ISSN-e: 2542-3401
Período: julio–septiembre, 2026

Universidad, Ciencia y Tecnología
Vol. 30, Núm. 132 (pp. 126–137)

Artículo de investigación

https://doi.org/10.47460/uct.v30i132.1113

Modelo predictivo basado en inteligencia artificial para la identificación temprana de complicaciones posquirúrgicas mediante indicadores clínicos y funcionales

Doris Elizabeth Zelada Chavarry*
https://orcid.org/0000-0003-0094-1380
dzeladach@gmail.com
Universidad Nacional de Cajamarca
Cajamarca–Perú

Ethel Paola Gonzalez Esparza
https://orcid.org/0009-0004-2075-3699
egonzales@unc.edu.pe
Universidad Nacional de Cajamarca
Cajamarca–Perú

Carlos Alexis Pajares Wong
https://orcid.org/0009-0003-6811-1604
solracpw@gmail.com
Universidad Nacional de Cajamarca
Cajamarca–Perú

Wilson Edgardo León Vilca
https://orcid.org/0000-0001-9560-3821
wilsonleon@unc.edu.pe
Universidad Nacional de Cajamarca
Cajamarca–Perú

**Autor de correspondencia: dzeladach@gmail.com*

Recibido: (07/05/2026), Aceptado: (23/07/2026)

Resumen. Se desarrolló un modelo predictivo basado en inteligencia artificial para identificar tempranamente complicaciones posquirúrgicas mediante indicadores clínicos y funcionales. Se analizaron 10.000 registros sintéticos del conjunto TERMINET mediante regresión logística, *Random Forest* y XGBoost, utilizando partición estratificada, validación cruzada y métricas de discriminación y calibración. Los modelos mostraron capacidad predictiva excepcional, con clasificación perfecta para regresión logística y *Random Forest*. El análisis SHAP identificó sodio, nitrógeno ureico, creatinina y parámetros relacionados con INR entre los principales predictores. Los resultados demuestran viabilidad computacional, aunque su elevado desempeño requiere interpretación cautelosa y validación externa posterior utilizando cohortes clínicas reales e independientes.

Palabras clave: inteligencia artificial, complicaciones posquirúrgicas, aprendizaje automático, predicción clínica.

Artificial Intelligence-Based Predictive Model for Early Identification of Postoperative Complications Using Clinical and Functional Indicators

Abstract. An artificial intelligence-based predictive model was developed for the early identification of postoperative complications using clinical and functional indicators. A total of 10,000 synthetic records from the TERMINET dataset were analyzed using logistic regression, Random Forest, and XGBoost, with stratified data splitting, cross-validation, and discrimination and calibration metrics. The models demonstrated exceptional predictive performance, with perfect classification achieved by logistic regression and Random Forest. SHAP analysis identified sodium, urea nitrogen, creatinine, and INR-related parameters among the main predictors. The findings demonstrate the computational feasibility of the proposed approach; however, the exceptionally high performance warrants cautious interpretation and subsequent external validation using real and independent clinical cohorts.

Keywords: artificial intelligence, postoperative complications, machine learning, clinical prediction.

I. INTRODUCCIÓN

Las complicaciones posquirúrgicas constituyen un problema relevante para la atención perioperatoria debido a su asociación con mayor morbimortalidad, prolongación de la estancia hospitalaria e incremento de los requerimientos asistenciales y costos sanitarios. Su identificación anticipada puede favorecer la vigilancia y adopción de medidas preventivas en pacientes con mayor riesgo de eventos adversos. Sin embargo, dicho riesgo depende de múltiples factores demográficos, funcionales, clínicos y fisiológicos cuya interacción resulta difícil de representar mediante aproximaciones convencionales [1], [2], [3]. En este contexto, la inteligencia artificial (IA) y los algoritmos de aprendizaje automático permiten procesar simultáneamente múltiples características e identificar relaciones complejas entre predictores y desenlaces clínicos [4], [5], con aplicaciones que abarcan desde la evaluación preoperatoria hasta la vigilancia posoperatoria [6], [7].

La evidencia disponible respalda la aplicación del aprendizaje automático en la predicción de complicaciones posoperatorias. Xue et al. [1] demostraron, mediante el análisis de 111 888 procedimientos, que la combinación de información preoperatoria e intraoperatoria permitía anticipar diferentes eventos adversos. Asimismo, se han desarrollado modelos predictivos en cirugía gastrointestinal [2], grandes procedimientos quirúrgicos [3] y otros contextos clínicos mediante algoritmos tradicionales y arquitecturas de aprendizaje profundo [8]. No obstante, el desempeño discriminativo no constituye el único criterio de utilidad de estos modelos. La incorporación de estrategias de IA explicable permite determinar la contribución de los predictores al riesgo estimado [9], [10], [11], mientras que la validación, generalización e interpretabilidad continúan siendo requisitos fundamentales antes de trasladar estos sistemas a escenarios asistenciales [12], [13], [14].

El acceso restringido a grandes conjuntos de datos clínicos representa una dificultad adicional para el desarrollo reproducible de modelos predictivos. Los datos sintéticos constituyen una alternativa para la experimentación computacional y la preservación de la privacidad, siempre que su procedencia sea transparente y sus resultados no se interpreten como evidencia de validación clínica. El conjunto *TERMINET eHealth post-operation complications synthetic dataset* fue desarrollado a partir de información clínica del estudio SUPERO y contiene 10 000 registros sintéticos con características demográficas, clínicas, funcionales y conductuales, además del desenlace relacionado con la presencia de complicaciones posteriores a la intervención [15]. Estas características permiten evaluar algoritmos predictivos en un entorno controlado y reproducible.

