UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the :

Valentín Aragón, Leo; Ramos Terrades, Oriol (tut.). *Aplicación de técnicas de machine learning para la predicción precoz de insuficiencia renal aguda en pacientes hospitalizados.* (Universitat Autònoma de Barcelona), 2026. 11 pàgines.

This version is available at https://ddd.uab.cat/record/339015

under the terms of the  license.

Aplicación de técnicas de *machine learning* para la predicción precoz de insuficiencia renal aguda en pacientes hospitalizados

Leo Valentín Aragón

Resumen — La insuficiencia renal aguda (IRA) es una enfermedad multifactorial que se detecta mediante la subida de creatinina sérica o descenso de la diuresis, cuando el daño renal ya está instaurado. Una detección precoz permitiría tomar medidas preventivas adecuadas, siendo especialmente interesante en el ámbito hospitalario, donde se estima que hasta un 20% de los pacientes desarrolla IRA. Se propone desarrollar y validar un modelo basado en analíticas realizadas a pacientes hospitalizados, con el fin de estimar su probabilidad de desarrollar IRA (según criterios KDIGO) en un horizonte temporal de dos días. Se ha desarrollado un estudio retrospectivo utilizando una cohorte de pacientes del Hospital Universitario Reina Sofía (Córdoba, España). Se compararon LightGBM, XGBoost, CatBoost, Random Forest y LASSO con validación anidada separando pacientes (5 particiones por tres semillas), recalibración y *bootstrap* por paciente. Del total de ingresos, el 12,1% desarrollaron IRA y un 2.61% de las analíticas fueron positivas con un horizonte temporal de 2 días. El modelo que mejor AUROC consiguió fue LightGBM (0,777 [IC 95 % 0,752–0,799]). Según los valores SHAP, pesaron más la edad, los días desde el ingreso, la creatinina, parámetros derivados de ésta y los linfocitos; el riesgo aumentó con la edad, creatinina alta o en aumento, y los linfocitos y proteínas totales bajos. Con un umbral del 5% el 40% de los ingresos con IRA tuvo una alerta.

Palabras clave. Insuficiencia renal aguda, detección precoz, machine learning, analíticas de laboratorio

Abstract— Acute kidney injury (AKI) is a multifactorial condition detected by a rise in serum creatinine or a decrease in urine output, once renal damage is already established. Early detection would allow appropriate preventive measures to be taken, which is especially relevant in the hospital setting, where up to 20% of patients are estimated to develop AKI. We propose to develop and validate a model based on laboratory tests performed on hospitalized patients, in order to estimate their probability of developing AKI (according to KDIGO criteria) within a two-day time horizon. A retrospective study was conducted using a cohort of patients from the Reina Sofía University Hospital (Córdoba, Spain), admitted during 2025. LightGBM, XGBoost, CatBoost, Random Forest and LASSO were compared using nested validation with patient-level separation (5 folds × 3 seeds), recalibration and patient-level bootstrap. Of all admissions, 12.1% developed AKI, and 2.61% of laboratory requests were positive at a two-day horizon. The best-performing model in terms of AUROC was LightGBM (0.777 [95% CI 0.752–0.799]). According to SHAP values, age, days since admission, creatinine, creatinine-derived features and lymphocytes had the greatest weight; risk increased with age, high or rising creatinine, and low lymphocytes and total protein. With a 5% threshold, 40% of AKI cases received an alert.

Index Terms—Acute kidney injury, early detection, machine learning, laboratory tests

1 Introducción

La insuficiencia renal aguda (IRA) constituye un problema clínico relevante en el ámbito de la hospitalización, ya que se estima que aproximadamente un 20% de los pacientes hospitalizados la desarrollará [1]. Su aparición se relaciona con diversos factores propios del entorno hospitalario, como la exposición a determinados fármacos, la realización de intervenciones quirúrgicas o la presencia de infecciones [2]. El desarrollo de IRA se asocia a un incremento de las complicaciones clínicas y a una mayor duración de la estancia hospitalaria, lo que se traduce en un aumento de la morbimortalidad [3]. En este contexto, la detección precoz de la lesión renal puede facilitar la instauración de medidas terapéuticas oportunas que eviten la progresión del daño renal. En la actualidad, la IRA se define principalmente según los criterios establecidos por las guías de la organización Kidney Disease: Improving Global Outcomes (KDIGO), que se basan en la subida de creatinina sérica o descenso de la diuresis [4]. No obstante, uno de los principales inconvenientes de este enfoque diagnóstico es que las alteraciones en la creatinina suelen manifestarse de forma relativamente tardía en el curso de la lesión renal, lo que puede retrasar la identificación del problema y la instauración de medidas preventivas o terapéuticas [5].

Estudios recientes [6]-[11] sugieren que los modelos de aprendizaje automático pueden predecir de forma temprana la aparición de IRA en pacientes hospitalizados y podrían orientar intervenciones tempranas a partir de datos que los hospitales ya generan de forma rutinaria. En este contexto, en el presente trabajo se plantea como objetivo

- *E-mail de contacto: leovaragon@gmail.com*
- *Trabajo tutorizado por: Oriol Ramos Terrades*
- *Curso: 2026*

general desarrollar y evaluar modelos de aprendizaje automático que predigan con una antelación de hasta dos días la primera IRA de cada ingreso hospitalario, a partir de datos demográficos y de la evolución temporal de los parámetros bioquímicos y hematológicos incluidos en las analíticas de pacientes hospitalizados. Como objetivos específicos, desarrollar y comparar el rendimiento de cinco algoritmos con validación anidada, determinar las variables más relevantes y qué aportan el resto de pruebas de laboratorio al margen de la creatinina para la predicción de IRA y describir la utilidad del algoritmo elegido según el umbral de aviso.

1.1 Estado del arte

En los últimos años se han publicado numerosos modelos de aprendizaje automático para predecir IRA en pacientes hospitalizados [6]–[11]. Los más próximos a este trabajo, desarrollados basándose en datos de planta general hospitalaria, se detallan a continuación. Cho et al. [6] emplearon una definición KDIGO más estricta (creatinina basal igual al mínimo de los 7 días previos y aumento mínimo de 0,3 mg/dL de forma simultánea) para predecir IRA entre 1 y 3 días antes en 135.068 pacientes de tres hospitales distintos de Corea del Sur. Además de datos de laboratorio, usaron constantes vitales, fármacos, cirugías y pruebas con contraste, y obtuvieron un AUROC de 0,886 en validación externa. Liu et al. [7], con 76.957 ingresos, una incidencia de IRA del 9,4 % y sin equilibrar las clases, estimaron el riesgo un día antes mediante un modelo personalizado con aprendizaje por transferencia (AUROC 0,78 en validación cruzada de 5 pliegues, frente a 0,76 del modelo global). Guo et al. [9], con más de 560.000 ingresos de 15 hospitales, probaron el modelo con los pacientes de cuatro hospitales no usados en el entrenamiento y mostraron que se puede predecir IRA a 24, 48 y 72 h incluso sin variables derivadas de la creatinina (AUROC 0,860–0,986). Yu et al. [8] emplearon 59.936 pacientes con un caso por cada control y obtuvieron un AUROC de 0,93 en validación interna, que bajó a 0,81–0,83 en un hospital externo con casos sin equilibrar.

