EVIDENCIAS EN PEDIATRÍA
Toma de decisiones clínicas basadas en las mejores pruebas científicas
www.evidenciasenpediatria.es

Fundamentos de medicina basada en la evidencia

Preprocesamiento de datos para uso en aprendizaje automático

Molina Arias M
Servicio de Gastroenterología. Hospital Universitario La Paz. Madrid. España.

Correspondencia: Manuel Molina Arias, mma1961@gmail.com

Palabras clave en español: aprendizaje automático; datos de salud; datos faltantes; estandarización; ingeniería de variables; preparación de datos.
Palabras clave en inglés: machine learning; health data; missing data; standardization; features engineering; data preparation.

Fecha de recepción: 27 de agosto de 2025 • **Fecha de aceptación:** 15 de septiembre de 2025
Fecha de publicación del artículo: 1 de octubre de 2025

Evid Pediatr. 2025;21:36.

CÓMO CITAR ESTE ARTÍCULO

Molina Arias M. Preprocesamiento de datos para uso en aprendizaje automático. Evid Pediatr. 2025;21:36.

Para recibir Evidencias en Pediatría en su correo electrónico debe darse de alta en nuestro boletín de novedades en
http://www.evidenciasenpediatria.es

Este artículo está disponible en: http://www.evidenciasenpediatria.es/EnlaceArticulo?ref=2025;21:36.
©2005-25 • ISSN: 1885-7388

Este es un artículo Open Access bajo la licencia
CC BY-NC-ND (Reconocimiento-No comercial-Sin obras derivadas): https://creativecommons.org/licenses/by-nc-nd/4.0/deed.es

Preprocesamiento de datos para uso en aprendizaje automático

Molina Arias M
Servicio de Gastroenterología. Hospital Universitario La Paz. Madrid. España.

Correspondencia: Manuel Molina Arias, mma1961@gmail.com

IINTRODUCCIÓN

En los últimos años, el crecimiento exponencial del volumen y variedad de datos disponibles en salud, junto al desarrollo de los nuevos modelos de aprendizaje automático e inteligencia artificial, han supuesto un cambio de paradigma en la forma en que se diseñan los estudios clínicos, se construyen modelos predictivos y se toman decisiones asistenciales[1]. Desde las historias clínicas electrónicas hasta los dispositivos de monitorización continua, pasando por registros administrativos, encuestas poblacionales, estudios observacionales y sistemas de triaje[2], los datos están cada vez más presentes en la actividad clínica diaria.

No obstante, disponer de grandes volúmenes de datos no garantiza su utilidad. El conocido aforismo de que "los científicos de datos dedican el 80% de su tiempo a limpiar datos y el otro 20% a quejarse por tener que hacerlo" refleja una realidad técnica incuestionable: los datos clínicos, en bruto, no suelen estar listos para ser analizados[3]. Con frecuencia contienen errores, valores faltantes, unidades inconsistentes, codificaciones heterogéneas o estructuras poco adecuadas para su análisis con los modelos disponibles. Esta situación puede llevar a interpretaciones erróneas, modelos sesgados o decisiones clínicas inadecuadas si no se abordan correctamente[4].

En este contexto, la manipulación y preparación de los datos, ya utilizada con los modelos estadísticos "convencionales", resurge como una competencia clave. Comprende un conjunto de técnicas destinadas a mejorar la calidad de la información y a transformarla en un formato útil para su análisis. En el caso particular de la Pediatría, donde la edad y el desarrollo del paciente condicionan los rangos fisiológicos y la presentación clínica, este proceso debe contemplar, además, elementos como la variabilidad por edad, la dinámica del crecimiento o las diferencias entre cohortes pediátricas.

En este artículo se revisan los conceptos clave y las estrategias prácticas para la limpieza, transformación e imputación de datos clínicos. Además, se abordan los tipos de datos faltantes y sus implicaciones analíticas, así como las técnicas de codificación, normalización y reducción de dimensionalidad necesarias para aplicar algoritmos de aprendizaje automático o realizar análisis estadísticos válidos.

Una adecuada preparación de los datos es indispensable para garantizar modelos predictivos fiables, decisiones clínicas informadas y una práctica ética basada en evidencia sólida.