En función de lo anterior, el objetivo de esta investigación fue desarrollar y evaluar un modelo predictivo basado en inteligencia artificial para la identificación temprana de complicaciones posquirúrgicas mediante indicadores clínicos y funcionales, utilizando el conjunto de datos sintéticos TERMINET. Se compararon algoritmos de aprendizaje automático considerando su capacidad discriminativa, sensibilidad, especificidad, precisión y calibración, complementando el análisis mediante técnicas de interpretabilidad para identificar las variables con mayor contribución predictiva. El estudio se concibió como una prueba de concepto computacional, por lo que sus resultados buscan establecer la viabilidad del enfoque y proporcionar una base metodológica para posteriores procesos de validación mediante cohortes clínicas independientes.

II. MARCO TEÓRICO

A. Riesgo y complicaciones posquirúrgicas

El riesgo posquirúrgico puede entenderse como la probabilidad de aparición de eventos adversos posteriores a una intervención, determinada por la interacción entre las características basales del paciente, su estado funcional, las condiciones clínicas y las características del procedimiento. La literatura ha demostrado que estos desenlaces no responden necesariamente a factores independientes, sino a configuraciones multidimensionales cuya identificación puede requerir el análisis simultáneo de numerosas variables [1], [3].

Los sistemas tradicionales de estratificación de riesgo ofrecen herramientas útiles para la práctica clínica, pero suelen depender de un conjunto limitado de variables y de relaciones previamente especificadas. Los métodos de aprendizaje automático amplían este enfoque al permitir la exploración de asociaciones no lineales e interacciones entre predictores sin exigir que todas las relaciones funcionales

sean definidas de antemano [4]. Esta característica ha favorecido su utilización para estimar complicaciones en diferentes poblaciones quirúrgicas y para complementar los mecanismos convencionales de valoración del riesgo [5], [13].

La identificación temprana adquiere particular importancia cuando los factores utilizados para estimar el riesgo están disponibles antes de la intervención. En tales circunstancias, la predicción no pretende diagnosticar una complicación ya desarrollada, sino reconocer configuraciones clínicas y funcionales asociadas con una mayor probabilidad de aparición posterior. Esta distinción resulta esencial para evitar fuga de información entre predictores y desenlace y para preservar la utilidad prospectiva del modelo.

B. Inteligencia artificial y aprendizaje automático en la predicción quirúrgica

La IA aplicada a cirugía comprende sistemas capaces de procesar información clínica para apoyar diferentes componentes de la atención quirúrgica. Dentro de este campo, el aprendizaje automático permite construir funciones predictivas a partir de patrones presentes en los datos y ha adquirido creciente relevancia en las etapas preoperatoria, intraoperatoria y posoperatoria [6], [7].

Entre los algoritmos utilizados para problemas de clasificación clínica se encuentran la regresión logística como referencia estadística, árboles de decisión, *Random Forest*, métodos de *gradient boosting* y redes neuronales. La comparación entre diferentes familias de algoritmos permite determinar si el incremento de complejidad proporciona ventajas reales frente a modelos más parsimoniosos. Estudios previos han mostrado que los métodos basados en árboles potenciados y otros algoritmos de aprendizaje automático pueden alcanzar una capacidad discriminativa elevada para diferentes complicaciones posoperatorias [1], mientras que las redes neuronales profundas también han sido evaluadas para este propósito [8].

La selección del algoritmo no debe sustentarse exclusivamente en la exactitud global. En problemas clínicos donde la frecuencia del desenlace puede ser considerablemente inferior a la ausencia de eventos, esta medida puede producir una percepción incompleta del rendimiento. Por ello, la evaluación debe considerar simultáneamente capacidad discriminativa, sensibilidad, especificidad, precisión y equilibrio entre clases, además de estudiar la calibración de las probabilidades estimadas. Las revisiones existentes sobre IA aplicada a predicción quirúrgica han destacado precisamente la heterogeneidad metodológica y la importancia de procedimientos de validación apropiados [12], [13], [14].

C. Interpretabilidad de los modelos predictivos

La creciente complejidad de algunos algoritmos genera una tensión entre capacidad predictiva e interpretabilidad. Mientras que modelos relativamente simples permiten reconocer directamente la dirección y magnitud de las asociaciones, métodos como *Random Forest*, *gradient boosting* o redes neuronales presentan estructuras internas cuya interpretación resulta menos inmediata. Esta característica adquiere especial relevancia en aplicaciones sanitarias, donde conocer las variables que sustentan una predicción puede ser tan importante como conocer la clasificación final.

Las técnicas de inteligencia artificial explicable permiten analizar esta dimensión. Entre ellas, SHAP (*SHapley Additive exPlanations*) proporciona estimaciones de la contribución de cada característica a una determinada predicción. Su aplicación en modelos de complicaciones posoperatorias ha permitido identificar los factores con mayor influencia global y explicar predicciones individuales [1], [9], [10], [11].

No obstante, la interpretabilidad no convierte automáticamente una asociación predictiva en una relación causal. La importancia asignada por el modelo expresa la contribución de una variable al comportamiento predictivo del algoritmo dentro del conjunto analizado. Por ello, los resultados de explicabilidad deben interpretarse como mecanismos para comprender el funcionamiento del modelo y generar hipótesis clínicas, no como demostración independiente de causalidad.

D. Indicadores clínicos, funcionales y conductuales en la estimación del riesgo

El estado preoperatorio del paciente constituye una fuente multidimensional de información potencialmente relevante para la estimación del riesgo. Además de variables demográficas y parámetros clínicos, la capacidad funcional, movilidad, estado cognitivo, nutrición, polifarmacia y actividad física pueden reflejar dimensiones de vulnerabilidad que no quedan completamente representadas mediante un único indicador fisiológico.