Otros trabajos se centran en poblaciones concretas y se alejan en mayor medida del diseño de este trabajo. Li et al. [10] estudiaron 1.457 pacientes de UCI con diabetes e insuficiencia cardiaca, de los que el 83 % desarrolló IRA, y obtuvieron con LightGBM un AUROC de 0,973 en entrenamiento y de 0,804 en validación. Zahr et al. [11], en 1.178 episodios de pacientes con anemia falciforme (82 con IRA), usaron constantes vitales medidas minuto a minuto y lograron, con 10 particiones aleatorias 70/30, un AUROC de 0,94 a 6 h y de 0,82 a 48 h.

La comparación entre estos estudios es difícil, ya que algunos equilibran casos y controles [8], [11]; en otros, se usan otros predictores aparte de la analítica de laboratorio [6], [7]; en algunos la validación es interna y en otros externa, las poblaciones y definiciones de IRA difieren y la utilidad clínica se evalúa de forma desigual. Cho et al. [6] es el más cercano en diseño a este trabajo y su definición de IRA inspira uno de los análisis de sensibilidad realizados, mientras que Liu et al. [7] es el más próximo en resultado. Sin embargo, hay pocos trabajos que usen solo las analíticas de hospitalización en España, con población general no seleccionada, que midan qué aportan las pruebas más allá de la creatinina.

2 MATERIAL Y MÉTODOS

2.1 Diseño del estudio y origen de los datos

Se realizó un estudio observacional, retrospectivo y unicéntrico con datos de pacientes ingresados en el Hospital Universitario Reina Sofía (Córdoba) entre el 1 de enero y el 31 de diciembre de 2025. Se emplearon los datos de las analíticas solicitadas al laboratorio desde los distintos servicios de hospitalización y la información demográfica y administrativa asociada a ellas. Los datos se pseudonimizaron antes del análisis: los identificadores de paciente y de petición se sustituyeron por códigos aleatorios y se eliminaron la fecha de nacimiento y el diagnóstico al ingreso. Las tablas de equivalencia se almacenaron por separado y no se utilizaron en el análisis. El estudio fue aprobado por el Comité de Ética de la Investigación Provincial de Córdoba (CEIm), que lo eximió del consentimiento informado (expediente SICEIA-2026-001098).

2.2 Preprocesamiento de la base de datos

Antes de definir la cohorte se preprocesó la base de datos usando Python 3.9.12 y JupyterNotebook 6.4.8. La tabla contenía una fila por prueba analítica. Se eliminaron las filas correspondientes a controles de calidad del laboratorio (que usaban su propia denominación característica) y las peticiones con códigos del sistema correspondientes a peticiones de prueba que no se correspondían con pacientes reales. Se eliminaron las copias exactas de una misma prueba dentro de una petición y se conservaron, numeradas, las repeticiones con resultado distinto. Los resultados que estaban registrados en formato texto se separaron como valor numérico y, cuando el laboratorio los informaba como inferior o superior a un límite (por ejemplo, <0,5 o >1000) o como aproximado, se conservó el valor numérico del límite y el símbolo o caracter se almacenó aparte, sin utilizarse como predictor. Cuando el laboratorio no informó el filtrado glomerular en una analítica, se calculó con CKD-EPI 2009. Los resultados no numéricos (cualitativos, no válidos o comentarios) no se incluyeron. Finalmente, la tabla se reestructuró (pivotación) para que cada fila correspondiera a una petición analítica.

2.3 Selección de la cohorte

Los datos no incluyen la fecha de ingreso, por lo que se consideró que comenzaba un nuevo ingreso cuando el paciente pasaba más de 15 días sin ninguna analítica. Se incluyeron los ingresos de pacientes de 18 años o más, con al menos 48 horas entre la primera y la última analítica y con al menos dos determinaciones de creatinina. Se excluyeron los ingresos con una primera creatinina superior a 4 mg/dL o un filtrado glomerular estimado (CKD-EPI 2009) inferior a 60 mL/min/1,73 m², por enfermedad renal previa, y aquellos cuyo primer diagnóstico informativo, registrado el día 0 o 1, fuera renal, al considerarse que la IRA ya estaba diagnosticada. Los ingresos con la primera creatinina tardía (más de 7 días tras la primera petición) o con diagnóstico renal registrado tardíamente se conservaron y se marcaron, para evaluarlos después en el análisis de sensibilidad. La cohorte final comprendió 6.144 pacientes, 6.966 ingresos y 56.770 peticiones utilizables (Fig. 1).

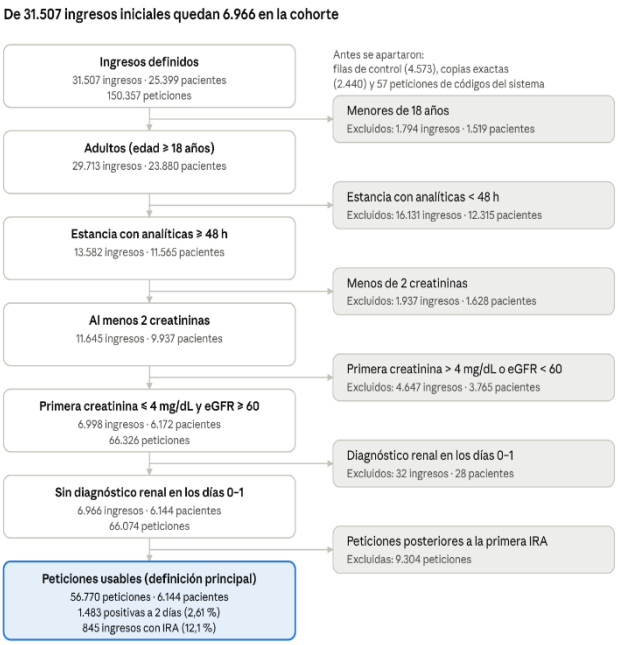


Fig. 1. Diagrama de flujo del preprocesado de la base de datos y selección de la cohorte.

2.4 Definición de IRA y de la variable respuesta

La IRA se definió exclusivamente por creatinina, según los criterios KDIGO [4], sin utilizar la diuresis por la ausencia de este parámetro en las analíticas recogidas. Como las fechas de las analíticas no incluyen la hora, para cada ingreso y día se consideró la creatinina máxima. Un día se clasificó como IRA si cumplía alguno de dos criterios: un aumento de la creatinina de al menos 0,3 mg/dL respecto al mínimo de los dos días previos o un valor al menos 1,5 veces el mínimo de los siete días previos. Solo se consideró la primera IRA de cada ingreso. Se calculó adicionalmente el estadio KDIGO.

Como análisis de sensibilidad se construyó una etiqueta alternativa (etiqueta B), inspirada en la definición de Cho et al. [6], en la que el criterio exige obligatoriamente el aumento de al menos 0,3 mg/dL, para no contar subidas mínimas cuando la creatinina de partida es muy baja.

Cada petición analítica se trató como un momento de predicción independiente. Los predictores se construyeron únicamente con la información disponible hasta ese instante (de modo que el modelo nunca ve información posterior al desenlace), y la variable respuesta indica si la primera IRA del ingreso aparece en los 1–2 días siguientes. Para cada horizonte temporal (de 1 a 4 días), una petición se etiquetó como positiva si la primera IRA ocurría entre 1 y x días después.