LOS DATOS EN EL ÁMBITO SANITARIO

La estructura de los datos sanitarios es intrínsecamente compleja. Por un lado, las fuentes de datos son diversas: desde registros electrónicos y sistemas de gestión hospitalaria, hasta dispositivos portátiles como pulsioxímetros o monitores de actividad física[5]. Estos datos pueden estar estructurados, como en una base de datos tabular donde se registran edad, peso o diagnóstico codificado, o no estructurados, como notas clínicas en texto libre, informes de laboratorio escaneados o imágenes médicas[6].

Además, como ya se ha comentado, la variabilidad biológica en Pediatría es sustancial. A diferencia de los adultos, los parámetros clínicos de referencia cambian según la edad, el sexo y el estadio de desarrollo. Una frecuencia cardíaca considerada normal en un lactante puede representar taquicardia en un escolar. Esto requiere adaptar los algoritmos de validación de datos para detectar valores atípicos o imposibles sin incurrir en errores clínicos.

Otro factor relevante es la fragmentación de los registros[7]. Los pacientes pediátricos pueden ser atendidos en distintos niveles asistenciales (atención primaria, urgencias, hospitalización) y es frecuente que los sistemas informáticos no estén plenamente integrados. Esta fragmentación complica la unificación de fuentes, genera duplicaciones y dificulta el seguimiento longitudinal del paciente.

Todo lo anterior pone de manifiesto que, antes de aplicar cualquier técnica analítica o modelo de aprendizaje automático, es imprescindible realizar un análisis exploratorio inicial que permita detectar inconsistencias, errores sistemáticos y patrones de datos faltantes. Esta exploración debe ir seguida de un proceso de limpieza y preprocesamiento cuidadosamente diseñado, cuya complejidad y profundidad dependerán en buena medida del tipo de modelo que se pretenda utilizar, desde regresiones simples hasta redes neuronales profundas[8,9]. Solo así se garantiza que los datos alimenten modelos válidos, fiables y clínicamente útiles.

ANÁLISIS EXPLORATORIO Y LIMPIEZA DE LOS DATOS

Antes de cualquier análisis estadístico o aplicación de modelos de aprendizaje automático, los datos deben someterse a una etapa de limpieza. Este proceso busca identificar y corregir errores sistemáticos, inconsistencias y datos faltantes, así como validar que las variables estén correctamente definidas[10].

La limpieza de datos comienza con una exploración inicial: se calculan estadísticas descriptivas como medias, medianas, máximos y mínimos, y se revisa la frecuencia de los valores para detectar posibles errores de codificación o entradas erróneas[11]. También se identifican duplicados, se verifican las unidades de medida y se detectan valores imposibles (por ejemplo, un peso de 500 kg en un neonato).

A esta etapa se le suma el análisis gráfico: histogramas, diagramas de caja o gráficos de dispersión permiten visualizar la distribución de las variables, identificar valores atípicos y entender la relación entre los distintos parámetros. En Pediatría, este análisis es fundamental para ajustar los rangos según la edad, y evitar eliminar registros válidos por considerarlos anómalos desde una perspectiva "adulta".

IDENTIFICACIÓN Y MANEJO DE DATOS FALTANTES

Uno de los problemas más frecuentes (y potencialmente más peligrosos) en el análisis de datos clínicos es la presencia de datos faltantes[12]. Su origen puede ser técnico (fallo en la captura del dato), organizativo (registro incompleto) o comportamental (rechazo del paciente o familia a responder una pregunta). Otras razones posibles, especialmente en nuestro entorno, son la falta de información por cuestiones de acceso, tiempo limitado en la consulta o incluso factores culturales.

