Modelo de Inteligencia Artificial – Machine Learning explicable para la predicción de sepsis.

Miguel Angel Duarte Millán

Resumen — La sepsis es un problema sanitario que ocasiona alta morbi-mortalidad. La predicción de su desarrollo por técnicas de machine learning (ML) es prometedora, pero afronta aún problemas para la implantación generalizada. En este trabajo, creamos un modelo predictivo de desarrollo de sepsis, ingreso en cuidados intensivos (UCI) o mortalidad en las 48 horas siguientes al inicio de la atención en urgencias. El estudio se llevó a cabo sobre una población final de 51,681 episodios, con una prevalencia de la variable principal del 9.92%. Tras aplicar procesado de los datos, técnicas de imputación, y tratamiento del desbalanceo de los datos, varios modelos (random forest, XGBoost, red neuronal recurrente) fueron entrenados sobre 51 variables seleccionadas (edad, sexo, comorbilidades, constantes vitales y datos analíticos) con variaciones de los hiperparámetros. Se analizó el rendimiento mediante área bajo la curva (AUROC), área bajo la curva *precision-recall* (AUPRC), *recall* o sensibilidad (S), especificidad (E), *precision* o valor predictivo positivo (VPP) y F-1 score (F1). El mejor modelo considerado, obtuvo una AUROC 0.85, AUPRC 0.40, S 0.80, E 0.75, VPP 0.26, F1 0.40. Se aplicaron métodos para evaluar la importancia de las variables en las prediciones -como valores Shapley- y crear un módulo de predicción individualizado que sea clínicamente interpretable.

Palabras clave (entre 4 y 6). Machine Learning, Random Forest, XGBoost, SHAP, Sepsis.

Abstract— Sepsis remains a major healthcare challenge, characterized by high morbidity and mortality. Predicting its onset using machine learning (ML) techniques is promising but still faces obstacles to widespread implementation. In this study, we developed a predictive model for sepsis onset, intensive care unit (ICU) admission, or death within 48 hours of emergency department admission. We analyzed a final cohort of 51 681 episodes, in which the primary composite outcome occurred in 9.92 % of cases. After data preprocessing, imputation, and class-imbalance correction, we trained several models—random forest, XGBoost and a recurrent neural network—on 51 selected variables (including age, sex, comorbidities, vital signs and laboratory parameters), with systematic hyperparameter tuning. Model performance was evaluated by the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), recall or sensitivity, specificity, precision or positive predictive value (PPV) and F1-score. The best model achieved an AUROC of 0.85, AUPRC of 0.40, sensitivity of 0.80, specificity of 0.75, PPV of 0.26 and F1-score of 0.40. Finally, we applied feature-importance techniques -as Shapley values- in order to build a clinically interpretable, individualized prediction module.

Index Terms—Machine Learning, Random Forest, XGBoost, SHAP, Sepsis.

1 Introducción.

La sepsis es una disfunción orgánica secundaria a una infección (1), que provoca alta morbimortalidad. Se estima que provoca el 20% de las defunciones a nivel mundial (2) y que el 18% de los pacientes que la desarrollan fallecerán (3). La identificación de pacientes en riesgo es esencial, ya que por cada hora de retraso en la prescripción de antibiótico la mortalidad aumenta (4). Sin embargo, la predicción individual puede ser difícil, debido a la heterogeneidad en la presentación (5). Algunas escalas -siendo la más utilizada *sepsis organic failure assesment* (SOFA) – se diseñaron como herramientas de screening, pero tienen ciertas limitaciones en la detección precoz o en generalización fuera de cuidados intensivos (UCI) (6).

Las tareas de predicción en medicina cuentan con un nuevo aliado. Los algoritmos de *machine learning* (ML), con su capacidad para integrar grandes cantidades de datos procedentes de diferentes fuentes, se postulan para superar las limitaciones previas (7). La tarea de predicción de sepsis, por su relevancia, ha sido un objeto de interés creciente (8). En este artículo, proponemos un modelo de ML diseñado para predecir el riesgo de desarrollar un evento compuesto por sepsis, ingreso en UCI, o fallecimiento. Además, realizamos una revisión narrativa sobre el tema y abordamos las dificultades a superar para la generalización del uso de estos modelos.

2 Estado actual del tema.

Aunque el uso de ML en la predicción de sepsis es indudablemente emergente, con un meta-análisis reciente que identificó 28 artículos centrados en esta tarea (8), muchos aspectos deben ser revisados antes de considerar su implantación. Una revisión de estos ítems a través de algunos estudios relevantes en la literatura se muestra en la Tabla 1, que discutiremos a continuación.

- *E-mail de contacto: duartemillan.miguelangel@gmail.com*
- *Trabajo tutorizado por: Antonio Lozano Bagen*
- *Curso: 2025*

TABLA 1
ASPECTOS DESTACADOS DE MODELOS PREVIOS.