Esta perspectiva resulta particularmente pertinente en poblaciones de adultos mayores sometidos a cirugía. TERMINET integra variables demográficas con información clínica, cuestionarios funcionales y estadísticas derivadas de actividad física registrada durante las dos semanas previas a la hospitalización [15]. En concreto, el repositorio describe seis características derivadas del número de pasos, veinte atributos clínicos, dos demográficos, doce atributos procedentes de cuestionarios y una variable de desenlace. Los datos originales proceden de pacientes participantes en el estudio SUPERO y combinan información clínica hospitalaria con información conductual recopilada mediante una plataforma digital [15].

La integración de estos dominios posibilita plantear el riesgo posquirúrgico como un problema de clasificación multivariable. En lugar de atribuir el desenlace a un único marcador, el modelo puede explorar configuraciones simultáneas de estado funcional, actividad, características demográficas y parámetros clínicos. Desde la perspectiva del aprendizaje automático, esta heterogeneidad constituye una oportunidad para evaluar qué combinación de características aporta mayor información predictiva.

E. Datos sintéticos y validación de modelos en salud

El empleo de datos sintéticos permite desarrollar experimentos computacionales en escenarios donde el acceso a información clínica individual se encuentra restringido por razones éticas, regulatorias o de privacidad. Sin embargo, los registros sintéticos no deben considerarse equivalentes a observaciones independientes procedentes de pacientes reales. Su utilidad depende del procedimiento mediante el cual fueron generados y del grado en que preservan las propiedades relevantes del conjunto de origen.

Esta consideración es particularmente importante en TERMINET. El repositorio informa que la generación del conjunto partió de datos reales de un número reducido de pacientes; posteriormente se crearon vectores adicionales mediante incorporación de ruido gaussiano, se realizó agrupamiento aglomerativo y cada grupo fue modelado mediante modelos de mezcla gaussiana, a partir de los cuales se generaron finalmente los 10 000 vectores sintéticos [15]. Por consiguiente, el tamaño nominal del conjunto proporciona capacidad para experimentación algorítmica, pero no equivale a evidencia clínica obtenida de una cohorte independiente de 10 000 pacientes.

Esta característica determina el alcance de cualquier modelo desarrollado con estos datos. Los resultados pueden utilizarse para comparar algoritmos, analizar patrones predictivos, estudiar la contribución de las características y demostrar la factibilidad computacional del enfoque, pero la generalización clínica requiere validación posterior mediante datos observacionales independientes. Este principio coincide con la literatura reciente, que identifica la validación externa, reproducibilidad, interpretabilidad y evaluación prospectiva como elementos necesarios para avanzar desde modelos experimentales hacia herramientas de apoyo clínico [6], [12].

En consecuencia, el modelo propuesto se conceptualiza como una prueba de concepto predictiva, en la cual los datos sintéticos permiten desarrollar y evaluar de manera reproducible una arquitectura de aprendizaje automático orientada a la identificación temprana de complicaciones posquirúrgicas. La contribución del estudio reside tanto en determinar el rendimiento comparativo de los algoritmos como en establecer cuáles dimensiones clínicas y funcionales concentran mayor información predictiva, manteniendo explícita la necesidad de validación posterior en poblaciones quirúrgicas reales.

III. METODOLOGÍA

A. Diseño del estudio y fuente de datos

Se desarrolló un estudio cuantitativo, predictivo y de carácter computacional, orientado al desarrollo y evaluación de modelos de aprendizaje automático para la identificación de complicaciones posquirúrgicas. El análisis se realizó utilizando el conjunto público *TERMINET eHealth post-operation complications synthetic dataset*, desarrollado por Lekka et al. [15] y disponible en el repositorio Zenodo. El conjunto fue generado con fines de investigación a partir de información clínica original del estudio SUPERO y contiene 10 000 registros sintéticos que reproducen características demográficas, clínicas, funcionales y conductuales relacionadas con pacientes sometidos a procedimientos quirúrgicos.

La unidad de análisis correspondió a cada registro sintético contenido en la base. No se realizó selección muestral adicional, debido a que se utilizaron los 10 000 registros disponibles. La variable

de resultado fue *Complications*, codificada de forma binaria como ausencia (0) o presencia (1) de complicaciones posquirúrgicas. El conjunto presentó 7570 registros sin complicaciones (75,7 %) y 2430 con complicaciones (24,3 %). Dado el carácter sintético de los datos, el estudio se planteó como una prueba de concepto computacional y no como una validación clínica del modelo. En consecuencia, las estimaciones obtenidas se interpretaron exclusivamente en términos de capacidad predictiva dentro de las propiedades representadas por el conjunto TERMINET.

B. Variables predictoras

Se consideraron como potenciales predictores las variables disponibles antes del desenlace y pertenecientes a cuatro dominios: características demográficas, indicadores funcionales, actividad física y parámetros clínicos. Entre ellas se incluyeron edad y sexo; componentes de evaluación funcional relacionados con nutrición, movilidad, estado neuropsicológico, índice de masa corporal, consumo de medicamentos y percepción de salud; indicadores derivados de SPPB y Mini-Cog; medidas de actividad física basadas en el número de pasos; y parámetros clínicos y de laboratorio, entre ellos hemoglobina, hematocrito, leucocitos, neutrófilos, linfocitos, plaquetas, sodio, potasio, creatinina, bilirrubina, ALT, AST, INR y nitrógeno ureico.