2.5 Variables predictoras

Como predictores se utilizaron únicamente resultados de pruebas de laboratorio y algunos datos demográficos (edad en años y sexo) del paciente. En cada petición se tomó, para cada prueba, el último valor medido durante el ingreso hasta ese momento. Se seleccionaron las pruebas para las que se disponía de algún valor previo en al menos el 20 % de las peticiones utilizables. Se excluyeron los índices de hemólisis, ictericia y lipemia y las pruebas derivadas de otras mediciones (como el filtrado glomerular estimado, el calcio corregido o el hiato aniónico), que se analizaron aparte. Quedaron 85 pruebas de laboratorio.

De cada prueba se generaron cinco variables: el último valor medido, el cambio entre las dos últimas mediciones, el cambio respecto al primer valor del ingreso, y el mínimo y el máximo hasta ese momento. A ellas se añadieron seis variables de contexto: edad, sexo, carácter urgente de la petición, días desde el inicio del ingreso, y la diferencia y el cociente entre la última creatinina y su mínimo previo, lo que en total supone 431 variables (consultar tablas A1 y A2 del Anexo). No se utilizó el servicio solicitante, pero se usó posteriormente en los análisis adicionales. Tampoco se incluyeron en el conjunto principal el número de mediciones previas ni el tiempo desde la última medición, que reflejan cuánto se ha monitorizado al paciente (a partir de ahora variables indicadoras de monitorización) y podrían delatar una sospecha clínica previa de IRA.

Se compararon cinco conjuntos de variables: uno básico con solo creatinina y datos del paciente (C), el conjunto principal con todas las pruebas (P), el principal sin creatinina ni filtrado glomerular (P sin creatinina), y dos de sensibilidad que añaden las variables indicadoras de monitorización (P_mon) y las pruebas derivadas de otras (P_der).

2.6 Modelos y validación

Se compararon cinco algoritmos: regresión logística con penalización L1 (LASSO), random forest y tres modelos de árboles potenciados (XGBoost, LightGBM y CatBoost). No se remuestreó la muestra. El desequilibrio de clases se abordó dando más peso a los casos positivos: en los árboles potenciados, el peso de la clase positiva (1, 3 o 6) se incluyó en la búsqueda de hiperparámetros, y en el ran-

dom forest y el LASSO se usaron pesos de clase que compensaron la menor frecuencia de los casos positivos. Los valores ausentes los manejan directamente los modelos de árboles potenciados; en el random forest y en el LASSO se imputaron con la mediana y, en caso del LASSO, además se añadió un indicador de ausencia y se estandarizaron las variables para que estuvieran en la misma escala. Los modelos se implementaron en Python 3.9.12 mediante JupyterNotebook 6.4.8 con scikit-learn 1.0.2 (random forest y LASSO), LightGBM 4.6.0, XGBoost 2.1.4 y CatBoost 1.2.10.

La validación fue anidada y por paciente, para que ningún paciente apareciera a la vez en entrenamiento y en prueba. Los pacientes se repartieron aleatoriamente en cinco grupos, de modo que cada grupo actuó una vez como prueba mientras los otros cuatro se usaban para entrenar; este procedimiento se repitió con tres semillas aleatorias, lo que dio lugar a 15 repeticiones.

Dentro de cada repetición, el entrenamiento externo se dividió en 3 grupos internos (también por paciente y manteniendo la proporción de ingresos con y sin IRA), con los que se buscaron los mejores hiperparámetros mediante Optuna con el estimador de Parzen estructurado en árbol (TPE) (método también empleado por Zheng et al. [12]), maximizando el AUPRC, con 15 iteraciones de búsqueda en LightGBM, 12 en XGBoost, 10 en CatBoost y 8 en random forest y LASSO. Por su coste de cálculo, el LASSO se evaluó solo con la primera semilla (5 repeticiones), y sus resultados se compararon con los de los demás algoritmos sobre esas repeticiones comunes. Además, como ya se ha comentado, durante el entrenamiento se dio más peso a los casos positivos para compensar la baja frecuencia de IRA. Esto sobreestima las probabilidades, por lo que se recalibraron con una regresión logística ajustada sobre predicciones de pacientes que el modelo no había visto durante su entrenamiento.

El modelo principal se eligió como el de mayor AUROC entre los cinco algoritmos en la validación anidada. El conjunto principal de variables (P) se evaluó con los cinco algoritmos, y los demás conjuntos descritos en el apartado 2.5 se evaluaron solo con LightGBM. Además, se evaluaron modelos parsimoniosos con las 10, 20 y 40 variables más importantes, seleccionadas en cada repetición solo con los datos de entrenamiento, sin usar el grupo de prueba.

El grupo de prueba externo no participó en ninguna decisión del proceso.

2.7 Métricas de evaluación

Discriminación. Se calcularon el área bajo la curva ROC (AUROC) y el AUPRC en cada grupo de prueba de cada repetición y se promediaron, en lugar de calcularlos sobre todas las predicciones juntas, porque las escalas de riesgo difieren entre repeticiones.

Calibración. La concordancia entre el riesgo predicho y la frecuencia observada de IRA se evaluó con un gráfico de calibración, que compara ambos por tramos de riesgo, y con dos medidas resumen: la pendiente y la ordenada en el origen de la calibración, cuyo valor ideal es 1 y 0, respectivamente, y el error de Brier, que indica el error cuadrático medio entre la probabilidad predicha y el resultado observado.

Utilidad. Se calcularon la sensibilidad, la especificidad, el valor predictivo positivo (VPP) y la proporción de peticiones con alerta para umbrales de riesgo del 3, 5 y 10 %. Se calcularon por petición y por ingreso, y en este último caso se midió la antelación del aviso, los días con alerta y la proporción de ingresos sin IRA con alguna alerta.

Intervalos de confianza y comparaciones. Los intervalos de confianza al 95 % se obtuvieron por bootstrap de pacientes (1.000 réplicas, con los mismos pacientes remuestreados para todos los modelos, lo que permite emparejar las diferencias entre ellos). Los algoritmos y los conjuntos de variables se compararon mediante la diferencia de AUROC frente al modelo principal. En el análisis de sensibilidad con la etiqueta B se usaron 500 réplicas por coste de cálculo. Los intervalos no incluyen la variabilidad debida al reentrenamiento de los modelos.

Análisis adicionales. Se evaluaron los horizontes de 1, 3 y 4 días (LightGBM con el conjunto P), la exclusión de los ingresos con primera creatinina tardía, y el rendimiento por sexo, por edad (menos de 50, 50 a 64, 65 a 74 y 75 años o más) y por servicio solicitante. Este último análisis fue exploratorio y se realizó tras ver los resultados. No se aplicó ninguna medida para corregir las diferencias entre grupos. La interpretación del modelo se realizó con valores SHAP (versión 0.49.1) sobre un modelo entrenado con todos los datos.

3 Resultados

3.1 Estadística descriptiva

La cohorte final incluyó 6.966 ingresos de 6.144 pacientes distintos, con 56.770 peticiones analíticas utilizables (Fig. 1). Desarrollaron IRA 845 ingresos (12,1 %), la mayoría de estadio 1 (72,7 %), seguidos del estadio 2 (19,4 %) y el 3 (7,9 %); con la definición alternativa de IRA (el conjunto B, definición de Cho et al. [6]) fueron 720. Como el desenlace se evaluó por petición, la prevalencia de peticiones positivas fue del 2,61 % a 2 días (1.483 de 56.770) y aumentó al alargar el horizonte, desde el 1,34 % a 1 día (760) hasta el 4,26 % a 4 días (2.421). Los ingresos con IRA correspondían a pacientes de más edad (mediana de 66,7 frente a 63,7 años), con un eGFR algo menor y con mayor frecuencia en Anestesiología y Reanimación (18,8 % frente a 11,6 %), sin diferencias por sexo (Tabla I).