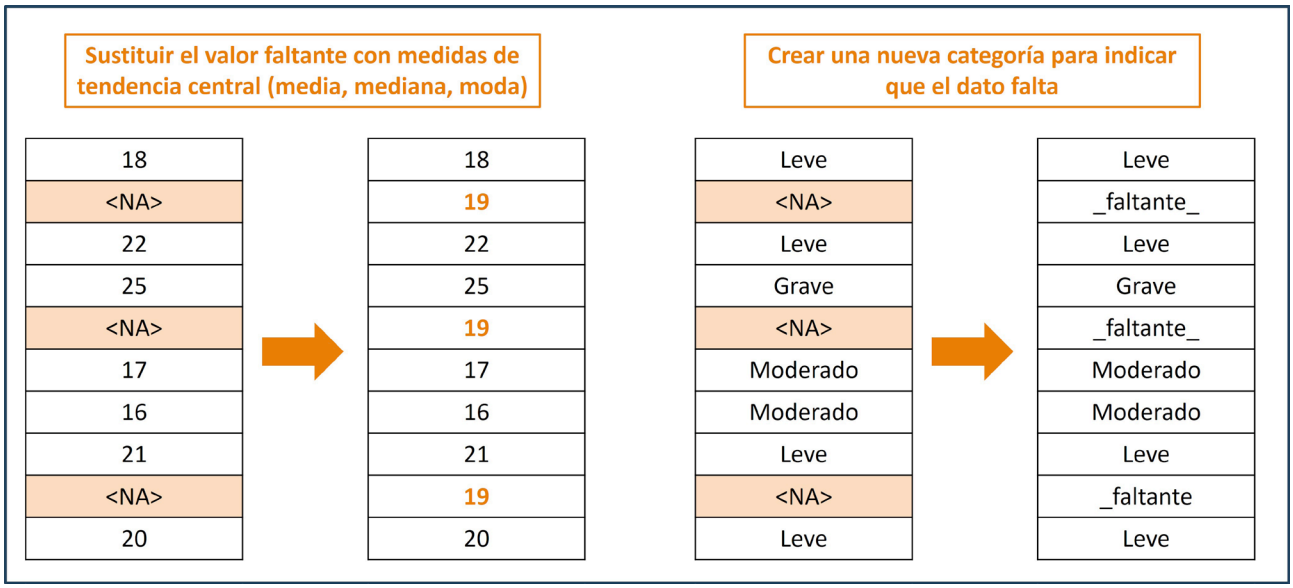
Desde el punto de vista metodológico, es fundamental distinguir el mecanismo por el cual los datos están ausentes[13]. Si los datos faltan completamente al azar (*missing completely at random*, MCAR), su ausencia no introduce sesgo, sino imprecisión en los resultados, y puede manejarse mediante técnicas simples. Si faltan al azar condicional (*missing at random*, MAR), es decir, su ausencia depende de otras variables observadas (por ejemplo, nivel socioeconómico ausente en pacientes más jóvenes), se requiere modelar dicha dependencia para minimizar el sesgo. Finalmente, si los datos están ausentes no al azar (*missing not at random*, MNAR), como en el caso de preguntas estigmatizantes que son evitadas por los pacientes, se requieren técnicas avanzadas y supuestos explícitos para abordar su impacto.

Existen múltiples estrategias para tratar los datos faltantes (**Figura 1**), cuya elección depende del tipo de ausencia, su proporción y el contexto clínico[14]. Las estrategias más sencillas, como eliminar registros incompletos, pueden resultar en una pérdida importante de información y reducir la potencia estadística del estudio. Esto puede ocurrir de manera inadvertida para el investigador durante el análisis de los datos, ya que muchos paquetes estadísticos realizan los cálculos considerando únicamente los registros completos, reduciendo así el tamaño de la muestra.

Las técnicas de imputación más básicas incluyen la sustitución de valores faltantes por la media, mediana o moda de la variable correspondiente. Estas técnicas son rápidas y fáciles de implementar, pero tienen la desventaja de subestimar la variabilidad y suponer implícitamente que los datos ausentes tienen características similares a los presentes, algo no que no siempre puede afirmarse con seguridad.

En contraposición, las técnicas avanzadas como la imputación por regresión, la imputación múltiple o el uso de algoritmos

Figura 1. Técnicas de imputación de datos faltantes



de aprendizaje automático (k-vecinos más cercanos, árboles de decisión, redes neuronales…) permiten capturar la complejidad de la relación entre variables y estimar los valores ausentes de forma más robusta. Estas técnicas pueden resultar computacionalmente exigentes, pero su uso es recomendable cuando los modelos se destinan a apoyar decisiones clínicas.