Autor / Año	Población (n)	Outcome	Sepsis (%)	Tiempo predicción	Predictores	Procesamiento	Modelo predictivo utilizado	Métricas rendimiento.
Calvert et al. – 2016	MIMIC - II. n = 1,394.	Codificación CIE-9: 995.9. Criterios SIRS.	11.4	3 h.	9 (constantes y laboratorio)		InSight: riesgo como sumatorio ponderado de indicadores.	- AUROC 0.92 (0.86–0.93). - S 0.90 (0.89–0.91). - E 0.81 (0.80–0.82).
Nemati et al. – 2018	UCI. n = 27,527.	Sepsis-3 (SOFA ≥2).	8.6	12 h.	65 (constantes, laboratorio, demográficas).		Modified Weibull-Cox proportional hazards.	- AUROC 0.85 E 0.67 (S fijada 0.85) (4h) - AUROC 0.83 E 0.63 (S fijada 0.85). (4h)
Mao et al. – 2018	UCSF. n = 90,353	Codificación CIE-9: 995.9. Criterios SIRS.	1.3	0 - 4 h.	6 variables clínicas.	Missings: imputación "carry-forward".	GTB.	- AUROC 0.92. S 0.98 (E fijada 0.80) (0h)
Scherpf et al. – 2019	MIMIC - III. n = 30,000	Codificación CIE-9: 995.9. Criterios SIRS.	0.3 - 1.	3/6/12 h.		Missings: "carry forward/backward".	RNN.	- AUROC 0.76 (0.73 - 0.79) - E 38.8 (31.7 - 45.8) (S fijada 0.90). (0h)
Lauritsen et al. – 2020	Hosp. n = 163,050	Sepsis-3 (SOFA ≥2).	2.4	0/3/6/12/ 24 h.	33: 27 laboratorio + 6 constantes.	Miissings: "carry-forward". Desbalanceo: oversampling.	RNC Módulo explicativo Deep Taylor decomposition.	- AUROC 0.92 (0.9–0.95) (0h). - AUPRC 0.43 (0.36–0.51) (0h) - AUROC 0.8 (0.78–83) (24h) - AUPRC 0.08 (0.07-0.09) (24h)
Goh et al. – 2021	UCI. n = 3,900	Codificación CIE-10.	6.2	4/6/12/ 24/48 h.	Datos estructurados (vitales, laboratorio, tratamientos) y no estructurados (NPL).	Desbalanceo: SMOTE.	RL + RF	- AUROC 0.94, S 0.89, E 0.85, VPP 0.85. (0h) - AUROC 0.94 (12h) / AUROC 0.90 (24h) / AUROC 0.87 (48h).
Persson et al. – 2021	MIMIC- III. n = 7681	Sepsis-3 (SOFA ≥2).	18	0 - 3 h.	20 (clínicas y laboratorio).	Missings: "carry-forward". Desbalanceo: Sampling hasta sepsis = 20%.	RNC	- AUROC 0.84. AUPRC 0.68. S 0.74 E 0.83 VPP 0.5 (3h).
Rosnati et al. 2021	MIMIC-III. n = 58,976	Sepsis-3 (SOFA ≥2).		0 - 5 h.	Selección variables según % válidos.	Missings: Imputación a valores no patológicos. Desbalanceo: Oversampling1:1.	Multi modelo : TCN, red feedforward, RL.	- AUROC 0.66 AUPRC 0.48 (5h).
Shashikumar et al. – 2021	UCI y urgencias. n = 515,720	Sepsis-3 (SOFA ≥2).			34 clínicas - y derivadas temporales- y 6 demográficas.	Missings: Imputacion mediana.	COMPOSER: 3 módulos (RN + RL)	- AUROC 0.925 VPP 0.24 (Val. ext. UCI) - AUROC 0.938 VPP 0.13 (Val. Ext. Urg)
Bhargava et al. – 2024	Hosp. n = 3457	Sepsis-3 (SOFA ≥2) en 24 h.	32.2	24 h.	22 clínicas.	Missings: Imputación con bagged trees.	RF	- AUROC 0.84 (0.78 - 0.89) (Diagnostico) - AUROC 0.76 (0.68 - 0.84) (Predicción)
Zhou et al. – 2024.	UCI. n = 2385	Sepsis-3 (SOFA ≥2).	15.3	24 h.	18 (selección algoritmo a priori)	Missings: Imputación múltiple y media. Desbalanceo: SMOTE.	RF (comparado con RL, XGBoost, RN)	- AUROC 0.87 F1-score 0.77 S = 0.66 VPP 0.88.
Liu et al. – 2025.	UCI. n = 2329	Sepsis-3 (SOFA ≥2).	10.22		36 -> 27 por test univariado -> 13 por recursive feature elimination (SVM).	Missings: Imputación KNN.	RF (comparado con RL, RF, MLP y LightGBM).	- AUROC 0.82, F1-score 0.38, S 0.75, E 0.75, ACC 0.75.

Tabla 1. Aspectos destacados de modelos previos. Abreviaturas: UCI: unidad de cuidados intensivos; Hosp: hospitalización; Urg: urgencias. SOFA: 'sepsis- organic failure assesment'; CIE : codificación internacional de enfermedades, GTB : 'gradient tree boost'; RNN : red neuronal recurrente; RNC : red neuronal convolucional; RL : regresión logística; RF : random forest; TCN : red convolucional temporal.

2.1 Procedencia de los datos.

Más del 30% de los modelos publicados, proceden del uso de la base de datos MIMIC-III(8), que contiene la información de pacientes ingresados en la UCI de un hospital norteamericano (9–12). A partir de este dataset, Calvert et al. desarrollaron uno de los primeros modelos publicados, que predice sepsis 3 horas antes del inicio, a partir de las tendencias temporales de 9 parámetros vitales, con una AUROC de 0.92, S de 0.90 y E de 0.80 (9). Estos datos han sido posteriormente utilizados por muchos autores para optimizar los modelos o como cohortes de validación externa (13,14). Sin embargo, el uso de estos datasets públicos plantea dos cuestiones: si son aplicables a otros entornos como urgencias u hospitalización, o cómo responden en entornos menos monitorizados y con menos datos disponibles. Por ello, son de interés los modelos desarrollados en otros ámbitos y cohortes locales, más asimilables al concepto de *"real-world data"*.

2.2 Definición de objetivos.

A partir de la aparición del consenso Sepsis-3, la mayoría de los estudios adoptan estos criterios como base para la inclusión y desenlace (1). La sospecha de infección se establece como la solicitud de una prueba microbiológica y la prescripción de un antibiótico, con unas determinadas reglas temporales. Algunos autores has utilizado modificaciones en la inclusión, la solicitud de hemocultivos (15), o utilizar sólo casos con cultivos bacterianos positivos (16). El desarrollo de sepsis se define como un aumento ≥2 puntos en SOFA. Esto facilita identificar el momento exacto del inicio ("sepsis-onset") y establecer horizontes predictivos. Algunos modelos (9,11,17) predicen a corto plazo (<6 horas), mientras que otros (13,18) amplían el horizonte hasta 12-24 horas. En general, el rendimiento decrece conforme aumenta el tiempo predictivo (19). En el ejemplo de Lauritsen et al., se puede ver como la AUROC baja de 0.92 en el tiempo de "sepsis-onset" a 0.80 en la predicción a 24 horas(18). Goh et al. consiguen una meritoria AUROC de 0.87 en una predicción a 48 horas (20).

2.3 Detección o predicción de sepsis.

Un aspecto clave es cómo abordar a pacientes con SOFA ≥2 al ingreso, situación que darse en situaciones de sepsis ya instaurada, o por presencia de comorbilidades previas. Muchos autores excluyen estos casos, centrando la tarea del modelo únicamente en una tares de predicción del desarrollo de sepsis, mientras que otros incluyen a todos los pacientes, y el modelo ejerce como una herramienta de detección (13,17,21). Es el caso de Bhargava et al., aunque posteriormente diferencia el rendimiento del modelo para todos los casos -detección (AUROC 0.84)- o excluyendo a los que presentan la variable desenlace desde el inicio -predicción (AUROC 0.76)- (15).

2.4 Variables elegidas para el entrenamiento.

El número de parámetros utilizados para entrenamiento varía considerablemente. Mao et al. utilizan un modelo basado únicamente en 6 signos vitales medidos cada hora, con una AUROC de 0.92 (17). Por otra parte, algunos modelos como el de Shashikumar et al. utiliza más de 100 variables, incluyendo sus derivadas temporales (21). En algunos casos, se muestra la interesante opción de aplicar métodos de preselección de variables (contrastes univariados, algoritmo a priori) (14,16). Aunque la mayoría son datos estructurados, Goh et al. incorporan procesamiento de datos no estructurados para mejorar su rendimiento, y encuentran que a mayor distancia del momento predictivo, mayor relevancia tienen estos datos, suponiendo su aplicación mejoras en AUROC (0.77 a 0.87) y S (0.71 a 0.78) para predicciones a 48 h (20).

Un aspecto a considerar cuando se emplea el incremento de SOFA para definir el desarrollo de sepsis es que su cálculo se realiza a partir de parámetros que, si se emplean como variables predictoras del modelo, suponen un riesgo de "data leakage". Mao et al. demostraron que al eliminar dichos parámetros el AUROC de su modelo pasa de 0.92 a 0.84 (17). En el estudio de Zhou et al. se muestra en los gráficos de dependencia parcial como el valor de SOFA, incluido en el modelo como predictor, tiene una correlación lineal perfecta con el peso de las predicciones (14).