3.2 Comparativa de modelos y elección del modelo principal

El AUROC de los cinco algoritmos osciló entre 0,746 y 0,777 (Tabla II). LightGBM obtuvo el mayor AUROC, 0,777 (IC 95 % 0,752–0,799), con un AUPRC de 0,122 frente a una prevalencia del 2,61 % (Fig.2), por lo que se

TABLA I

CARACTERÍSTICAS DE LOS INGRESOS SEGÚN TENGAN O NO IRA

Variable	Total (n = 6.966)	Con IRA (n = 845)	Sin IRA (n = 6.121)	p
Edad (años)	64,1 [53,2–73,9]	66,7 [58,0–75,3]	63,7 [52,7–73,7]	<0,001
Sexo masculino, n (%)	4.096 (58,8)	505 (59,8)	3.591 (58,7)	0,569
Primer servicio: Anestesiología, n (%)	868 (12,5)	159 (18,8)	709 (11,6)	<0,001
Creatinina (mg/dL)	0,8 [0,7–0,9]	0,8 [0,6–0,9]	0,8 [0,7–0,9]	0,888
Urea (mg/dL)	34,0 [26,0–45,0]	35,0 [27,0–48,0]	34,0 [26,0–45,0]	0,016
Proteína C reactiva (mg/L)	40,6 [9,0–121,4]	48,7 [12,6–138,8]	39,3 [8,7–118,7]	0,002
Linfocitos (%)	15,2 [9,1–23,4]	12,9 [7,2–20,3]	15,5 [9,4–23,7]	<0,001
Proteínas totales (g/dL)	6,3 [5,7–6,8]	6,0 [5,4–6,5]	6,3 [5,8–6,8]	<0,001
Potasio (mmol/L)	4,0 [3,6–4,3]	3,9 [3,5–4,2]	4,0 [3,6–4,3]	<0,001

Mediana [rango intercuartílico] salvo indicación. Edad, sexo y servicio son los de la primera petición del ingreso; las pruebas son el primer valor medido en el ingreso (disponible en el 100 % de los ingresos, salvo proteína C reactiva, 88 %, y proteínas totales, 82 %). p: Mann-Whitney (continuas) y chi-cuadrado (proporciones); valores descriptivos.

eligió como modelo principal. XGBoost rindió de forma equivalente (AUROC 0,774; diferencia con LightGBM de −0,003, IC 95 % −0,008 a 0,001) y con un AUPRC algo mayor (0,127), mientras que CatBoost, el random forest y el LASSO rindieron algo menos (diferencias de −0,011, −0,014 y −0,035, respectivamente). La mejora, por tanto, no dependió de la elección de un algoritmo concreto.

TABLA II

COMPARACIÓN DE ALGORITMOS (HORIZONTE DE 2 DÍAS)

Algoritmo	AUROC (IC 95 %)	AUPRC	Dif. AUROC con LightGBM (IC 95 %)
LightGBM	0,777 (0,752–0,799)	0,122	Referencia
XGBoost	0,774 (0,748–0,796)	0,127	−0,003 (−0,008 a 0,001)
CatBoost	0,766 (0,740–0,790)	0,103	−0,011 (−0,019 a −0,003)
Random Forest	0,763 (0,735–0,788)	0,106	−0,014 (−0,023 a −0,004)
LASSO*	0,746 (0,718–0,772)	0,086	−0,035 (−0,053 a −0,018)

*LASSO solo con la semilla 7 (5 repeticiones). Prevalencia: 2,61 %.

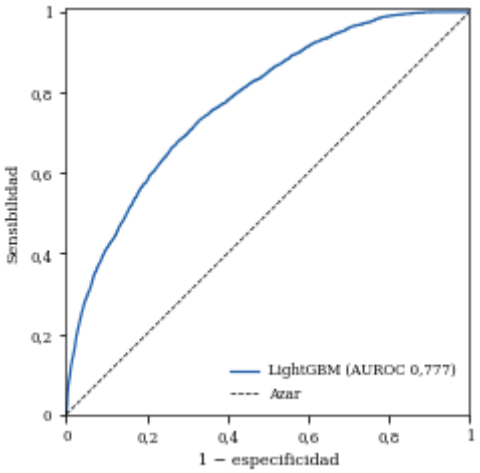


Fig. 2. Curva ROC de LightGBM (horizonte de 2 días) con las predicciones fuera de muestra de las 15 repeticiones. El AUROC indicado es la media de las 15 repeticiones.

3.3 Aportación de las pruebas distintas a la creatinina

Con solo edad, sexo, días de ingreso y formas de creatinina (conjunto C), el AUROC fue de 0,751, 0,026 menos que con las 431 variables del conjunto principal (Tabla III). Al quitar la creatinina y los cálculos derivados de ésta del conjunto principal, el AUROC bajó solo 0,020 (0,757), de modo que el resto de la analítica de laboratorio aporta por sí sola gran parte de la señal, aunque no sustituye por completo a la creatinina. Las pruebas derivadas de cálculos (como el eGFR o el calcio corregido por proteínas) no mejoraron el rendimiento (0,778), mientras que la monitorización, es decir, cuántas veces se midió cada prueba, lo mejoró (0,788), pero se excluyó del modelo principal porque refleja la sospecha clínica y no el estado del paciente. Un modelo con las 20 variables más importantes alcanzó prácticamente el mismo rendimiento (0,771), mientras que con 10 variables el AUROC bajó a 0,762.

TABLA III

APORTACIÓN DE LOS CONJUNTOS DE VARIABLES (LIGHTGBM)

Conjunto de variables	AUROC (IC 95 %)	Dif. con P (IC 95 %)
P: 431 variables	0,777 (0,752–0,799)	Referencia
C: edad, sexo, días y creatinina	0,751 (0,729–0,772)	−0,026 (−0,045 a −0,007)
P sin formas de creatinina	0,757 (0,731–0,780)	−0,020 (−0,029 a −0,011)
P con variables derivadas	0,778 (0,752–0,800)	+0,001 (−0,004 a 0,005)
P con monitorización	0,788 (0,763–0,809)	+0,011 (0,005 a 0,017)
Parsimonioso, 40 variables	0,772 (0,747–0,793)	−0,005 (−0,011 a 0,000)
Parsimonioso, 20 variables	0,771 (0,746–0,793)	−0,006 (−0,014 a 0,001)
Parsimonioso, 10 variables	0,762 (0,738–0,786)	−0,015 (−0,026 a −0,004)

3.4 Aportación de las variables a la predicción

Para entender mejor el modelo se calcularon valores SHAP sobre el modelo LightGBM entrenado con las 20 variables más frecuentes en las 15 repeticiones. Las variables de contexto que más pesaron fueron la edad (SHAP medio absoluto de 0,31) y los días desde el ingreso (0,20); y entre las analíticas, las formas derivadas de creatinina (en conjunto, 0,37) y los linfocitos (0,36), seguidos de las proteínas totales (0,16), la PCR (0,086) y el potasio (0,074) (Fig. 3). El riesgo aumentó con mayor edad, creatinina alta o en aumento, y PCR alta, y con porcentaje de linfocitos y proteínas totales bajos. Este patrón podría ser compatible con inflamación o gravedad general, aunque el cálculo de SHAP no refleja relaciones causales.

