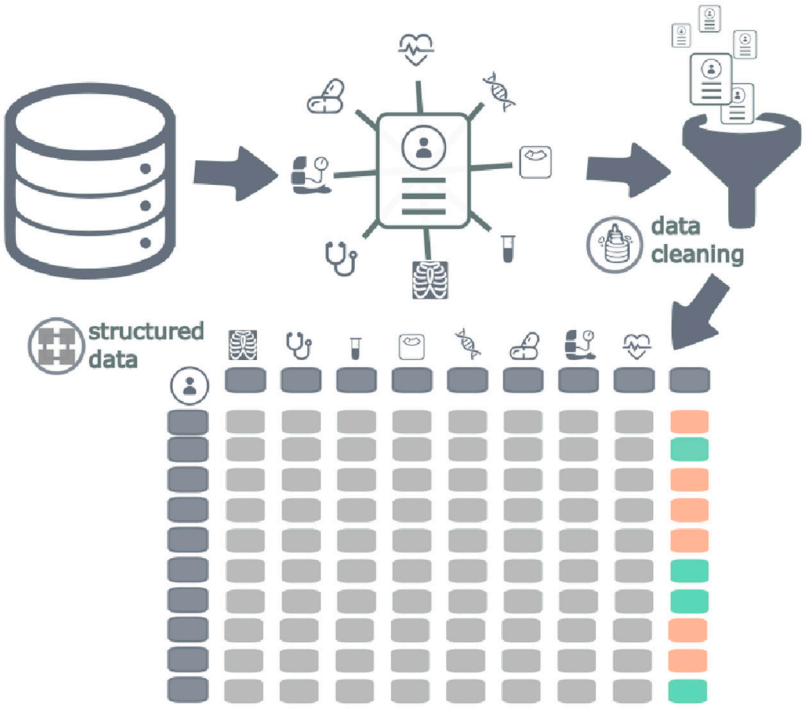

Figura 1. Esquema de recolección de datos y su análisis. El proceso se conoce como ETL (extraer, transformar y cargar, por sus siglas en inglés). Es el proceso de compilación de datos a partir de un número de fuentes, su posterior organización (*structured data*) y centralización en una base de datos.

Bayes, árboles de decisión, bosques aleatorios (*random forest*) y máquinas de soporte vectorial (SVM).

Regresión logística

Es un método estadístico que aplica la lógica de la regresión lineal a la predicción de etiquetas a partir de modelizar los *odds ratio*. El resultado que arroja el modelo es una medida de la probabilidad de que un conjunto de datos corresponda a una etiqueta determinada, por lo que también es posible interpretarlo como una medida de la incertidumbre en la clasificación.

Clasificador *naïve* Bayes

Se puede utilizar para determinar las probabilidades de las clases dada una serie de observaciones diferentes aplicando el principio de probabilidad condicional (es decir, cuál es la probabilidad de que la salida sea de una determinada etiqueta dadas las observaciones de entrada). La simplicidad del modelo radica en asumir que las variables de entrada son condicionalmente independientes dada la clase, lo que también representa su mayor debilidad.

Árboles de decisión

Se utilizan para clasificar mediante la partición sucesiva de los datos de entrada de acuerdo con métricas de pureza y homogeneidad (entropía, chi-cuadrado, entre otras). El método busca de esta manera encontrar las reglas de partición que permite separar óptimamente los datos. Esta técnica posee la ventaja de ser altamente interpretable, ya que cada nodo del árbol divide los datos en ramas de acuerdo con umbrales o condiciones en alguna de las variables de entrada.

Random forest (bosque aleatorio)

Como su nombre lo indica, es la combinación de una gran cantidad de árboles de decisión independientes probados sobre conjuntos de datos aleatorios y aplicado sobre un subconjunto aleatorio de variables de entrada. El resultado es un ensamble de cientos de árboles cuyos votos permiten clasificar por mayoría a qué clase pertenece un dato de entrada.