2.5 Valores perdidos y desbalanceo de clases.

El tratamiento de valores perdidos puede variar según el entorno. En ámbitos altamente monitorizados (UCI), prevalecen métodos como "carry-forward" del último valor conocido. En contextos con más valores perdidos, se han utilizado con éxito métodos como imputación por media, mediana, bagged trees o K-nearest neighbors (KNN). Un estudio, por ejemplo, informa de que la imputación múltiple mejora el AUROC de 0,81 a 0,84 (14).

Un problema relevante en modelos de ML para predecir sepsis es su baja prevalencia, lo que genera un marcado desbalanceo en las clases y afecta las métricas. Además, existe una alta variabilidad entre las series (ver Tabla 1, donde la prevalencia varía entre el 0.3 y 35.2 % (15,18). Las técnicas de corrección empleadas incluyen ajuste de prevalencia(11), oversampling con relación 1:1 (12) o SMOTE (14,20). Goh et al. reportan una mejora del valor predictivo positivo (VPP) de 0.04 a 0.77 a 48 h tras aplicar SMOTE (20). Zhou et al. obtienen una mejora en AUROC de 0.77 a 0.84 con la misma técnica (14).

2.6 Elección del modelo.

Diversos algoritmos han sido utilizados para esta tarea: random Forest (RF), regresión logística (RL), máquinas de soporte vectorial (SVM), Gradient Boosting (GB) o XGBoost (XGB), redes neuronales recurrentes (RNN) y redes convolucionales temporales (TCN) (8). En los primeros momentos, RF y GB fueron de los primeros modelos utilizados (17). Posteriormente, se generalizó el uso de RNN. Las arquitecturas más recientes incorporan redes convolucionales temporales (TCN) que permiten capturar la evolución dinámica del riesgo y hacer predicciones a tiempo real(18). Algunos autores han realizado comparaciones de varios modelos con los mismos datos, concluyendo que RF presenta mejor equilibrio entre rendimiento y simplicidad(14,16).

2.7 Métricas de rendimiento.

Aunque la mayoría de estudios informan la AUROC como métrica principal, este parámetro puede no reflejar adecuadamente el rendimiento del modelo, cuando la clase positiva (sepsis) es minoritaria y existe un problema de desbalanceo. Lauritsen et al, cuyo modelo tiene un AUROC elevado (0.92 para predicción a 0h y 0.80 a 24h), destacan la dificultad de trabajar con conjuntos de datos altamente desbalanceados (su prevalencia de sepsis es del 2.44%). Proponen como una métrica más informativa la curva *precision-recall* (AUPRC), que establece el compromiso entre *recall* (equivalente a sensibilidad) y *precision* (valor predictivo positivo). Reportan una AUPRC de 0.43 en predicciones a 0h y tan baja como 0.08 a 24h (18).

Esta dificultad para lograr un buen compromiso *recall-precision* es constante. Persson et al. obtienen una de las mejores AUPRC reportadas, de 0.68, pero en un intervalo de predicción cercano al inicio de la sepsis (0–3 h) (11). Esta relación compleja queda patente en estudios como el de Shashikumar et al., que a pesar de AUROC superiores a 0.90 incluso en validación externa, reporta una *precision* de solo 0.24 en UCI y 0.13 en urgencias, lo que supone una alta tasa de falsos positivos(21).

Recall y *precision* también dependen de los umbrales de predicción elegidos por los investigadores. Una opción mayoritaria es fijar la S en torno a 0.80-0.90. Por ejemplo, Nemati et al. obtienen una E de 0.63 con S fijada a 0.8, mientras que Scherpf et al. obtienen una E de 0.39 para una S de 0.90 (10)(13).

Métricas como el *precision* y AUPRC dependen directamente de la prevalencia de la sepsis, por lo que pueden mejorar tras corregir el desbalanceo de clases. Zhou et al. logran un buen equilibrio en su modelo entre S 0,74 y VPP 0,74 en una cohorte externa tras aplicar SMOTE (14).

2.8 Problemas de generalización.

Un problema significativo en la aplicación clínica de modelos de ML es su limitada capacidad de generalización entre diferentes escenarios asistenciales. Calvert et al. desarrollaron el modelo Insight con un AUROC de 0.92 en predicciones a 3h; sin embargo, cuando Scherpf compara este modelo con una RNN encuentra que, en sus datos, el rendimiento de Insight disminuye a un AUROC de 0.72 en predicciones para el mismo intervalo. Además, Insight originalmente reporta una S de 0.90 y E de 0.80, pero cuando Scherpf decide fijar la S en 0.90, la E de Insight cae a 0.24-0.34 (9,10).

De hecho, algún estudio ha señalado incluso la discrepancia en resultados de prevalencia de sepsis utilizando

los mismos criterios Sepsis-3 sobre los mismos datasets (MIMIC-III). Utilizando diferentes métodos de imputación y tratamiento de los valores perdidos, Rosnati et al. identifican hasta cuatro veces más casos de sepsis comparado con un estudio sobre el mismo conjunto de datos. Sobre dicha disparidad para definir los casos de sepsis, no es extraño que al aplicar Insight en sus datos, reporten un rendimiento considerablemente inferior (AUROC de 0.49 y AUPRC de 0.35) al de desarrollo original (12).

2.9 Explicabilidad.

Para contrarrestar la percepción de "caja negra" asociada a los algoritmos de ML, recientemente se han incorporado métodos de explicabilidad. Estos métodos permiten desglosar y entender la contribución de las variables en la predicción individual y global. Diversos métodos explicativos, destacando por su capacidad ilustrativa los valores Shapley (SHAP) o métodos como Deep Taylor decomposition que ilustra como el peso de las variables se modifica en la predicción a lo largo del tiempo, se han utilizado para mejorar la confiabilidad clínica del modelo (15,18).

A pesar de las limitaciones mencionadas, algunos estudios han evaluado el impacto clínico real de sistemas predictivos basados en ML. Burdick et al. reportan unos resultados optimistas, en los que la implementación de un modelo basado en XGBoost con siete variables logró reducir la mortalidad del 3.86% al 2.34%, lo que representa una reducción relativa del 39.5% (22).

2.9 Perspectivas del presente trabajo.

En resumen, los algoritmos de ML para predicción de sepsis se postulan como útiles. Sin embargo, es importante contextualizar el ámbito de uso, el número de variables requeridas para su rendimiento, el tiempo al que funcionan las predicciones y el tratamiento del problema del disbalanceo de clases. Más allá de la AUROC, se debe ser exigente al interpretar otras métricas de rendimiento como la relación entre *recall* y *precision*, para hacer el modelo útil en la práctica. Los módulos de explicabilidad, parecen necesarios para entender las predicciones individualizadas que puedan ser interpretadas por los clínicos.