Antes del modelado se verificó la naturaleza y codificación de cada variable, su rango de valores, distribución, frecuencia de categorías y presencia de datos ausentes. También se examinaron variables constantes o con variabilidad insuficiente y posibles redundancias entre predictores. La variable *Complications* se excluyó de la matriz de características y se utilizó únicamente como variable objetivo. Para prevenir fuga de información (*data leakage*), se excluyó cualquier característica cuya determinación dependiera temporal o conceptualmente de la aparición de la complicación. Las decisiones de exclusión se realizaron antes del entrenamiento de los algoritmos y se documentaron para garantizar la reproducibilidad del análisis.

C. Preparación y partición de los datos

Las variables categóricas se transformaron a una representación numérica apropiada cuando fue necesario. Las variables continuas utilizadas por la regresión logística se estandarizaron utilizando la media y desviación estándar calculadas exclusivamente sobre los datos de entrenamiento. Los modelos basados en árboles se entrenaron con las variables en su escala original, dado que su funcionamiento no requiere estandarización. El conjunto completo se dividió mediante muestreo aleatorio estratificado en 80 % para entrenamiento y 20 % para prueba, preservando aproximadamente la proporción original de registros con y sin complicaciones. El conjunto de prueba permaneció completamente separado durante las etapas de selección y ajuste de los modelos y se utilizó una única vez para la evaluación final.

Debido a la distribución de la variable objetivo, no se realizó sobremuestreo sintético de manera predeterminada. El desbalance se abordó mediante ponderación de clases en los algoritmos que permitieron esta configuración. Esta decisión evitó introducir observaciones artificiales adicionales en un conjunto que ya posee naturaleza sintética.

D. Desarrollo de los modelos predictivos

Se evaluaron tres algoritmos de clasificación supervisada: regresión logística, *Random Forest* y XGBoost. La regresión logística se utilizó como modelo de referencia debido a su interpretabilidad y amplia utilización en predicción clínica. *Random Forest* permitió representar interacciones y relaciones no lineales mediante múltiples árboles de decisión, mientras que XGBoost se incorporó como algoritmo de potenciación basado en árboles por su capacidad para modelar estructuras predictivas complejas y su utilización previa en problemas de riesgo clínico.

El ajuste de hiperparámetros se realizó exclusivamente sobre el conjunto de entrenamiento mediante búsqueda en una cuadrícula acotada de valores (*grid search*) y validación cruzada estratificada de cinco pliegues. Para *Random Forest* se ajustaron el número de árboles, profundidad máxima y número mínimo de observaciones requerido para dividir nodos. Para XGBoost se consideraron el número de estimadores, profundidad máxima, tasa de aprendizaje y proporción de observaciones utilizada durante el entrenamiento. En la regresión logística se evaluó la intensidad de regularización. El criterio principal para seleccionar la configuración de cada algoritmo durante la validación cruzada fue el área bajo la curva ROC (ROC-AUC). Finalizado el ajuste, cada modelo fue reentrenado utilizando la totalidad del conjunto de entrenamiento con los hiperparámetros seleccionados.

E. Evaluación del desempeño predictivo

Los tres modelos finales se evaluaron sobre el conjunto de prueba independiente. Se calcularon ROC-AUC, área bajo la curva precisión-recall (PR-AUC), sensibilidad, especificidad, precisión, valor predictivo negativo, F1-score y exactitud. La sensibilidad se consideró una métrica de particular interés debido al objetivo de identificar oportunamente registros con riesgo de complicaciones. La especificidad permitió evaluar la capacidad para reconocer correctamente los casos sin complicaciones, mientras que PR-AUC y F1-score complementaron la evaluación frente al desbalance existente entre las clases.

Para analizar la calidad de las probabilidades generadas se evaluó además la calibración mediante el *Brier score* y una curva de calibración. La comparación integral entre discriminación, sensibilidad, comportamiento frente al desbalance y calibración permitió seleccionar el modelo con mejor desempeño global, evitando establecer la superioridad exclusivamente a partir de la exactitud.

F. Interpretabilidad del modelo

Una vez identificado el algoritmo con mejor desempeño, se analizó la contribución de las variables mediante SHAP (*SHapley Additive exPlanations*). Se estimaron valores SHAP para determinar la importancia global de los predictores y la dirección de su contribución a las probabilidades estimadas. El análisis de interpretabilidad se utilizó para identificar las características clínicas y funcionales con mayor influencia en las predicciones y comprender el comportamiento interno del modelo. Las contribuciones SHAP fueron interpretadas como asociaciones predictivas dentro del modelo y no como estimaciones de efectos causales.

G. Análisis estadístico y entorno computacional

El procesamiento y análisis de los datos se realizó en Python. Para la manipulación y preparación de la información se utilizaron *pandas* y NumPy; el preprocesamiento, validación y evaluación de los modelos se realizó mediante *scikit-learn*; XGBoost se implementó mediante la biblioteca *xgboost*; y el análisis de interpretabilidad mediante *shap*. Las variables continuas se describieron mediante media y desviación estándar o mediana y rango intercuartílico, de acuerdo con su distribución. Las variables categóricas se expresaron mediante frecuencias absolutas y relativas. Se estableció una semilla aleatoria fija para las operaciones de partición, validación y entrenamiento que involucraron aleatoriedad, con el propósito de favorecer la reproducibilidad computacional.

H. Consideraciones éticas y reproducibilidad

El estudio utilizó exclusivamente un conjunto de datos sintéticos de acceso público y no implicó contacto con participantes, intervención clínica ni acceso a información personal identificable. La generación y procedencia del conjunto TERMINET fueron documentadas por Lekka et al. [15]. Por esta razón, el presente análisis no implicó recolección primaria de información clínica. Los resultados se interpretaron considerando explícitamente la naturaleza sintética de los registros. En particular, el desempeño obtenido no se consideró equivalente a una validación externa en pacientes reales ni suficiente para recomendar la utilización clínica del modelo. La metodología, criterios de preparación, partición de datos, algoritmos y métricas se definieron de forma reproducible para permitir su posterior aplicación y validación en cohortes clínicas independientes.

