



# Nueva metodología basada en 'machine learning' para mejorar la explotación de datos médicos

- Un equipo del Instituto de Salud Carlos III (ISCIII) ha utilizado una nueva herramienta complementaria a la inteligencia artificial, basada en el denominado 'machine learning', para generar una nueva metodología capaz de mejorar la extracción, selección y tratamiento de datos biomédicos y sanitarios.

**26 de junio de 2025.** Un artículo publicado por un equipo del Instituto de Salud Carlos III (ISCIII) presenta una innovadora metodología para extraer conocimiento científico de bases de datos médicas gracias al uso de técnicas de machine learning.

[El estudio aparece en la revista \*Mathematics MDPI\*.](#)

Los investigadores Ricardo Sánchez de Madariaga, Mario Pascual Carrasco y Adolfo Muñoz Carrero, de la Unidad de Investigación en Salud Digital (UITeS) del ISCIII, han trabajado con un nuevo enfoque complementario a la inteligencia artificial, denominado Dataset Feature Splitting (DFS). Esta herramienta permite generar diferentes distribuciones de probabilidad a partir de un conjunto de datos (dataset) original, identificando qué algoritmos basados en machine learning se adaptan mejor a la naturaleza estadística de los datos. El resultado final es una mejor extracción, selección y tratamiento de los datos biomédicos y sanitarios, con capacidades superiores a métodos clásicos como la regresión logística.

Esta nueva metodología se apoya en modelos de lenguaje, como ChatGPT y Google Gemini, que ayudan a codificar conocimiento médico con una cantidad mínima de datos, permitiendo así identificar automáticamente los algoritmos más eficaces para cada caso de búsqueda y procesamiento de la información para la extracción de conocimiento. Los autores explican que se trata de "un importante avance para mejorar la explotación inteligente de datos médicos, con gran potencial para la investigación en salud".



Por ejemplo, la metodología que presenta el equipo del ISCIII puede determinar si un conjunto específico de datos médicos es válido para representar y extraer un determinado conocimiento, utilizando un subconjunto de algoritmos estándar supervisados de machine learning. Si el análisis estadístico y la inteligencia artificial considera que el proceso es viable para los objetivos planteados, se determina el subconjunto de algoritmos más adecuado para extraer y sistematizar el conocimiento médico deseado del conjunto de datos.

De esta forma, a partir de un dataset con una serie de características médicas como pueden ser analíticas, comorbilidades o medicamentos de un conjunto de pacientes, la metodología devuelve cuáles de esas características tienen relación con la outcome (enfermedad) del dataset, desde el punto de vista del conocimiento médico. Esto puede ser de mucha utilidad para los médicos al manejar un conjunto reducido de características para diagnosticar rápidamente la outcome médica considerada.

Estos algoritmos sugieren una serie de correlaciones que pueden ser posibles causalidades, por lo que, a partir de este punto, los profesionales médicos deben validarlas partiendo del conocimiento científico que les confiere su formación y experiencia.

- **Referencia del artículo:** *Sánchez-de-Madariaga, R.; Pascual Carrasco, M.; Muñoz Carrero, A. A Methodology to Extract Knowledge from Datasets Using ML. Mathematics 2025, 13, 1807. <https://doi.org/10.3390/math13111807>.*