En dicho contexto, en nuestro trabajo se pretende desarrollar un modelo que se centre en la tarea de predicción de un compuesto de pronóstico desfavorable: sepsis, UCI y mortalidad; que se desarrolle a lo largo de un intervalo de 48 h desde el inicio de la atención, utilizando para la predicción los primeros datos disponibles en el propio servicio de urgencias. Además, se pretende explorar la aplicación de un módulo de explicabilidad que ayude a la comprensión racional de las predicciones.

3 MATERIAL Y MÉTODOS.

3.1 Objetivo principal. Definiciones.

Predicción de una variable combinada de desarrollo de sepsis, ingreso en UCI, o mortalidad, en pacientes con sospecha de infección, en las 48 h posteriores al inicio de la atención en urgencias.

Se define desarrollo de sepsis como un SOFA ≥ 2, en pacientes que no lo presentan inicialmente. La creación de una variable combinada con UCI y mortalidad suple la ausencia de datos respecto a algunos elementos de la escala SOFA (como uso de ventilación mecánica o soporte vasoactivo). En nuestro entorno, el ingreso en UCI se considera un equivalente a estos ítems, por lo que se decidió combinar, junto a la mortalidad, para conformar un modelo que informe sobre riesgo de pronóstico desfavorable.

3.2 Población a estudio. Recogida de datos.

Se diseñó un estudio retrospectivo, con datos extraídos de la historia clínica electrónica (HCE) de episodios atendidos en las urgencias del Hospital Universitario de Fuenlabrada, en el periodo comprendido entre 1 de Enero de 2016 hasta el 31 de Diciembre de 2024.

Los criterios de inclusión, establecen un marco de sospecha de infección modificado respecto a los criterios Sepsis-3 (1).

a) Episodios de pacientes > 18 años atendidos en el servicio de urgencias del HUF, con sospecha de infección.

b) Sospecha de infección: se define como prescripción de un antibiótico en las primeras 24 h de atención.

Se eliminar el requisito de solicitud de pruebas microbiológicas. Esta decisión se toma para favorecer la reproducibilidad del modelo, según la práctica observada en nuestro medio de pacientes con sospecha de infección sin solicitud microbiológica asociada.

Para la realización del estudio, se extrajeron de la HCE datos demográficos, comorbilidades (basadas en la codificación CIE-10), datos de laboratorio y constantes vitales registradas en formularios.

3.3 Procedimientos estadísticos.

Los datos fueron extraídos de un repositorio central del HUF a través de consultas SQL, tras un proceso de selección de las variables requeridas. Para el procesamiento de datos, se ha utilizado lenguaje Python (v 3.11.10) y el software Spyder 6. Varias librerías se han utilizado en el estudio: pandas (v 2.2.3) y numpy (2.1.2) para la preparación de los datos; scipy (v 1.14.1) para los procedimientos estadísticos; sklearn (v 1.6.1), xgboost (v 3.0.2) y tensorflow (v 2.19.0) para los modelos de ML; matplotlib (v 3.9.2) para gráficos; eli5 (v 0.16.0) y shap (v 0.48.0) para los módulos de explicabilidad.

En la descripción de variables continuas se empleó la media y la desviación estándar (SD), y se compararon a través del test t de Student para muestras independientes. Las variables categóricas se describieron por porcentajes, y se compararon con los test estadísticos de Chi-cuadrado o test exacto de Fisher cuando correspondía.

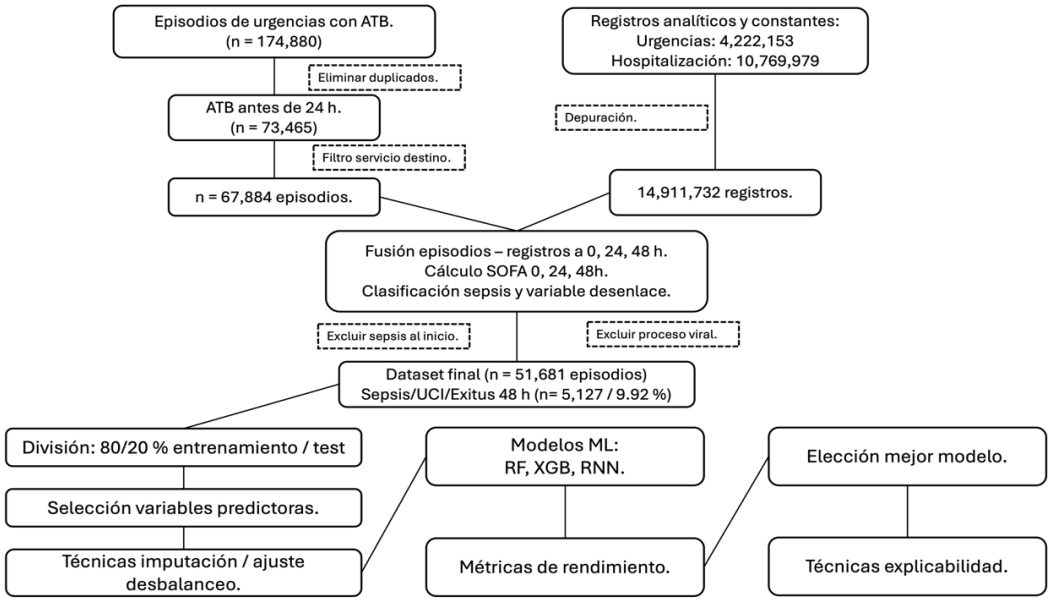

Fig. 1. Flujo de pacientes y procedimientos. Abreviaturas: ATB: antibiótico; SOFA: sepsis organic failure assessment; RF: random forest, XGB: xgboost; RNN: red neuronal recurrente.

3.4 Preparación de los datos.

3.4.1 Flujo de pacientes.

El flujo total de pacientes y datos filtrados a lo largo del proceso, se muestra en la Figura 1. Inicialmente, 174,880 episodios de urgencias fueron identificados como procesos susceptibles de sospecha de infección en el periodo establecido (2016-2024). Se eliminaron duplicidades y el conjunto se redujo a sólo antibióticos prescritos en las primeras 24 horas. A criterio de los investigadores, se excluyeron pacientes que ingresaban en servicios que habitualmente no tratan infecciones (p ej. algunos servicios quirúrgicos y obstétricos, etc.), para evitar que prescripciones de antibiótico como profilaxis prequirúrgicas, incluyeran procesos sin sospecha de infección. Tras la asignación de valores SOFA se descartaron los pacientes con sepsis presente al inicio, y aquellos que tienen una codificación de proceso viral (COVID19 o gripe) por CIE-10, por considerarse confusores respecto al proceso de sepsis bacteriana. El dataset final contenía 51,681 episodios.

3.4.2 Procesado de variables clínicas y de laboratorio.

Para todos los episodios, se obtuvieron los datos analíticos y constantes vitales asociadas, de urgencias y de hospitalización (4,222,153 y 10,769,979 registros distintos en cada área, respectivamente). Se redefinieron variables que aludían a los mismos parámetros y se eliminaron aquellos indicadores con poca representación o considerados irrelevantes. Tras la supresión, se mantuvieron un total de 14,911,732 registros. Una exploración de datos preliminar detectó valores anómalos. Se realizó una exclusión de dichos valores, según unos rangos definidos por los investigadores (disponibles en Anexo I).