IV. RESULTADOS

A. Caracterización del conjunto de datos

El análisis incluyó los 10 000 registros sintéticos disponibles en TERMINET y 39 variables predictoras, además de la variable de desenlace (Tabla 1). No se identificaron valores ausentes ni registros duplicados. La variable *Complications* presentó 7570 registros sin complicaciones (75,7 %) y 2430 con complicaciones (24,3 %), lo que evidencia un desbalance moderado entre las clases. La partición estratificada destinó 8000 registros al entrenamiento y 2000 a la prueba independiente; esta última quedó constituida por 1514 registros negativos y 486 positivos.

Tabla 1. Distribución de la variable de resultado

Resultado	**n**	%
Sin complicaciones	7570	75,7
Con complicaciones	2430	24,3
Total	10 000	100,0

B. Asociación preliminar de los predictores con las complicaciones

La exploración bivariada mostró una separación marcada entre las clases para varios indicadores clínicos (Tabla 2). Las mayores correlaciones absolutas con el desenlace correspondieron a nitrógeno ureico ($r = 0,744$), INR ($r = -0,729$), segundos de INR ($r = -0,725$), sodio ($r = 0,721$), porcentaje de INR ($r = 0,709$), creatinina ($r = 0,582$), MiniCog ($r = -0,570$) y plaquetas ($r = -0,455$). Estas asociaciones se utilizaron con finalidad descriptiva y no se interpretaron como relaciones causales.

Tabla 2. Predictores con mayor asociación lineal con *Complications*

Variable	r
`Urea_Nitrogen`	0,744
INR	-0,729
`INR_Seconds`	-0,725
`Sodium`	0,721
`INR_Perc`	0,709
`Creatinine`	0,582
`MiniCog`	-0,570
`Platelets`	-0,455

C. Desempeño comparativo de los modelos

Los modelos mostraron una capacidad discriminativa excepcionalmente alta en el conjunto de prueba (Tabla 3). La regresión logística y *Random Forest* clasificaron correctamente los 2000 registros, mientras que XGBoost produjo un único falso positivo y ningún falso negativo. Los tres algoritmos alcanzaron ROC-AUC y PR-AUC redondeadas a 1,000. La diferencia más visible apareció en calibración: la regresión logística obtuvo el menor *Brier score*, seguida de *Random Forest* y XGBoost.

Tabla 3. Rendimiento de los modelos en el conjunto de prueba

Modelo	**ROC-AUC**	**PR-AUC**	**Sens.**	**Espec.**	**Prec.**	**F1**	**Exact.**	**Brier**
Regresión logística	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000
Random Forest	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,0000
XGBoost	1,0000	1,0000	1,0000	0,9993	0,9979	0,9990	0,9995	0,0002

La capacidad discriminativa de los modelos se examinó mediante las curvas ROC obtenidas sobre el conjunto de prueba. Esta representación permite comparar gráficamente la relación entre sensibilidad y 1-especificidad para los tres algoritmos evaluados y observar su comportamiento frente a la línea de referencia. La Figura 1 presenta las curvas correspondientes a la regresión logística, *Random Forest* y XGBoost, cuyos desempeños mostraron una elevada superposición debido a los valores de AUC obtenidos.

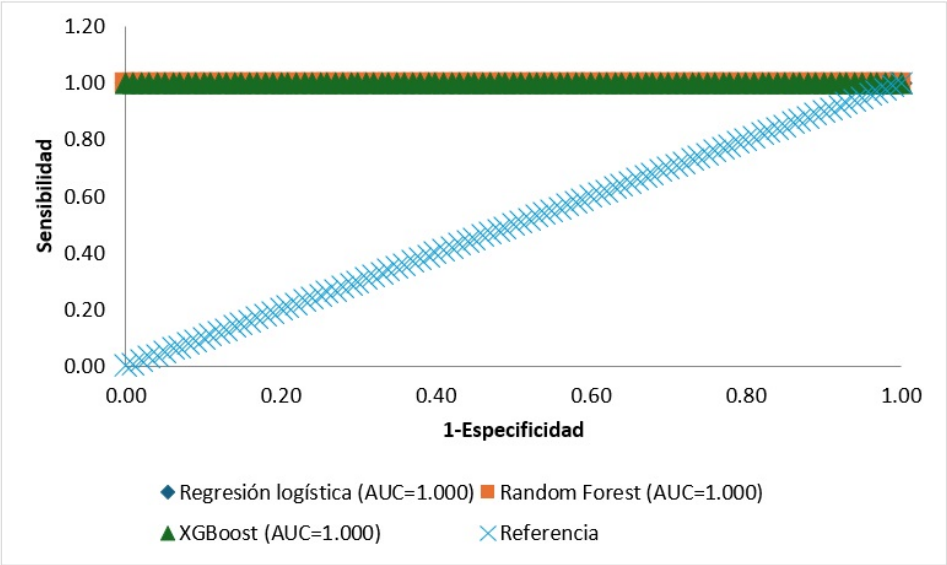

Fig. 1. Curvas ROC de los tres modelos en el conjunto de prueba.