SVM (máquinas de soporte vectorial)

Este modelo es geométrico no probabilístico, en el sentido de que considera los datos de entrada simplemente como puntos en un espacio de características, e intenta encontrar un hiperplano que separe de la mejor manera posible, dejando la mayor brecha entre aquellos de una clase respecto de la otra.

VALIDACIÓN DEL MODELO

Una vez desarrollado e implementado el modelo hay que validarlo, es decir, corroborar si funciona igual en otros grupos distintos de aquellos que se han empleado para su desarrollo. Básicamente se evalúa cuán bien se predice la variable resultado en nuevos grupos.

Validación interna: incluye diferentes técnicas, *data splitting*, *cross-validation* y *bootstrap-validation.* Estas utilizan los datos de los pacientes que se utilizaron para el desarrollo (entrenamiento y prueba) del modelo.

- *data splitting*

La totalidad de los pacientes para incluir (conjunto completo) se dividirá en dos conjuntos no superpuestos: el primero *training dataset* (conjunto de entrenamiento) estará compuesto del 70-80% del conjunto completo y de aquí el algoritmo seleccionado extraerá los parámetros para su ajuste de entrenamiento. El porcentaje restante se le asigna al *testing dataset* (conjunto de prueba), el cual contiene datos que nunca se han utilizado en el entrenamiento y contribuye a la evaluación final del modelo[13]; en la figura 2 se puede visualizar un esquema simplificado del proceso. Para verificar la reproducibilidad del modelo se utilizará submuestreo aleatorio simple sin reposición.

- *cross-validation o validación cruzada*

Consiste en dividir el conjunto de datos en dos partes del mismo tamaño, se entrena con el *training dataset* y se evalúa con en el *testing dataset*. Se repite el proceso intercambiando los grupos para entrenar y para evaluar. El rendimiento total del modelo se obtiene como promedio del rendimiento en cada etapa del conjunto de evaluación. Si se repite k cantidad de veces, siendo k los grupos, se lo conoce como *k-folds cross validation*.

- *bootstrap-validation*

Es una técnica de simulación propuesta para generar observaciones a partir de las distribuciones de la muestra original (conjunto [*set*] de datos completo). Es recomendable utilizar esta técnica cuando tenemos *set* de datos pequeños.

CONCLUSIÓN

En este artículo hemos introducido al lector en un campo ya conocido como es el aprendizaje automático pero con un objetivo en salud, mencionando la recolección y análisis de datos, la selección y entrenamiento de modelos de aprendizaje automático de tipo supervisado y los métodos de validación interna. En el siguiente artículo comentaremos cómo se realiza la validación externa del modelo, además de las métricas de evaluación para decidir cuál es el mejor modelo que se ajusta a nuestro objetivo.