Posteriormente, se ejecutó una fusión de datos, de tal forma que por cada episodio y variable, se ejecutó una función que buscaba el valor más cercano disponible a los puntos temporales de 0h, 24h y 48h (con intervalo de ± 12h) considerando el inicio de atención en urgencias como 0h. El número de registros válidos por cada variable en dichos puntos temporales, se muestra en la Tabla 2.

3.4.3 Cálculo de SOFA y variables desenlace.

Se calcularon para cada episodio los valores SOFA a las 0, 24 y 48 h. La saturación de oxígeno periférica (SpO2) era un parámetro más disponible que la saturación por gasometría, por lo que para el SOFA respiratorio se utilizó el cociente SpO2/FiO2. La FiO2 se calculó mediante una fórmula FiO2 = 0,21 + (4 * Flujo O2). Se determinaron reglas para ajustar el flujo de O2 y el tipo de dispositivo de oxigenoterapia utilizado. Como se mencionó, la ausencia de datos sobre si se precisó ventilación mecánica o soporte vasoactivo, impedía que hubiera pacientes con puntuación de SOFA respiratorio mayor de 2, o de SOFA hemodinámico mayor de 1. Esta carencia se suple al incorporar como casos a los pacientes con ingreso en UCI, asumiendo la equivalencia a sepsis.

Se calculó el SOFA total para episodios con al menos un valor válido de alguno de sus componentes. Al igual que Rosnati et al (12), para pacientes sin datos de SOFA pero que seguían activos, se determinó una imputación al último valor disponible (estrategia "carry-forward"), asumiendo que en la práctica, el deterioro clínico suele asociar nuevos datos capturados, mientras que la ausencia de datos es más probable en un escenario de estabilidad clínica.

A los pacientes con SOFA ≥ 2 a las 0 h (11.76 % del total), se les consideró pacientes con "sepsis presente al inicio" y se excluyeron. El resto (SOFA < 2 en la primera valoración) fueron considerados los pacientes subsidiarios de predicción del desenlace principal. En ellos, se etiquetó como positivo el desenlace en caso de presentar durante las 48 h siguientes un SOFA ≥ 2, ingreso en UCI o fallecimiento.

TABLA 2
ANÁLISIS DE VALORES VÁLIDOS.

Variable / % valores válidos.	0 h	24 h	48 h
Constantes			
Temperatura	85	91	94
TA sistólica	97	98	98
TA diastólica	97	98	98
Frecuencia cardiaca	97	98	98
Frecuencia respiratoria	4	5	1
Sat O2 per. (%)	69	87	93
Flujo O2 (l/min)	59	69	63
Puntos Glasgow	2	1	0
Laboratorio			
Act. Protrombina (%)	71	22	21
Albumina	11	19	35
ALT/GPT	40	27	37
AST/GOT	7	5	9
Bilirrubina total	27	21	31
Bicarbonato art.	16	4	3
Bicarbonato ven.	40	19	12
Creatinina	76	44	44
Fibrinogeno	71	21	20
GGT	39	18	17
Glucosa	92	48	42
Hemoglobina	95	50	44
Linfocitos	78	42	41
Lactato art.	16	4	3
Lactato ven.	35	17	11
Leucocitos	94	50	44
Neutrofilos	79	43	42
Potasio	75	44	43
Proteina c reactiva	77	40	38
Procalcitonina	15	6	7
Plaquetas	94	49	44
Sodio	89	48	43
Saturación O2 art.	16	4	3
Saturacion O2 ven.	39	19	12
TTPA	60	19	19
Urea	87	48	41
pcO2 art.	20	5	3
pcO2 ven.	40	19	12
pH art.	20	5	3
pH ven.	40	19	12
pO2 art.	19	5	3
pO2 ven.	40	19	12

Tabla 2. Mapa de calor sobre la distribución de los valores válidos por cada variable e intervalos temporales. Se mapean los valores según 4 etapas: <8%; 8-25%; 25-75%; >75%.

3.5 MACHINE LEARNING.

3.5.1 Selección de variables de entrenamiento.

Para el entrenamiento de los modelos, se configuró un conjunto de variables que incluían edad, sexo, un listado de comorbilidades previas y el conjunto de constantes vitales y valores de laboratorio asignados al punto temporal 0 h. Se eliminaron las comorbilidades que presentaban prevalencias inferiores al 1 %, así como las variables clínicas que tenían un porcentaje inferior al 10 % de valores válidos a las 0 h. El total de variables seleccionadas para el entrenamiento fue 51. El listado se encuentra en el Anexo I.

3.5.2 Imputación y desbalanceo.

A fin de conseguir el modelo más óptimo para la tarea de predicción, se probaron diferentes técnicas de tratamiento de valores perdidos, de ajuste del desbalanceo de clases, y se entrenaron varios tipos de algoritmo predictivo (RF, XGBoost, RNN) con variaciones de los hiperparámetros, y un ajuste de umbrales para la toma de decisiones del modelo. Se realizó una división entre cohorte de entrenamiento (80% de los datos) y de test (20%), por división aleatoria, para el desarrollo de los modelos.

Para el tratamiento de valores perdidos, se probó sin imputar ningún valor (cuando el modelo lo permitía), con métodos de imputación por mediana, o por *knn – neighbours*. Para el desbalanceo de clases en la variable de desenlace, se ensayaron modelos sin ningún tipo de ajuste, y otros donde se introdujo en la configuración del modelo un ajuste de pesos ("*class weight = balanced*"). Además, se ensayaron técnicas de oversampling (duplicando los casos, creando una relación 1:1, mediante el método SMOTE) y de undersampling (relación 1:1). Sobre estos modelos se ejecutó una iteración de distintos parámetros, a fin de evaluar los cambios en las métricas de rendimiento.

Para RF, se diseñó una iteración para 200 o 300 árboles, con profundidad 8,10 o 12. Para XGBoost se diseñó una iteración de 50 o 100 árboles, con profundidad de 4 o 6 con una *learning rate* fijada en 0.1. Para RNN se ejecutaron diseños con 3 o 6 capas ocultas, a intervalos de 10, 20,50 o 100 neuronas por capa, con optimizador '*adam*' y función de coste '*binary cross-entropy*'.

3.5.3 Definición métricas de rendimiento.

Por cada modelo entrenado y evaluado se obtuvieron AUROC, AUPRC, *accuracy* global (ACC), recall (S), precision (VPP), especificidad (E) y F1-score, por cada configuración del modelo y por cada ajuste del umbral de decisión. De cara a elegir el modelo más adecuado, se estableció un objetivo de $S > 0.80$. Entre los modelos que cumplian dicho criterio, se valoró aquel con AUPRCy F1-score más altos, en búsqueda del mejor compromiso posible entre S y VPP.

3.5.4 Módulos de explicabilidad.

Para el mejor modelo considerado de RF y XGBoost, se ejecutaron técnicas de explicabilidad para evaluar el peso de las variables en el entrenamiento del modelo. Se ejecutó una clasificación de importancia según el Índice Gini, por

importancia de permutación, y se calcularon los valores de *shapley additive explanation* (SHAP) de forma global. Además, se configuró un módulo de predicción individual, basada en los valores SHAP, que permitía un análisis caso a caso.