D. Interpretabilidad y contribución de los predictores

Dado que *Random Forest* mantuvo clasificación perfecta en la prueba y permite obtener una medida directa de importancia de características, se utilizó como modelo no lineal de referencia para examinar la contribución global de los predictores (Figura 2). El sodio presentó la mayor importancia relativa (0,212), seguido de creatinina (0,113), nitrógeno ureico (0,097), segundos de INR (0,070), INR (0,062), porcentaje de INR (0,048), hemoglobina (0,047), plaquetas (0,045), linfocitos (0,039) y edad (0,035). La concentración de importancia en un grupo reducido de variables fue consistente con la fuerte separación observada durante la exploración bivariada.

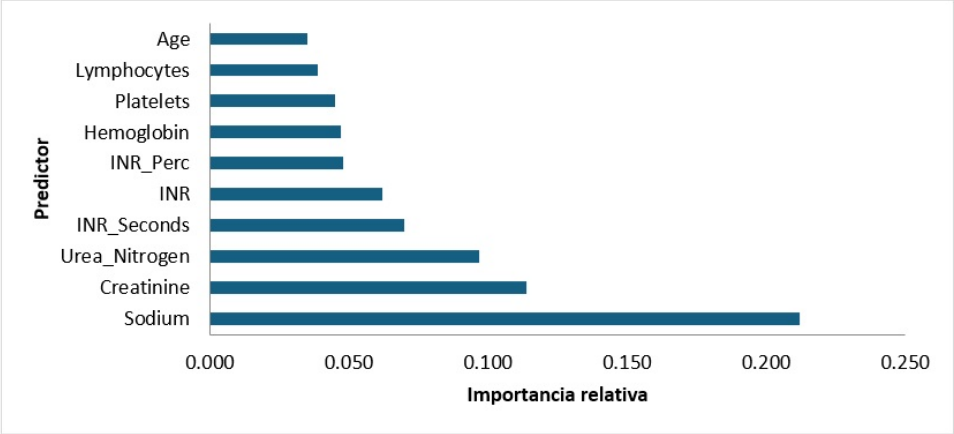

Fig. 2. Diez variables con mayor importancia relativa en *Random Forest*.

E. Separabilidad del conjunto sintético

El desempeño casi perfecto observado en los tres algoritmos indica que las clases del conjunto TERMINET son altamente separables en el espacio de características. Este comportamiento no debe interpretarse como evidencia de una capacidad clínica equivalente en pacientes reales. En particular, la obtención de resultados prácticamente idénticos mediante un modelo lineal y dos métodos basados en árboles sugiere que una parte sustancial de la discriminación se encuentra incorporada en la estructura estadística del conjunto sintético. Por ello, los resultados representan validación interna computacional y una prueba de concepto del enfoque predictivo; su generalización requiere evaluación posterior en cohortes clínicas independientes.

En términos operativos, el hallazgo central no fue la superioridad de un algoritmo complejo sobre otro, sino la existencia de un patrón multivariable suficientemente definido para separar los registros con y sin complicaciones dentro del entorno sintético. La ausencia de falsos negativos en los tres modelos resulta favorable para el objetivo de identificación temprana, pero debe interpretarse con cautela debido a la naturaleza del conjunto utilizado.

F. Calibración de las probabilidades predichas

Además de la capacidad discriminativa, se examinó la correspondencia entre las probabilidades estimadas por los modelos y la clasificación observada en el conjunto de prueba. Los valores del *Brier score* fueron muy bajos: 0,000005 para la regresión logística y 0,000014 para *Random Forest*, lo que indicó una elevada concordancia entre las probabilidades predichas y el desenlace registrado. Las probabilidades mostraron, además, una marcada separación entre las clases. En la regresión logística, los registros sin complicaciones presentaron probabilidades estimadas entre 0,00000002 y 0,0387, mientras que en los registros con complicaciones oscilaron entre 0,9834 y 0,999998. En *Random Forest*, los intervalos correspondientes fueron 0,000–0,038 y 0,950–1,000, respectivamente.

La ausencia de solapamiento entre los intervalos probabilísticos explica tanto los valores reducidos del *Brier score* como el comportamiento prácticamente perfecto observado en las curvas ROC. Sin embargo, esta calibración debe interpretarse dentro del conjunto sintético analizado y no como evidencia de calibración clínica externa. La Figura 3 presenta la relación entre las probabilidades predichas y la frecuencia observada de complicaciones. La proximidad al comportamiento ideal debe valorarse conjuntamente con la elevada separabilidad de las clases y con el origen sintético de los registros.

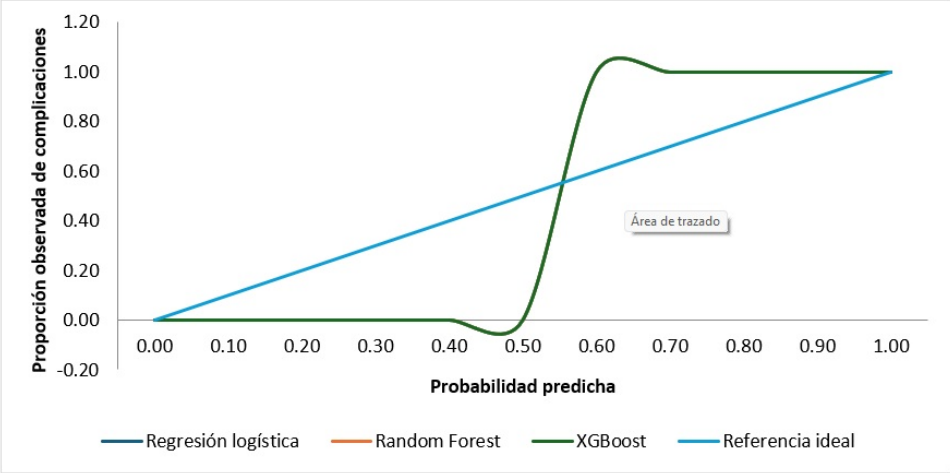

Fig. 3. Curvas de calibración de los modelos predictivos en el conjunto de prueba.