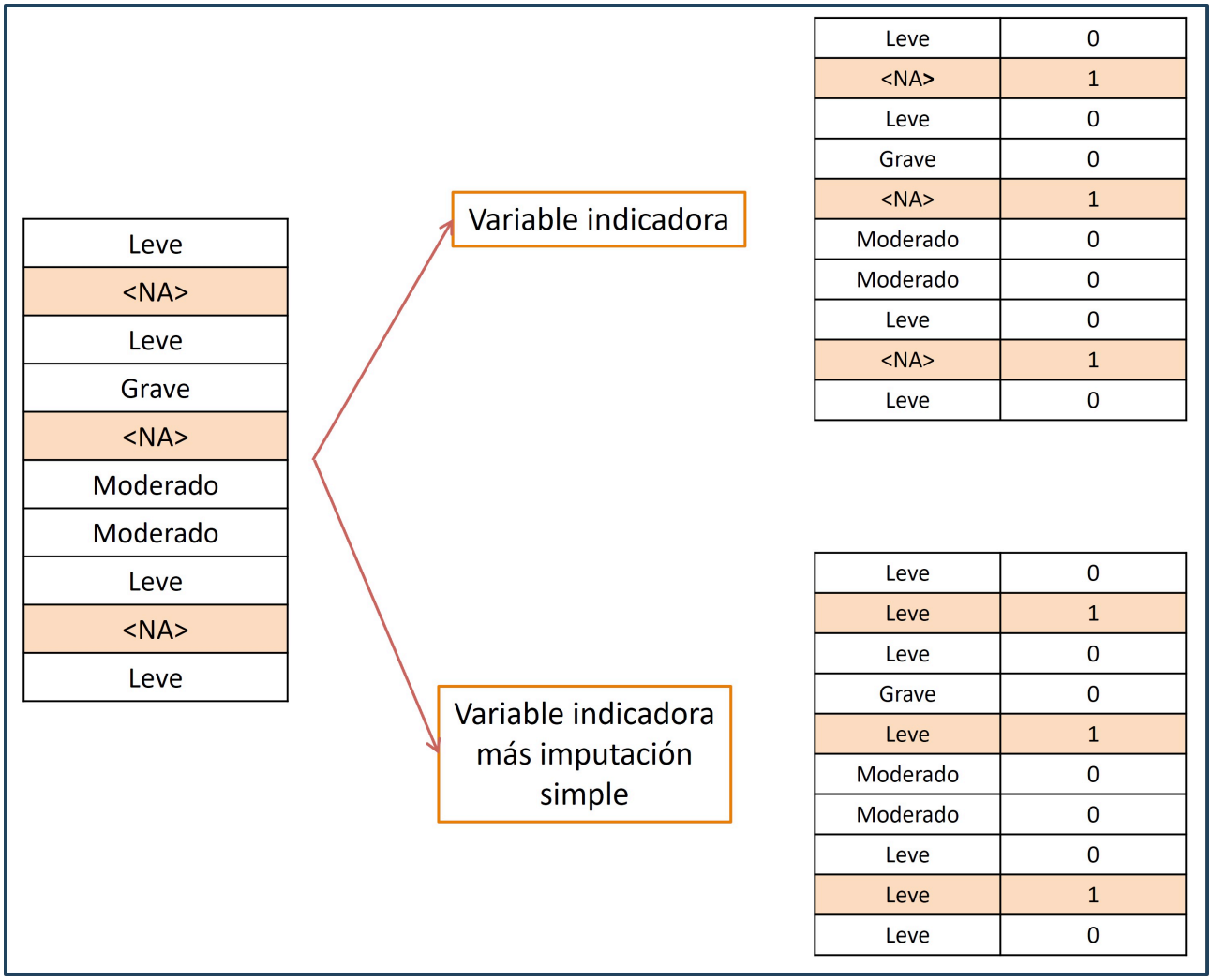
Una alternativa complementaria consiste en introducir una variable indicadora binaria que señale si un valor está ausente. Esto puede ser útil cuando la propia ausencia contiene información diagnóstica o pronóstica (por ejemplo, la falta de una medición de glucemia puede sugerir que no hay sospecha de diabetes). En algunas ocasiones, los modelos de aprendizaje automático pueden obtener más información de la ausencia o presencia del dato que del valor del dato imputado. Lógicamente, pueden combinarse las dos estrategias: imputar el dato y considerar su ausencia (**Figura 2**).

Estos métodos de imputación avanzados suelen requerir herramientas estadísticas especializadas para su correcta implementación. En la práctica, estos procedimientos se aplican habitualmente mediante paquetes específicos en lenguajes como R (por ejemplo, *mice* o *missForest*), o bien a través de plataformas estadísticas que permiten automatizar el proceso respetando los supuestos del análisis. Su uso adecuado exige conocimientos técnicos y una cuidadosa interpretación de los resultados imputados.

TRANSFORMACIÓN DE DATOS PARA EL ANÁLISIS

Una vez que los datos han sido limpiados e imputados, se procede a su transformación para ajustarlos a los requerimientos de los algoritmos analíticos[15]. Este paso incluye la normalización (escala los datos entre 0 y 1), la estandarización (centra los datos en media cero y desviación uno), y transformaciones logarítmicas para corregir asimetrías en las distribuciones.