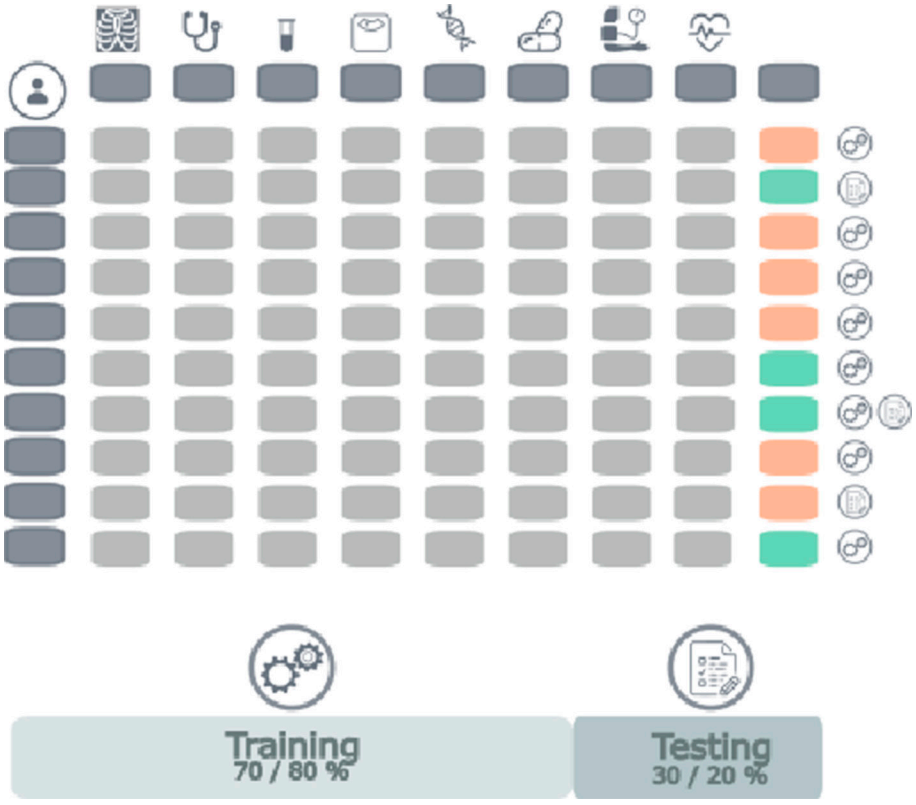

Figura 2. Esquema de entrenamiento y prueba.

Conflictos de interés: los autores declaran no tener conflictos de interés.

REFERENCIAS

1. Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: a literature review. Biomed Inform Insights. 2016;8:1-10. https://doi.org/10.4137/BII.S31559.
2. Holzinger A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform. 2016;3(2):119-131. https://doi.org/10.1007/s40708-016-0042-6.
3. Obermeyer Z, Lee TH. Lost in thought: the limits of the human mind and the future of medicine. N Engl J Med. 2017;377(13):1209-1211. https://doi.org/10.1056/NEJMp1705348.
4. Read-Brown S, Hribar MR, Reznick LG, et al. Time requirements for electronic health record use in an academic ophthalmology center. JAMA Ophthalmol. 2017;135(11):1250-1257. https://doi.org/10.1001/jamaophthalmol.2017.4187.
5. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-1358. https://doi.org/10.1056/NEJMra1814259.
6. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216-1219. https://doi.org/10.1056/NEJMp1606181.
7. Chowriappa P, Dua S, Todorov Y. Introduction to machine learning in healthcare informatics. En: Dua S, Acharya UR, Dua P, eds. Machine learning in healthcare informatics [Internet]. Berlin: Springer Verlag; 2014 [citado 2021 nov 10]. p. 1-23. https://doi.org/10.1007/978-3-642-40017-9_1.
8. Baştanlar Y, Özuysal M. Introduction to machine learning. En: Yousef M, Allmer J, eds miRNomics: microRNA biology and computational analysis [Internet]. Totowa, NJ: Humana Press; 2014 [citado 2021 nov 10]. p. 105-128. https://doi.org/10.1007/978-1-62703-748-8_7.
9. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920-1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593.
10. Natarajan P. Healthcare and the big data V's. En: Natarajan P, Frenzel JC, Smaltz DH. Demystifying big data and machine learning for healthcare [Internet] Boca Raton: CRC Press; 2017 [citado 2021 nov 10]. p. 11-30. https://doi.org/10.1201/9781315389325-2.
11. Stephens ZD, Lee SY, Faghri F, et al. Big data: astronomical or genomical? PLoS Biol. 2015;13(7):e1002195. https://doi.org/10.1371/journal.pbio.1002195.
12. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis. 2018;66(1):149-153. https://doi.org/10.1093/cid/cix731.
13. Suthaharan S. Supervised learning models. En: Machine learning models and algorithms for big data classification [Internet]. Boston: Springer; 2016 [citado 2021 nov 10]. p. 145-181. https://doi.org/10.1007/978-1-4899-7641-3_7.