G. Interpretabilidad del modelo mediante SHAP

Para complementar la importancia relativa obtenida mediante *Random Forest*, se aplicó el método SHAP con el propósito de cuantificar la contribución de cada predictor a las estimaciones individuales. El análisis confirmó que la predicción estuvo concentrada principalmente en un grupo reducido de indicadores clínicos. Las mayores magnitudes medias absolutas de los valores SHAP correspondieron a `Sodium` (0,1153), `Urea_Nitrogen` (0,0549), `Creatinine` (0,0478), `INR_Seconds` (0,0380), `Hemoglobin` (0,0288), INR (0,0283), `INR_Perc` (0,0254), `Platelets` (0,0238), `Bilirubin` (0,0223) y `Lymphocytes` (0,0217).

El análisis de la dirección de las contribuciones mostró que valores mayores de `Urea_Nitrogen` y `Creatinine` tendieron a desplazar la predicción hacia la clase correspondiente a complicaciones, mientras que valores mayores de INR, `INR_Seconds`, `Platelets`, `Bilirubin` y `Lymphocytes` mostraron predominantemente contribuciones en sentido contrario. `Sodium` presentó la mayor contribución global al comportamiento del modelo. Estos patrones representan relaciones predictivas internas del algoritmo y no deben interpretarse como efectos causales de los indicadores sobre la aparición de complicaciones.

La Figura 4 permite visualizar simultáneamente la magnitud y dirección de las contribuciones de las características sobre la predicción. Los resultados corroboran que el rendimiento del modelo no se distribuyó homogéneamente entre las variables disponibles, sino que estuvo determinado principalmente por un subconjunto de indicadores clínicos.

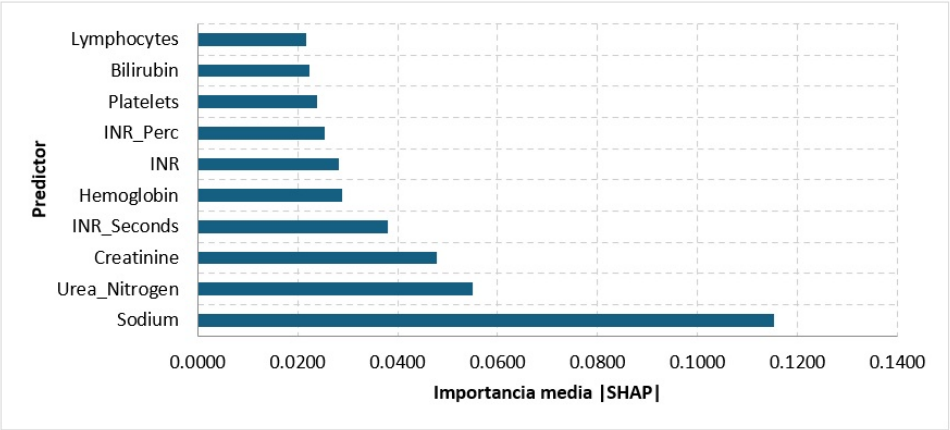

Fig. 4. Distribución de los valores SHAP de los principales predictores del modelo *Random Forest*.

H. Separabilidad del conjunto sintético y alcance de la validación interna

El análisis integral evidenció una característica particularmente relevante del conjunto TERMINET: las dos clases presentaron una separabilidad extraordinariamente elevada (Figura 5). En los 2.000 registros reservados para prueba, la regresión logística y *Random Forest* clasificaron correctamente los 1.514 registros sin complicaciones y los 486 registros con complicaciones, sin falsos positivos ni falsos negativos. Esta capacidad de separación también fue consistente con los valores de ROC-AUC, PR-AUC y las probabilidades estimadas.

Para comprobar que este comportamiento no dependiera exclusivamente de la complejidad de los algoritmos empleados, durante la auditoría se evaluó adicionalmente un árbol de decisión de profundidad limitada. Con una profundidad máxima de tres niveles alcanzó un ROC-AUC aproximado de 0,989 y una exactitud del 99,3 %, mientras que con cinco niveles el ROC-AUC se aproximó a 0,999 y la exactitud al 99,75 %. Por tanto, la elevada discriminación persistió incluso mediante estructuras predictivas relativamente simples.

	Predicho: sin complicación	**Predicho: complicación**
Real: sin complicación	1514	0
Real: complicación	0	486

Fig. 5. Matriz de confusión del modelo *Random Forest* en el conjunto de prueba.

La ausencia de errores de clasificación en la validación interna no debe interpretarse como evidencia de rendimiento equivalente en pacientes reales. Considerando que TERMINET fue generado mediante procedimientos de síntesis destinados a reproducir características del conjunto clínico de origen, los resultados sugieren que parte de la elevada capacidad predictiva puede estar relacionada con la estructura matemática preservada durante dicho proceso. En consecuencia, los valores obtenidos demuestran la factibilidad computacional del enfoque y la existencia de patrones altamente discriminantes dentro del conjunto sintético, pero requieren validación externa con datos clínicos independientes antes de establecer conclusiones sobre su desempeño en escenarios asistenciales.

CONCLUSIONES

Los modelos de inteligencia artificial desarrollados permitieron identificar con elevada capacidad discriminativa los registros asociados con complicaciones posquirúrgicas a partir de indicadores clínicos y funcionales del conjunto sintético TERMINET. La regresión logística y *Random Forest* clasificaron correctamente los 2000 registros del conjunto de prueba, mientras que XGBoost presentó un desempeño prácticamente equivalente. Estos resultados confirmaron la viabilidad computacional del enfoque predictivo propuesto.