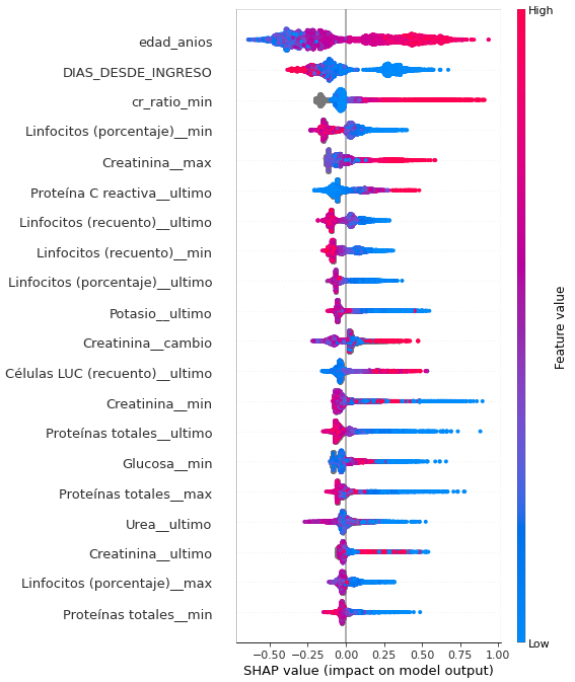


Fig. 3. Valores SHAP de las 20 variables más frecuentes en el modelo parsimonioso de LightGBM (horizonte de 2 días).

3.5 Calibración y utilidad por umbral

Tras recalibrar, el modelo principal estuvo bien calibrado entre el 0,8 % y el 10 % de riesgo predicho (pendiente 1,01; ordenada en el origen 0,00), mientras que sin recalibrar sobrestimaba el riesgo (pendiente 0,84; ordenada −0,91) (Fig. 4). Por debajo del 0,7 % siguió sobrestimándolo, y por encima del 10 % no hubo suficientes peticiones para evaluarlo. Con un umbral del 5 %, el modelo avisó en el 13,2 % de las peticiones, con una sensibilidad del 44,7 % y un VPP del 9,1 % (una de cada once alertas fue seguida de IRA); el 40,1 % de los ingresos con IRA recibió un aviso en la ventana previa, con una antelación mediana de 2 días, y el 25,6 % de los ingresos sin IRA recibió alguna alerta. Un umbral del 3 % detectó más casos (65,1 % de los ingresos con IRA), a costa de alertar al 49,3 % de los ingresos sin IRA, mientras que uno del 10 % redujo las alertas, pero solo detectó el 16,0 % (Tabla IV; resultados ampliados en tabla A3 de Anexos). La utilidad por umbral de aviso con horizonte de 1 días se puede consultar en la tabla A4.

3.5 Análisis de sensibilidad y subgrupos

El AUROC disminuyó al alargar el horizonte de predicción: fue de 0,803 a 1 día (IC 95 % 0,772–0,827), 0,777 a 2 días, 0,752 a 3 días y 0,738 a 4 días (Tabla A5 de Anexos). El resultado no dependió de la definición de IRA: con la etiqueta B el AUROC fue de 0,776, frente a 0,777 con la principal (diferencia de 0,000; IC 95 % −0,012 a 0,009). En los ingresos con la primera creatinina pedida tras 7 días de ingreso (el 14,3 % de las peticiones), la IRA solo puede detectarse con retraso, y excluirlos redujo el AUROC a 0,755. Por subgrupos, no hubo diferencias entre hombres y mujeres, y el rendimiento fue menor en los pacientes más mayores (AUROC de 0,710 a partir de los 75 años, frente a 0,796 en menores de 50). Los resultados de los análisis de sensibilidad se pueden consultar en la tabla A6 de Anexos.

En el análisis exploratorio por servicio (Tabla A7), el modelo discriminó mejor en Anestesiología y Reanimación (AUROC 0,828; IC 95 % 0,784–0,866; prevalencia de peticiones positivas del 6,0 %) que en el conjunto del resto de servicios (0,739; 0,717–0,761; prevalencia del 2,1 %).

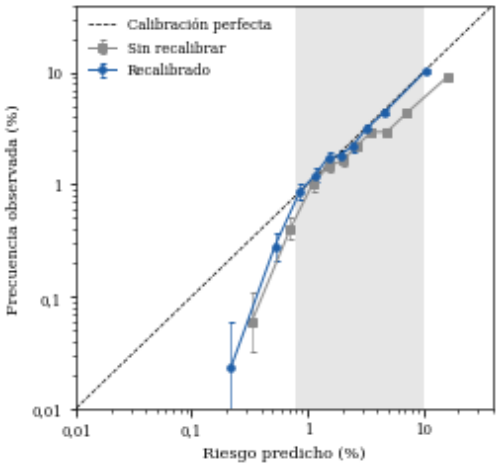


Fig. 4. Calibración de LightGBM (horizonte de 2 días) con las predicciones fuera de muestra de las 15 repeticiones, antes y después de la recalibración logística. La calibración es adecuada entre 0,8 y 10 % (zona sombreada).

4 Discusión

4.1 Interpretación de los resultados

Con solo analíticas de laboratorio y datos básicos del paciente (edad, sexo y urgencia de la petición), el modelo principal discriminó de forma moderada (AUROC 0,777) la primera IRA de los 1–2 días siguientes. LightGBM y

TABLA IV

UTILIDAD POR UMBRAL DE AVISO (HORIZONTE DE 2 DÍAS)

Umbral	Peticiones con aviso (%)	Sens. (%)	VPP (%)	IRA avisadas en la ventana (%)	Ingresos sin IRA con aviso (%)
2 %	43,0	79,6	4,9	80,7	69,4
3 %	27,4	66,3	6,4	65,1	49,3
5 %	13,2	44,7	9,1	40,1	25,6
10 %	3,3	20,9	17,4	16,0	6,5

Media de las 15 repeticiones. Sens.: sensibilidad; VPP: valor predictivo positivo (ambos por petición). «IRA avisadas en la ventana»: ingresos con IRA con al menos un aviso en los 1–2 días previos.

XGBoost rindieron de forma equivalente, y CatBoost, Random Forest y LASSO algo menos, por lo que la capacidad predictiva parece depender más de los datos que del algoritmo.

En cuanto a lo que aportan las pruebas más allá de la creatinina: un conjunto de solo 10 variables (creatinina y sus cambios, edad, sexo, días de ingreso y urgencia) alcanzó un AUROC de 0,751, y el resto de la analítica añadió 0,026 hasta llegar a 0,777. Incluso sin creatinina ni filtrado glomerular, el modelo conservó un AUROC de 0,757, por lo que la analítica contribuye por sí sola, aunque la creatinina sigue aportando. Tras la edad, los días de ingreso y la evolución de la creatinina, las variables más influyentes fueron los linfocitos y la PCR, un patrón que podría reflejar inflamación o gravedad general, sin implicar causalidad. Un modelo de 20 variables rindió casi igual que el de 431 (0,771 frente a 0,777), lo que facilitaría su revisión y su posible uso.