3.6 Aspectos éticos y de confidencialidad.

Este es un estudio retrospectivo sobre información clínica que no introduce modificación asistencial ni riesgo en los pacientes cuya información se analiza. Los datos utilizados han sido anonimizados con la exclusión de identificadores, lo que exime la necesidad de consentimiento informado. La información extraída cuenta con el permiso de la dirección del centro. Todos los investigadores han firmado el compromiso de acceso responsable a los datos clínicos. El estudio cuenta con la aprobación del Comité Ético de Investigación del HUF con el código 25/28, otorgado el 20/05/2025.

4 RESULTADOS.

4.1 Análisis descriptivo del dataset final.

Tras excluir los pacientes con sepsis de inicio y con codificación para un proceso viral, el dataset final contenía 51,681 episodios, de los cuales 30,515 (59.04%) derivaron en hospitalización. Se produjo el fallecimiento en el 3.04% (el 0.32% en las primeras 48 h). El 1.14% de los episodios acabaron en UCI, de los cuales el 57,9% ingresaron en las primeras 48 h. A las 48 h, el 9.92 % había cumplido en algún momento la variable compuesta de SOFA ≥ 2, ingreso en UCI o fallecimiento.

Se llevó a cabo un análisis comparativo de las características demográficas, comorbilidades, y valores clínicos y analíticos a tiempo 0h, entre los pacientes que no desarrollarán la variable combinada, y los que sí. La comparación se muestra en la Tabla 3. En resumen, los pacientes que desarrollan complicaciones son mayores, con más prevalencia de varones y con más comorbilidades (esencialmente diabetes, enfermedad renal crónica, oncológica, EPOC, cardiopatía isquémica, o insuficiencia cardiaca, con prevalencias superiores al doble en todas ellas). En los valores analíticos y constantes a tiempo 0 h, se observan diferencias significativas en casi todas las variables, excepto en temperatura, fibrinógeno, y pO2 arterial.

Los pacientes que desarrollan sepsis, tienen una estancia media de 9.6 ± 9.6 días, más del doble que los que no la desarrollan (4.4 ± 6.2 días).

TABLA 3
ANÁLISIS COMPARATIVO ENTRE GRUPOS.

Variable	Sepsis No (n=46554)	Sepsis Sí (n=5127)	p-valor
Edad (años)	62.7 ± 20.5	74.6 ± 16.0	< 0.05
Sexo (H), n (%)	22227 (47.7 %)	2836 (55.3 %)	< 0.05
Comorbilidades, n (%)			
Cardiopatia isquemica,	1535 (3.3 %)	514 (10.0 %)	< 0.05
Insuficiencia cardiaca,	2557 (5.5 %)	1037 (20.2 %)	< 0.05
Arteropatia perifercia,	355 (0.8 %)	107 (2.1 %)	< 0.05
Epoc,	2971 (6.4 %)	758 (14.8 %)	< 0.05
Asma,	2522 (5.4 %)	403 (7.9 %)	< 0.05
Demencia,	1240 (2.7 %)	262 (5.1 %)	< 0.05
Alzheimer,	529 (1.1 %)	115 (2.2 %)	< 0.05
Enf oncologica,	3248 (7.0 %)	790 (15.4 %)	< 0.05
Enf hematologica,	439 (0.9 %)	170 (3.3 %)	< 0.05
Enf autoinmune,	649 (1.4 %)	150 (2.9 %)	< 0.05
Hipertension,	9521 (20.5 %)	1629 (31.8 %)	< 0.05
Diabetes,	6380 (13.7 %)	1598 (31.2 %)	< 0.05
Cirrosis,	74 (0.2 %)	45 (0.9 %)	< 0.05
Enf renal cronica,	2808 (6.0 %)	1240 (24.2 %)	< 0.05
Vih,	95 (0.2 %)	19 (0.4 %)	< 0.05
Ictus,	151 (0.3 %)	33 (0.6 %)	< 0.05
Constantes (media, sd)			
Temperatura	36.65 ± 0.77	36.65 ± 0.84	0.8065
TA sistólica	75.59 ± 12.55	72.56 ± 13.32	< 0.05
TA diastólica	130.45 ± 23.45	128.70 ± 25.80	< 0.05
Frecuencia cardiaca	92.63 ± 19.51	95.36 ± 21.64	< 0.05
Frecuencia respiratoria	20.18 ± 5.57	23.38 ± 7.28	< 0.05
Sat O2 per. (%)	94.84 ± 3.17	93.03 ± 4.54	< 0.05
Flujo O2 (l/min)	0.41 ± 1.16	1.07 ± 2.13	< 0.05
Puntos Glasgow	14.84 ± 0.47	14.63 ± 0.65	< 0.05
Valores analíticos (media, sd)			
Act. Protrombina (%)	80.15 ± 20.54	73.25 ± 23.09	< 0.05
Albumina	3.84 ± 0.49	3.63 ± 0.53	< 0.05
ALT/GPT	31.63 ± 71.55	55.45 ± 146.99	< 0.05
AST/GOT	35.19 ± 40.11	58.45 ± 166.50	< 0.05
Bilirrubina total	0.66 ± 0.32	0.80 ± 0.40	< 0.05
Bicarbonato art.	26.05 ± 4.46	26.62 ± 5.96	< 0.05
Bicarbonato ven.	26.37 ± 3.89	26.10 ± 5.47	< 0.05
Creatinina	0.89 ± 0.27	1.06 ± 0.37	< 0.05
Fibrinogeno	637.47 ± 196.06	635.21 ± 199.90	0.4897
GGT	74.23 ± 157.22	133.10 ± 272.48	< 0.05
Glucosa	127.32 ± 57.09	151.91 ± 76.39	< 0.05
Hemoglobina	13.42 ± 1.91	12.88 ± 2.30	< 0.05
Linfocitos	1.63 ± 1.02	1.30 ± 0.93	< 0.05
Lactato art.	1.69 ± 0.90	1.95 ± 1.15	< 0.05
Lactato ven.	2.02 ± 1.01	2.51 ± 1.59	< 0.05
Leucocitos	11.45 ± 4.92	11.65 ± 5.77	< 0.05
Neutrofilos	8.71 ± 4.49	9.21 ± 5.02	< 0.05
Potasio	4.06 ± 0.51	4.18 ± 0.61	< 0.05
Proteina c reactiva	7.49 ± 8.34	9.36 ± 9.75	< 0.05
Procalcitonina	0.70 ± 3.54	2.02 ± 8.01	< 0.05
Plaquetas	260.51 ± 88.39	227.50 ± 94.11	< 0.05
Sodio	137.79 ± 3.79	137.62 ± 5.12	< 0.05
Saturación O2 art.	88.89 ± 11.52	87.97 ± 12.05	< 0.05
Saturacion O2 ven.	63.35 ± 21.64	65.63 ± 22.05	< 0.05
TTPA	33.09 ± 5.86	33.45 ± 6.99	< 0.05
Urea	40.26 ± 19.98	58.23 ± 35.24	< 0.05
pcO2 art.	40.28 ± 9.34	43.61 ± 15.07	< 0.05
pcO2 ven.	44.24 ± 8.46	45.24 ± 12.97	< 0.05
pH art.	7.42 ± 0.05	7.40 ± 0.08	< 0.05
pH ven.	7.39 ± 0.06	7.37 ± 0.08	< 0.05
pO2 art.	63.44 ± 19.09	62.99 ± 21.96	0.4291
pO2 ven.	37.34 ± 17.06	39.69 ± 18.93	< 0.05

Tabla 3. Análisis comparativo entre los pacientes que desarrollarán la variable de desenlace y los que no.