La comparación de los algoritmos mostró que una mayor complejidad computacional no produjo ventajas sustanciales en este conjunto de datos. El rendimiento alcanzado incluso mediante modelos relativamente simples evidenció una elevada separabilidad entre las clases. Asimismo, los análisis de importancia e interpretabilidad mediante SHAP identificaron al sodio, nitrógeno ureico, creatinina y parámetros relacionados con INR entre las características con mayor contribución a las predicciones, sin que estas asociaciones puedan interpretarse como relaciones causales.

El desempeño excepcional debe valorarse considerando la naturaleza sintética de TERMINET. La discriminación prácticamente perfecta, la elevada calibración y el reducido solapamiento entre las clases sugieren que parte del rendimiento puede estar relacionado con las propiedades estadísticas preservadas durante el proceso de generación de los registros. En consecuencia, los resultados constituyen evidencia de validación interna y factibilidad computacional, pero no demuestran un rendimiento equivalente en pacientes quirúrgicos reales.

El estudio proporciona un procedimiento reproducible para la identificación temprana de complicaciones posquirúrgicas mediante aprendizaje automático e inteligencia artificial explicable. Como continuidad, resulta necesario evaluar el modelo en cohortes clínicas independientes y heterogéneas, analizar su estabilidad entre diferentes procedimientos y poblaciones y realizar posteriormente una validación prospectiva. Estas etapas permitirán determinar si los patrones identificados conservan su capacidad predictiva y establecer su eventual utilidad como herramienta de apoyo para la estratificación del riesgo posquirúrgico.

REFERENCIAS

[1] B. Xue, D. Li, C. Lu, C. R. King, T. Wildes, M. S. Avidan, T. Kannampallil, and J. Abraham, "Use of machine learning to develop and evaluate models using preoperative and intraoperative data to identify risks of postoperative complications," *JAMA Network Open*, vol. 4, no. 3, p. e212240, 2021.

[2] K. Merath, J. M. Hyer, R. Mehta *et al.*, "Use of machine learning for prediction of patient risk of postoperative complications after liver, pancreatic, and colorectal surgery," *Journal of Gastrointestinal Surgery*, vol. 24, pp. 1843–1851, 2020.

[3] P. Thottakkara, T. Ozrazgat-Baslanti, B. B. Hupf, P. Rashidi, P. Pardalos, P. Momcilovic, and A. Bihorac, "Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications," *PLoS ONE*, vol. 11, no. 5, p. e0155705, 2016.

[4] A. M. Hassan, A. Rajesh, M. Asaad, J. A. Nelson, J. H. Coert, B. J. Mehrara, and C. E. Butler, "Artificial intelligence and machine learning in prediction of surgical complications: Current state, applications, and implications," *The American Surgeon*, vol. 89, no. 1, pp. 25–30, 2023.

[5] S. Kokkinakis, E. I. Kritsotakis, and K. Lasithiotakis, "Artificial intelligence in surgical risk prediction," *Journal of Clinical Medicine*, vol. 12, no. 12, p. 4016, 2023.

[6] C. Varghese, E. M. Harrison, G. O'Grady, and E. J. Topol, "Artificial intelligence in surgery," *Nature Medicine*, vol. 30, no. 5, pp. 1257–1268, 2024.

[7] A. Guni, P. Varma, J. Zhang, M. Fehervari, and H. Ashrafian, "Artificial intelligence in surgery: The future is now," *European Surgical Research*, vol. 65, no. 1, pp. 22–39, 2024.

[8] A. Bonde, K. M. Varadarajan, N. Bonde *et al.*, "Assessing the utility of deep neural networks in predicting postoperative surgical complications: A retrospective study," *The Lancet Digital Health*, vol. 3, no. 8, pp. e471–e485, 2021.

[9] S. Chen, T. Deng, Q. Yang *et al.*, "Development and validation of an explainable machine learning model for predicting postoperative pulmonary complications after lung cancer surgery: A machine learning study," *EClinicalMedicine*, vol. 86, 2025.

[10] H. Deng, Z. Eftekhari, C. Carlin *et al.*, "Development and validation of an explainable machine learning model for major complications after cytoreductive surgery," *JAMA Network Open*, vol. 5, no. 5, p. e2212930, 2022.

[11] P. Li, S. Gao, Y. Wang *et al.*, "Utilising intraoperative respiratory dynamic features for developing and validating an explainable machine learning model for postoperative pulmonary complications," *British Journal of Anaesthesia*, vol. 132, no. 6, pp. 1315–1326, 2024.

[12] N. Kenig, J. M. Echeverria, and A. M. Vives, "Artificial intelligence in surgery: A systematic review of use and validation," *Journal of Clinical Medicine*, vol. 13, no. 23, p. 7108, 2024.

[13] W. T. Stam, L. K. Goedknegt, E. W. Ingwersen, L. J. Schoonmade, E. R. Bruns, and F. Daams, "The prediction of surgical complications using artificial intelligence in patients undergoing major abdominal surgery: A systematic review," *Surgery*, vol. 171, no. 4, pp. 1014–1021, 2022.

[14] T. J. Loftus, P. J. Tighe, A. C. Filiberto *et al.*, "Artificial intelligence and surgical decision-making," *JAMA Surgery*, vol. 155, no. 2, pp. 148–158, 2020.

[15] D. Lekka, A. Pnevmatikakis, E. Kanavos, B. Gottardelli, A. M. Tudor, P. Cornacchione, M. de Angeli, and A. Bellieni, "TERMINET eHealth post-operation complications synthetic dataset," 2024, dataset.