El AUROC del modelo principal (0,777) es inferior al de la mayoría de los estudios citados [6]–[11], pero las cifras no son directamente comparables porque los diseños difieren en tres aspectos. Primero, Yu et al. [8] y Zahr et al. [11] equilibran casos y controles, lo que altera las probabilidades y puede sobrestimar el rendimiento según cómo se construya la muestra; en este estudio se mantuvo la prevalencia natural. Segundo, Cho et al. [6] y Guo et al. [9] incluyen además otras variables, como constantes vitales o medicación, mientras que este modelo usa solo analíticas y datos básicos. Tercero, solo Cho et al. [6] y Yu et al. [8] validaron el modelo fuera del centro, mientras que aquí la validación fue interna, separando pacientes. El resultado más cercano es el de Liu et al. [7] (0,78), que tampoco equilibra las clases.

En cuanto a la aplicabilidad, con un umbral del 5 % se avisa al 40 % de los ingresos con IRA con una antelación promedio de 2 días, pero el 26 % de los ingresos sin IRA recibe alguna alerta, y solo una de cada once alertas anticipa realmente una IRA. Por ello, el modelo se plantea como apoyo a la decisión clínica, por ejemplo para priorizar la revisión del estado de los pacientes o de su medicación, y el umbral debería elegirlo el equipo clínico según su capacidad de respuesta. El rendimiento fue menor en los mayores de 75 años y mejor en Anestesiología y Reanimación, aunque este último análisis fue exploratorio y es uno de los servicios que acumulan más IRA.

4.2 Limitaciones

Es un estudio unicéntrico y con datos de un solo año, por lo que no sabemos cómo se comportaría el modelo en otros hospitales, con otros laboratorios o en otros periodos de tiempo. Tampoco se puede comprobar si las alertas cambiarían las decisiones clínicas sobre el manejo de los pacientes. Los ingresos en los extremos del periodo pudieron tener un seguimiento incompleto: una IRA posterior al 31 de diciembre no consta, y ese ingreso cuenta como sin IRA.

El modelo solo usa parámetros de analíticas y sus demográficos asociados; no incorpora constantes vitales, medicación ni diagnósticos, de modo que no puede aprovechar la información que sí tiene el médico. La IRA se definió únicamente por creatinina (sin diuresis) y las fechas no incluyeron la hora, por lo que la antelación del aviso es solo una estimación: la fecha de la IRA es la de la analítica que cumple el criterio y puede ser posterior al inicio real del daño renal. Además, las analíticas no se piden con una periodicidad fija; el intervalo entre ellas varía de un paciente a otro y según la situación clínica, de modo que un mismo daño renal se detecta antes o después según cuándo se mida la creatinina. Los valores informados como inferiores o superiores a un límite se trataron como el valor de ese límite, lo que podría producir un pequeño sesgo. Se conservaron solo las pruebas con valor en al menos el 20 % de las peticiones, un criterio práctico que no se comparó con otros umbrales. La variable de urgencia de la petición refleja en parte la organización del hospital (por ejemplo, las analíticas de control pedidas en fin de semana), no solo la gravedad del paciente, lo que puede limitar su valor. Y la monitorización (cuántas veces se midió cada prueba) mejora el rendimiento, pero se excluyó del modelo principal por reflejar la sospecha clínica.

Las peticiones de un mismo paciente no son independientes, por lo que el tamaño efectivo es menor que el número de filas de la base de datos. LightGBM se eligió por ser el que mejor AUROC obtuvo entre los algoritmos comparados, y su rendimiento se informa con esos mismos datos, por lo que podría estar algo sobrestimado. La búsqueda de hiperparámetros fue limitada (entre 8 y 15 iteraciones según coste de cálculo), con más iteraciones en LightGBM, por lo que la comparación entre algoritmos podría no ser plenamente equitativa; además, el LASSO solo se evaluó con una semilla y los modelos con menos variables reutilizaron los hiperparámetros de la búsqueda con todas.

El análisis por servicio fue exploratorio y se eligió tras ver los resultados de 19 servicios, por lo que el rendimiento por edad y servicio debe interpretarse con cautela.

5 Conclusiones y líneas abiertas

Con solo las analíticas de laboratorio de pacientes hospitalizados y sus demográficos y sin otros datos clínicos, es posible predecir con una discriminación moderada (AUROC 0,777) la primera IRA de los 1–2 días siguientes en esta población, y las probabilidades, una vez recalibradas, se pueden leer como riesgo entre el 0,8 y el 10 %. LightGBM y XGBoost rindieron de forma parecida y el resto de algoritmos algo menos, lo que indica que la capacidad predictiva depende más de las variables que del algoritmo. Respecto al objetivo secundario, las pruebas distintas de la creatinina mejoran de forma pequeña pero medible la predicción. Sin las variables derivadas de creatinina el AUROC baja de 0,777 a 0,757, y el conjunto de 10 variables centrado en la creatinina alcanzó 0,751; la

información de los demás parámetros incluidos en las analíticas (por ejemplo, linfocitos, proteínas totales o PCR) aporta señal, pero no sustituyen a la creatinina. Un modelo con 20 variables mantiene prácticamente el mismo rendimiento (0,771), lo que facilitaría su implementación. En la práctica clínica, el modelo permitiría avisar con antelación en una parte de los casos, a costa de numerosas falsas alertas según el umbral, por lo que serviría como apoyo a la decisión clínica y no como herramienta autónoma.

El siguiente paso sería una validación externa y prospectiva en otros hospitales, junto con la evaluación de si las alertas modifican el manejo clínico y los desenlaces de los pacientes. Sería útil comparar este modelo con otros que incorporen constantes vitales, medicación y diagnósticos, para medir qué aportan más allá de la analítica; explorar umbrales distintos por servicio, y analizar con más detalle el mejor rendimiento observado en servicios como Anestesiología y Reanimación, que en este trabajo es solo exploratorio. Por último, convendría ampliar la búsqueda de hiperparámetros y evaluar todos los algoritmos con las mismas semillas para confirmar la comparación.

Declaración de uso de inteligencia artificial

Para este trabajo se utilizó el asistente de inteligencia artificial Claude (Anthropic; modelo Claude Sonnet 5.5) como apoyo en el diseño del protocolo de trabajo, la escritura de código en Python (preparación de datos, entrenamiento, análisis estadístico y figuras), la revisión de errores de código y el diseño de las tablas. La búsqueda bibliográfica, las decisiones de diseño del estudio, la definición de la cohorte y el etiquetado, la ejecución de todo el código, la comprobación e interpretación de los resultados y la redacción fueron realizadas por el autor. El asistente no tuvo acceso a los datos de los pacientes, solo al código y a los resultados agregados que el autor le comunicó. El autor asume la responsabilidad del contenido.

El uso de esta herramienta ha permitido un trabajo fluido y rápido, en especial en la elaboración de código en Python, que habría llevado muchos meses de trabajo a una persona no entrenada en programación. Sin embargo, el asistente tendió a cometer errores realizando cálculos internos, que tuvieron que ser constantemente contrastados con lo obtenido al ejecutar las celdas de código, dándose por válidos solo estos últimos. La herramienta tiende además a dar la razón a lo que propone el usuario y a presentar sus respuestas con seguridad, a pesar de estar cometiendo errores, por lo que, pese a ser una herramienta muy potente, necesitó supervisión constante por parte del autor. Aun así, el ahorro de tiempo fue notable, y las ideas que aportó sin que se le pidieran llevaron a realizar análisis comparativos adicionales, de los que se obtuvieron resultados interesantes.