Figura 2. Combinación de técnicas de imputación de datos faltantes



Esta etapa es especialmente relevante al combinar variables de distinta magnitud y naturaleza, como edad, presión arterial o niveles hormonales. Las transformaciones permiten garantizar que ninguna variable domine el análisis debido a su escala y mejoran el rendimiento de algoritmos sensibles como k-NN o redes neuronales.

CODIFICACIÓN DE VARIABLES CATEGÓRICAS

Los modelos matemáticos no pueden trabajar directamente con variables categóricas si no se transforman en representaciones numéricas. Este tipo de variables son frecuentes en nuestro entorno: sexo, tipo de vacuna, diagnóstico, nivel de gravedad… Las técnicas más comunes son la codificación por etiquetas y la codificación binaria (**Figura 3**)[16,17].

La codificación por etiquetas es bien conocida por su aplicación para el uso de modelos estadísticos más clásicos. Este método asigna un número entero único a cada categoría presente en la variable. Por ejemplo, una variable "tipo de alimentación" con categorías "lactancia materna", "lactancia artificial" y "mixta" podría codificarse como 0, 1 y 2, respectivamente.

Aunque es una técnica sencilla y eficiente en términos computacionales, la codificación por etiquetas introduce un orden numérico entre categorías que puede no reflejar ninguna jerarquía real. Esto puede generar problemas en modelos sensibles a la escala o al orden, como regresiones lineales o redes neuronales, donde el algoritmo podría interpretar erróneamente que una categoría "vale más" que otra por su valor numérico asignado. En entornos clínicos, donde muchas variables categóricas no implican relación de orden, es fundamental considerar si esta técnica es adecuada o si es preferible utilizar alternativas como la codificación binaria, que evitan este tipo de sesgos.

La codificación binaria, también conocida como *one-hot encoding*, es una técnica frecuentemente utilizada para modelos de aprendizaje automático. Consiste en crear una nueva variable para cada categoría distinta dentro de la variable original, de modo que cada observación recibe un valor de 1 en la nueva variable correspondiente a su categoría y 0 en las demás. Por ejemplo, si la variable "tipo de vacuna" incluye las categorías "triple vírica", "hexavalente" y "meningocócica", la codificación binaria generará tres nuevas columnas, una por cada tipo de vacuna, y marcará con un 1 la columna correspondiente a la categoría presente en cada registro. Este concepto se asemeja al de la creación de variables indicadoras (*dummies*) de los modelos estadísticos.

Esta técnica evita la introducción de un orden arbitrario entre categorías, lo que la hace especialmente útil cuando las categorías son nominales, es decir, sin jerarquía. Sin embargo, la codificación binaria también tiene inconvenientes. El más relevante es el aumento de la dimensionalidad del conjunto de

Figura 3. Técnicas de codificación de variables cualitativas

Codificación ordinal

Paciente	Intensidad
1	3
2	2
3	1
4	2
5	1
6	1
7	3
8	2

Codificación binaria

Paciente	Leve	Moderada	Grave
1	0	0	1
2	0	1	0
3	1	0	0
4	0	1	0
5	1	0	0
6	1	0	0
7	0	0	1
8	0	1	0

Intensidad: 1 leve, 2 moderada, 3 grave.

datos, especialmente cuando la variable original contiene muchas categorías. Esto puede dificultar la interpretación del modelo, incrementar el tiempo de procesamiento y aumentar el riesgo de sobreajuste, en particular en conjuntos de datos pequeños.

REDUCCIÓN DE DIMENSIONALIDAD

Los conjuntos de datos clínicos suelen contener muchas variables, algunas de las cuales pueden estar correlacionadas o no aportar información útil al análisis. La reducción de dimensionalidad, a través de técnicas como el análisis de componentes principales, permite transformar las variables originales en combinaciones lineales que concentran la mayor parte de la información. Esto reduce el tiempo de cálculo, mejora la interpretación y disminuye el riesgo de sobreajuste.

INGENIERÍA DE VARIABLES

La ingeniería de variables (ingeniería de características o *feature engineering*), consiste en crear, transformar o seleccionar variables que mejoren el rendimiento del modelo. Esto puede implicar la creación de nuevas variables combinadas, la transformación de variables existentes, o la selección de aquellas más relevantes según criterios estadísticos o clínicos. En Pediatría, donde la interacción entre variables como edad, peso, antecedentes o hábitos puede ser compleja, este paso puede resultar decisivo para construir modelos útiles y precisos. Además, dentro del ámbito del aprendizaje automático, constituye uno de los pilares del preprocesamiento necesario para optimizar el rendimiento de los futuros modelos.