** Las variables categóricas se expresan como número total y porcentaje sobre cada columna. La comparación se lleva a cabo a través de test Chi cuadrado. Las variables cuantitativas se expresan como media y desviación estándar (sd), y la comparación se lleva a cabo por test t de Student.*

4.2 Resultados por modelo.

4.2.1 Random Forest.

180 configuraciones distintas variando técnicas e hiperparámetros fueron entrenadas (ver Anexo I). Los modelos sin ajuste de pesos dentro de la configuración fueron rápidamente desechados por tener una detección preferente de la clase negativa (muy baja S, con alta E). Sobre estos modelos, técnicas como SMOTE, o sobre todo ovesampling 1:1 o undersampling 1:1 conseguían aumentar S y F-1 score.

Sin embargo, cuando el ajuste de pesos estaba indicado dentro de la propia configuración del random forest ("*class weight = balanced")*, para un mismo umbral, los modelos de por sí tenían mejor S y F-1 score, a costa de un descenso de E, VPP, AUROC y AUPRC. En este caso, las técnicas adicionales conseguían modificaciones más marginales de las métricas. Incluso algunas técnicas como SMOTE resultaban contraproducentes, disminuyendo la capacidad de detección.

Según las reglas establecidas, se seleccionó como método más óptimo el 'RF con imputación por mediana y undersampling' con 200 árboles, profundidad máxima de 10, y fijando el umbral en 0.5. En este caso, la S fue 0.80, la E 0.73 y el VPP 0.25. El AUROC fue 0.84, la AUPRC 0.35 y el F-1 0.38. En la tabla 4 figura una comparación entre el modelo elegido, y los modelos ejecutados sobre los mismos parámetros con diferentes técnicas, para visualizar su efecto sobre las métricas.

La imputación por *Knn*, respecto a la mediana, además de ser más costosa computacionalmente, no supuso un beneficio apreciable.

4.2.2 XGBoost.

En XGBoost, las métricas más óptimas se obtuvieron con menos árboles y menos profundidad. Por lo general, las cifras de AUROC y AUPRC fueron más elevadas que en RF. 120 modelos fueron entrenados con diferentes configuraciones (Anexo I). De nuevo, no hubo grandes diferencias entre los tipos de imputación -incluso Knn las empeoró respecto a la mediana. También en este caso, cuando el modelo estaba configurado para tener en cuenta el desbalanceo en su entrenamiento, las técnicas adicionales de desbalanceo aportaron un beneficio más marginal (ver Tabla 4).

El modelo seleccionado más óptimo fue el 'XGBoost con imputación a mediana y oversampling x2), que tenía 100 estimadores, profundidad de 6, learning rate fijada en 0.1 y el umbral se determinó en 0.4. Con ello, el modelo conseguía una S 0.80, E 0.75 y VPP 0.26; con AUROC 0.85 AUPRC 0.40 y F-1score 0.40. Este modelo consigue la mayor AUROC y AUPRC entre todos los modelos, con un compromiso *precision-recall* considerado aceptable y cumpliendo la capacidad de detección de una S mínima de 0.80.

TABLA 4
MÉTRICAS DE RENDIMIENTO SEGÚN MODELO.

Random Forest. Nº estimadores: 200 // Profundidad: 10 // Umbral: 0.5
Entrenamiento del modelo ajustado por pesos.

Imputación	Desbalanceo	AUROC	AUPRC	ACC	PREC	RECALL	SPEC	F1-S
Mediana	Sin ajuste adicional	0.83	0.32	0.82	0.29	0.61	0.84	0.4
	Oversampling x2	0.83	0.33	0.8	0.29	0.68	0.81	0.4
	Oversampling 1:1	0.83	0.33	0.79	0.28	0.7	0.81	0.4
	SMOTE	0.82	0.32	0.89	0.41	0.28	0.96	0.33
	Undersampling 1:1	0.84	0.35	0.74	0.25	0.8	0.73	0.38
KNN	Sin ajuste adicional	0.81	0.3	0.83	0.3	0.55	0.86	0.39
	Oversampling x2	0.82	0.31	0.8	0.28	0.63	0.82	0.39
	Oversampling 1:1	0.82	0.31	0.8	0.28	0.65	0.81	0.39
	SMOTE	0.81	0.3	0.88	0.37	0.28	0.95	0.32
	Undersampling 1:1	0.82	0.32	0.73	0.24	0.79	0.72	0,37

XGBoost. Nº estimadores: 100 // Profundidad: 6 // LR: 0.1 // Umbral: 0.4
Entrenamiento del modelo ajustado por pesos.

Imputación	Desbalanceo	AUROC	AUPRC	ACC	PREC	RECALL	SPEC	F1-S
Mediana	Sin ajuste adiconal	0.85	0.39	0.75	0.26	0.78	0.75	0.39
	Oversampling x2	0.85	0.40	0.76	0.26	0.80	0.75	0.40
	Oversampling 1:1	0.85	0.38	0.76	0.26	0.79	0.75	0.39
	SMOTE	0.84	0.37	0.89	0.45	0.31	0.96	0.37
	Undersampling 1:1	0.85	0.39	0.70	0.23	0.85	0.68	0.36
KNN	Sin ajuste adicional	0.84	0.38	0.76	0.26	0.79	0.75	0.39
	Oversampling x2	0.84	0.38	0.75	0.25	0.77	0.74	0.38
	Oversampling 1:1	0.84	0.38	0.75	0.25	0.78	0.74	0.38
	SMOTE	0.82	0.36	0.89	0.44	0.32	0.96	0.37
	Undersampling 1:1	0.84	0.37	0.69	0.22	0.85	0.67	0.35

Red neuronal // Batch_size_ 32 // Epochs: 20 // Umbral: 0.5.
Entrenamiento del modelo con Oversampling 1:1.

Capas ocultas	Neuronas	AUROC	AUPRC	ACC	PREC	RECALL	SPEC	F1-S
3	10	0.83	0.35	0.86	0.36	0.50	0.90	0.41
	20	0.83	0.35	0.79	0.28	0.68	0.80	0.39
	50	0.83	0.36	0.80	0.29	0.67	0.82	0.40
	100	0.83	0.35	0.84	0.32	0.55	0.87	0.41
5	10	0.83	0.36	0.77	0.26	0.71	0.78	0.39
	20	0.83	0.36	0.84	0.32	0.58	0.86	0.41
	50	0.82	0.35	0.80	0.28	0.64	0.82	0.39
	100	0.82	0.34	0.80	0.28	0.65	0.81	0.39

Tabla 4. Métricas de rendimiento de RF, XGB, RNN para hiperparámetros similares. Abreviaturas: AUROC: area under receiving operator curve; AUPRC: area under precision-recall curve; ACC: accuracy; PREC: precision; SPEC: specificity; F1-S: f1 score.