Referencias

[1] N. Ramos Terrades *et al.*, "Documento de consenso para el diagnóstico y tratamiento de la insuficiencia renal aguda," *Nefrología*, vol. 46, Art. no. 501360, 2026.

[2] E. A. J. Hoste *et al.*, "Epidemiology of acute kidney injury in critically ill patients: The multinational AKI-EPI study," *Intensive Care Med.*, vol. 41, no. 8, pp. 1411–1423, Aug. 2015, doi: 10.1007/s00134-015-3934-7.

[3] J. A. Kellum, F. E. Sileanu, A. Bihorac, E. A. J. Hoste, and L. S. Chawla, "Recovery after acute kidney injury," *Am. J. Respir. Crit. Care Med.*, vol. 195, no. 6, pp. 784–791, Mar. 2017, doi: 10.1164/rccm.201604-0799OC.

[4] P. Rodríguez Benítez, N. Ramos Terrades, and E. Poch, "Insuficiencia renal aguda," in *Nefrología al día*, V. Lorenzo and J. M. López Gómez, Eds., 2025. [Online]. Available: https://nefrologiaaldia.org/articulo/insuficiencia-renal-aguda/ (accessed Oct. 18, 2026).

[5] J. Mårtensson, C. R. Martling, and M. Bell, "Novel biomarkers of acute kidney injury and failure: Clinical applicability," *Br. J. Anaesth.*, vol. 109, no. 6, pp. 843–850, Dec. 2012, doi: 10.1093/bja/aes357.

[6] J. Cho *et al.*, "Machine learning to assist in managing acute kidney injury in general wards: Multicenter retrospective study," *J. Med. Internet Res.*, vol. 27, Art. no. e66568, Mar. 2025, doi: 10.2196/66568.

[7] K. Liu *et al.*, "Development and validation of a personalized model with transfer learning for acute kidney injury risk estimation using electronic health records," *JAMA Netw. Open*, vol. 5, no. 7, Art. no. e2219776, Jul. 2022, doi: 10.1001/jamanetworkopen.2022.19776.

[8] X. Yu, W. Wang, R. Wu, X. Gong, Y. Ji, and Z. Feng, "Construction of a machine learning-based interpretable prediction model for acute kidney injury in hospitalized patients," *Sci. Rep.*, vol. 15, Art. no. 9313, 2025, doi: 10.1038/s41598-025-90459-5.

[9] S. Guo *et al.*, "Developing machine learning-driven acute kidney injury predictive models using non-standard EMRs in resource-limited settings," *Med. Phys.*, vol. 52, Art. no. e70038, 2025, doi: 10.1002/mp.70038.

[10] G. Li, Z. Zhao, Z. Yu, J. Liao, and M. Zhang, "Machine learning for risk prediction of acute kidney injury in patients with diabetes mellitus combined with heart failure during hospitalization," *Sci. Rep.*, vol. 15, Art. no. 10728, 2025, doi: 10.1038/s41598-025-87268-1.

[11] R. S. Zahr *et al.*, "Machine learning predicts acute kidney injury in hospitalized patients with sickle cell disease," *Am. J. Nephrol.*, vol. 55, no. 1, pp. 18–24, 2024, doi: 10.1159/000534864.

[12] L. Zheng, J. Yang, L. Zhao, C. Li, K. Fang, S. Li, J. Wu, and M. Zheng, "Development and validation of the PHM-CPA model to predict in-hospital mortality for cirrhotic patients with acute kidney injury," *Dig. Liver Dis.*, vol. 57, pp. 485–493, 2025, doi: 10.1016/j.dld.2024.09.012.

Anexo. Material complementario.

TABLA A1

VARIABLES DE CONTEXTO DEL MODELO PRINCIPAL

Variable	Descripción
edad_anios	Edad (años) en el momento de la petición
sexo_M	Sexo masculino (1) o femenino (0)
DIAS_DESDE_INGRESO	Días transcurridos desde la primera petición del ingreso
cr_ratio_min	Cociente entre la última creatinina conocida y el mínimo del ingreso hasta ese momento
cr_menos_min	Diferencia entre la última creatinina conocida y el mínimo del ingreso hasta ese momento
urgente_n	Petición urgente (1) o no urgente (0)

Junto con las pruebas de la Tabla A2, forman las 431 variables del modelo principal (85 × 5 + 6).

TABLA A2

PRUEBAS ANALÍTICAS DEL MODELO PRINCIPAL Y COBERTURA

Prueba	Cobertura (%)	Prueba	Cobertura (%)
Basófilos (porcentaje)	81,0	Presión parcial de O2 (a)	50,0
Basófilos (recuento)	81,0	Calcio Iónico (a)	49,8
Concentración de hemoglobina corpuscular media	81,0	Hematocrito (gasómetro; a)	49,8
Células LUC (porcentaje)	81,0	Potasio (a)	49,8
Células LUC (recuento)	81,0	Sodio (a)	49,8
Dispersión de hematíes (volumen)	81,0	Ácido láctico (sangre total; arterial)	49,8
Eosinófilos (porcentaje)	81,0	Glucosa (a)	49,7
Eosinófilos (recuento)	81,0	Carboxihemoglobina (porcentaje; a)	49,6
Hematocrito	81,0	Hemoglobina total (hemoxímetro; a)	49,6
Hematíes (recuento)	81,0	Metahemoglobina (porcentaje; a)	49,6
Hemoglobina	81,0	Saturación de O2 (a)	49,6
Hemoglobina corpuscular media	81,0	Cloro (a)	49,5
Leucocitos (recuento)	81,0	Presión parcial de CO2 (v)	47,9
Linfocitos (porcentaje)	81,0	pH (v)	47,9
Linfocitos (recuento)	81,0	Presión parcial de O2 (v)	47,8
Monocitos (porcentaje)	81,0	Calcio iónico (v)	47,7
Monocitos (recuento)	81,0	Glucosa (v)	47,7
Neutrófilos (porcentaje)	81,0	Hematocrito (gasómetro; v)	47,7
Neutrófilos (recuento)	81,0	Potasio (v)	47,7
Plaquetas (recuento)	81,0	Sodio (v)	47,7
Volumen corpuscular medio	81,0	Ácido láctico (sangre total; venosa)	47,7
Volumen plaquetario medio	81,0	Hemoglobina total (hemoxímetro; v)	47,6
Creatinina	80,6	Carboxihemoglobina (porcentaje; v)	47,5
Sodio	80,6	Metahemoglobina (porcentaje; v)	47,5
Glucosa	80,4	Saturación de O2 (v)	47,5
Potasio	79,8	Cloro (v)	47,3
Urea	78,0	Lactato deshidrogenasa	40,7
Cloro	77,2	Creatina quinasa	31,8
Aspartato transaminasa	73,6	Albúmina	29,3
Alanina transaminasa	73,5	FIO2 (a)	28,8
Gamma glutamiltransferasa	72,0	Colesterol	27,3
Bilirrubina total	70,4	alfa-Amilasa	26,9
Fosfatasa alcalina	69,9	Acido úrico	26,8
Tiempo de protrombina (porcentaje)	67,6	Triglicéridos	26,4
Tiempo de protrombina (segundos)	67,6	Magnesio	24,0
Tiempo de protrombina normalizado (INR)	67,6	Bilirrubina directa	23,7
Tiempo de tromboplastina parcial activada (ratio)	67,6	Fósforo	23,1
Tiempo de tromboplastina parcial activada (segundos)	67,6	Fibrinógeno (coagulativo)	22,2
Proteína C reactiva	66,7	Bilirrubina indirecta	21,6
Proteínas totales	62,7	FIO2 (v)	21,3
Calcio	58,6	Procalcitonina	21,1
pH (a)	50,1	Ferritina	20,5
Presión parcial de CO2 (a)	50,0		

Cobertura: porcentaje de peticiones con al menos un valor conocido (el último valor medido en el ingreso, arrastrado). De cada prueba se construyeron cinco variables: último valor, cambio respecto a la medida anterior, cambio respecto a la primera medida del ingreso, mínimo y máximo. (a): arterial; (v): venosa.