Así, vemos que el preprocesamiento de los datos es una fase crítica e imprescindible de los proyectos de análisis clínico y elaboración de modelos predictivos[18]. En medicina en general, y en Pediatría en particular, esta etapa cobra una importancia aún mayor por la complejidad de los datos, la variabilidad de los pacientes y las implicaciones clínicas de los resultados. Una correcta manipulación de los datos no solo mejora la calidad del análisis, sino que constituye un acto de responsabilidad clínica y ética hacia los pacientes que nos confían su información.

BIBLIOGRAFÍA

1. Vesoulis ZA, Husain AN, Cole FS. Improving child health through Big Data and data science. Pediatr Res. 2022;93:342-9.
2. Halwani MA, Merdad G, Almasre M, Doman G, AlSharif S, Alshiakh SM, *et al*. Predicting triage of pediatric patients in the emergency department using machine learning approach. Int J Emerg Med. 2025;18:51.
3. Press J. Cleaning Big Data: most time-consuming, least enjoyable data science task, survey says. En: Forbes [en línea] [consultado el 23/09/2025]. Disponible en www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?utm_source=chatgpt.com
4. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: Implications for clinical decision-making. PLOS Digit Health. 2024;3:e0000651.
5. Ortiz BL, Gupta V, Kumar R, Jalin A, Cao X, Ziegenbein C, *et al*. Data preprocessing techniques for ai and machine learning readiness: scoping review of wearable sensor data in cancer care. JMIR Mhealth Uhealth. 2024;12:e59587.
6. Mohsen F, Ali H, El Hajj N, Shah Z. Artificial intelligence-based methods for fusion of electronic health records and imaging data. Sci Rep. 2022;12(1):17981.
7. Wei WQ, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, *et al.* Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. J Am Med Inform Assoc. 2012;19:219-24.
8. Ferrão JC, Duarte Oliveira M, Janela F, Martins HMG. Preprocessing structured clinical data for predictive modeling and decision support. Appl Clin Inform. 2016;7:1135-53.
9. Nan Y, Del Ser J, Walsh S, Schönlieb C, Roberts M, Selby I, *et al*. Data harmonisation for information fusion in digital healthcare: a state-of-the-art systematic review, meta-analysis and future research directions. Inf Fusion. 2022;82:99-122.
10. Pilowsky JK, Elliott R, Roche MA. Data cleaning for clinician researchers: Application and explanation of a data-quality framework. Aust Crit Care. 2024;37:827-33.
11. Exploring data. En: Zumel N, Mount J (eds.). Practical data science with R, 2ª ed. Shelter Island, NY: Manning Publications Co. 2020:51-87 [en línea] [consultado el 23/09/2025]. Disponible en https://mathematics.foi.hr/Rprojekti/knjige/Practical_dAta_Science_with_R.pdf
12. Liu M, Li S, Yuan H, Ong MEH, Ning Y, Xie F, *et al.* Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. Artif Intell Med. 2023;142:102587.
13. Afkanpour M, Hosseinzadeh E, Tabesh H. Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review. BMC Med Res Methodol. 2024;24:188.
14. Luo Y. Evaluating the state of the art in missing data imputation for clinical data. Brief Bioinform. 2022;23:bbab489.
15. Maharana K, Mondal S, Nemade B. A review: data preprocessing and data augmentation techniques. Glob Transit Proc. 2022;3:91-9.
16. Nicoomanesh A. Categorical features encoding advanced techniques: an overview. En: Medium [en línea] [consultado el 23/09/2025]. Disponible en https://medium.com/%40anicomanesh/categorical-encoding-methods-an-overview-ce90eec3cf86

17. Menear K. Comparing label encoding, one-hot encoding, and binary encoding for handling categorical variables in machine learning. En: Medium [en línea] [consultado el 23/09/2025]. Disponible en https://medium.com/%40kevin.menear/comparing-label-encoding-one-hot-encoding-and-binary-encoding-for-handling-categorical-variables-933544ccbd02

18. Huang HN, Chen HM, Lin WW, Huang CJ, Chen YC, Wang YH, *et al*. Employing feature engineering strategies to improve the performance of machine learning algorithms on echocardiogram dataset. Digit Health. 2023;9:20552076231207589.