TABLA A3

UTILIDAD POR UMBRAL DE AVISO (HORIZONTE DE 2 DÍAS)

Umbral	Peticiones con aviso (%)	Sens. (%)	VPP (%)	Falsos avisos por acierto	IRA avisadas en la ventana (%)	Ingresos sin IRA con aviso (%)
1 %	70,4	95,9	3,6	27,4	97,0	95,1
2 %	43,0	79,6	4,9	20,0	80,7	69,4
3 %	27,4	66,3	6,4	15,1	65,1	49,3
4 %	18,5	54,4	7,8	12,3	51,0	35,0
5 %	13,2	44,7	9,1	10,6	40,1	25,6
6 %	9,5	38,5	10,9	8,7	33,3	18,9
8 %	5,4	28,4	14,4	6,6	22,3	10,8
10 %	3,3	20,9	17,4	5,4	16,0	6,5
15 %	1,4	11,1	22,6	4,3	7,6	2,4

Media de las 15 repeticiones. Sens.: sensibilidad; VPP: valor predictivo positivo (ambos por petición). «IRA avisadas en la ventana»: ingresos con IRA con al menos un aviso en los 1–2 días previos.

TABLA A4

UTILIDAD POR UMBRAL DE AVISO (HORIZONTE DE 1 DÍA)

Umbral	Peticiones con aviso (%)	Sens. (%)	VPP (%)	Falsos avisos por acierto	IRA avisadas en la ventana (%)	Ingresos sin IRA con aviso (%)
0,5 %	65,9	95,7	2,0	51,5	94,0	88,8
1,0 %	38,5	79,5	2,8	36,2	73,8	61,1
1,5 %	24,9	68,3	3,8	27,3	59,7	42,5
2,0 %	17,5	59,0	4,7	22,4	48,6	31,3
3,0 %	9,7	45,5	6,4	15,9	34,4	19,2
4,0 %	6,1	35,1	7,9	12,6	25,1	12,7
5,0 %	4,2	29,5	9,7	10,6	19,9	8,8
8,0 %	2,0	20,1	14,0	7,3	11,3	3,9
10,0 %	1,4	16,0	16,3	6,6	8,1	2,7

Horizonte de 1 día: 760 peticiones positivas (1,34 %). Media de las 15 repeticiones; las probabilidades son más bajas que a 2 días, por lo que los umbrales se desplazan.

TABLA A5

AUROC Y AUPRC DE LIGHTGBM SEGÚN EL HORIZONTE DE PREDICCIÓN

Horizonte	Peticiones positivas	Prevalencia (%)	AUROC (IC 95 %)	AUPRC (IC 95 %)	Dif. AUROC con 2 días (IC 95 %)
1 día	760	1,34	0,803 (0,772–0,827)	0,108 (0,069–0,174)	+0,025 (+0,008 a +0,042)
2 días (principal)	1.483	2,61	0,777 (0,752–0,799)	0,122 (0,092–0,168)	Referencia
3 días	1.992	3,51	0,752 (0,726–0,774)	0,126 (0,101–0,163)	−0,025 (−0,036 a −0,015)
4 días	2.421	4,26	0,738 (0,712–0,761)	0,132 (0,110–0,166)	−0,039 (−0,054 a −0,024)

LightGBM con las 431 variables sobre 56.770 peticiones usables. IC 95 % por bootstrap por paciente (1.000 réplicas, mismas remuestras para todos los horizontes). El AUPRC no es comparable entre horizontes porque la prevalencia cambia.

TABLA A6

ANÁLISIS DE SENSIBILIDAD Y POR SUBGRUPOS (LIGHTGBM, CONJUNTO P, HORIZONTE DE 2 DÍAS)

Análisis	Peticiones	Positivas	AUROC (IC 95 %)
Modelo principal			
Etiqueta A (definición principal)	56.770	1.483	0,777 (0,752–0,799)
Definición alternativa de IRA			
Etiqueta B (criterio relativo con aumento ≥ 0,3 mg/dL)	57.742	1.355	0,776 (0,750–0,801)
Primera creatinina tardía y diagnóstico renal			
Sin ingresos con primera creatinina tardía (> 7 d)	48.635	1.451	0,755 (0,731–0,780)
Sexo			
Hombres	35.270	895	0,771 (0,741–0,799)
Mujeres	21.500	588	0,784 (0,747–0,815)
Edad			
< 50 años	12.722	159	0,796 (0,751–0,836)
50–64 años	20.925	419	0,784 (0,744–0,815)
65–74 años	14.336	503	0,744 (0,696–0,782)
≥ 75 años	8.787	402	0,710 (0,655–0,758)
Servicio solicitante (exploratorio)			
Anestesiología y Reanimación	8.054	483	0,828 (0,784–0,866)
Resto de servicios	48.716	1.000	0,739 (0,717–0,761)

IC 95 % por bootstrap por paciente (1.000 réplicas). Los IC de las exclusiones se recalcularon con una reimplementación del bootstrap, por lo que pueden diferir en unas milésimas de los del resto. Etiqueta B: recuentos con sus propias filas; el AUROC se calculó sobre las 56.770 peticiones comunes a ambas etiquetas, cada modelo frente a su etiqueta (diferencia con A: 0,000; IC 95 % −0,012 a +0,009; 500 réplicas). Los análisis por edad y servicio son exploratorios.

TABLA A7

AUROC POR SERVICIO Y UNIDAD SOLICITANTE (ANÁLISIS EXPLORATORIO, HORIZONTE DE 2 DÍAS)

Servicio / unidad	Peticiones	Positivas	AUROC (IC 95 %)
Anestesiología y Reanimación (total)	8.054	483	0,828 (0,784–0,866)
Anestesiología y Reanimación, rutina	6.494	369	0,847 (0,801–0,883)
Anestesiología y Reanimación, urgente	1.526	114	0,745 (0,695–0,788)
Resto de servicios	48.716	1.000	0,739 (0,717–0,761)
Aparato Digestivo, urgente	1.829	50	0,706 (0,601–0,801)
Cirugía General y Digestiva, urgente	2.858	85	0,700 (0,642–0,757)
Cirugía Cardiovascular, urgente	1.614	59	0,686 (0,618–0,751)
Hematología	3.003	114	0,664 (0,601–0,713)
Medicina Interna, urgente	2.375	59	0,593 (0,518–0,683)

Unidades con al menos 50 peticiones positivas. IC 95 % por bootstrap por paciente (500 réplicas).