4.2.3 Red neuronal recurrente.

Los modelos de red neuronal que no tenían un ajuste del desbalanceo previo, tenían S inferiores a 0.2 – 0.3. Por lo tanto, se entrenaron redes con los diseños previstos, sobre un conjunto sobre el que se había aplicado oversampling 1:1. Las AUROC obtenidas eran de 0.82 – 0.83 con una AUPRC 0.34-0.36, mejor que RF pero peor que XGBoost. Sin embargo, para cualquier caso, la S se veía penalizada respecto a RF y XGBoost, tal y como se aprecia en la Tabla 4. Incluso con modificaciones de los umbrales fue difícil obtener S por encima del objetivo de 0.80, sin penalizar demasiado el resto de métricas.

Respecto a la complejidad de la red, se observó como a partir de diseños más complejos que los mostrados, el AUROC y AUPRC comenzaban a descender, revelando un patrón de sobreentrenamiento. Por lo tanto, no se consideró que la optimización de los parámetros pudiera encontrar un modelo superior a los anteriores.

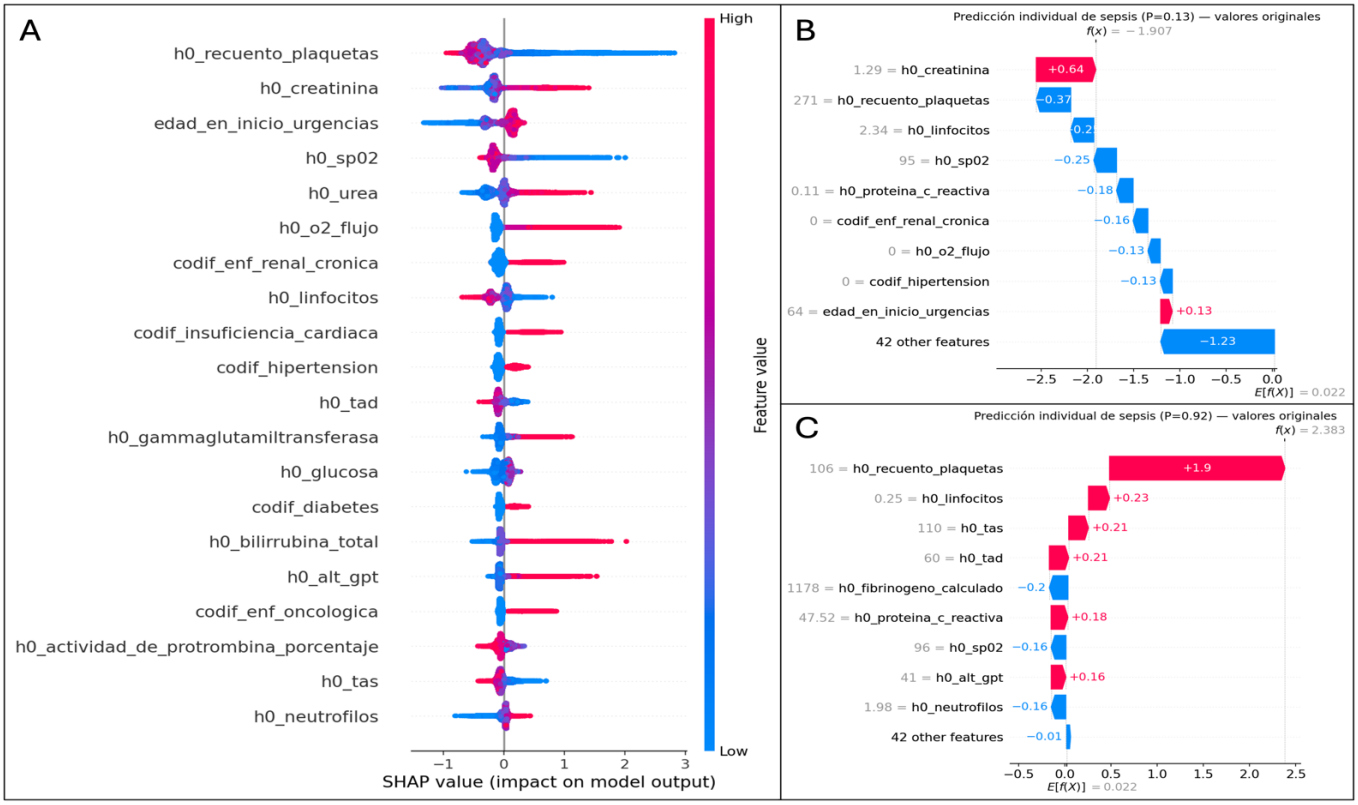

Figura 3. Gráficos con valores SHAP para (A) predicción global del modelo. En rojo se visualiza como el aumento de una variable influye en la predicción, aumentado o disminuyendo el riesgo del valor SHAP. (B) Predicción individual de un caso etiquetado como poco probable de presenter el desenlace (P: 0.13). Se puede observer como el valor de creatinina aumenta el riesgo; mientras que los valores en azul están contribuyendo a disminuirlo (C) Predicción individual sobre un caso de alta probabilidad (P: 0.92).

4.3 Módulos de explicabilidad.

Al aplicar los métodos de explicabilidad, encontramos que las variables con más importancia, por índice GINI, en el modelo de RF fueron: recuento de plaquetas, edad, urea, antecedente de ERC y saturación periférica de O2. Al aplicar *Permutation Importance* sobre el mismo modelo, igualmente el factor de mayor peso fue el recuento de plaquetas, seguido de antecedentes de ERC, insuficiencia cardiaca, y el flujo de O2.

Sobre el mejor modelo considerad (XGBoost) las variables más importante por GINI fueron también antecedente de ERC e insuficiencia cardiaca, flujo de O2, antecedente oncológico, edad, y el recuento de plaquetas. En *Permutation Importance*, el recuento de plaquetas volvió a ser el factor con más peso. Sobre este modelo se aplicaron los valores SHAP, que confirmaron la cifra de plaquetas como el parámetro más decisivo en nuestra predicción. En la Figura 3a se puede observar la influencia de las variables sobre todas las predicciones del modelo de forma conjunta, y observar su tendencia (por ejemplo, como una cifra de plaquetas baja, en azul, influye en aumentar la probabilidad de desenlace , o una cifra elevada de creatinina, en rojo, actúa en el mismo sentido). Las figuras 3b y 3c, son ejemplos de una predicción individual aplicada a casos de ejemplo clasificados de alto y bajo riesgo, con el peso correspondiente de cada variable. Estos gráficos constituyen el sistema de predicción explicable e individualizado que permite trasladar el funcionamiento del modelo a la interpretabilidad clínica.

5 Discusión.

En este trabajo, hemos desarrollado un modelo de ML para predecir el pronóstico desfavorable en los pacientes con sospecha de infección -sepsis, UCI, o fallecimiento-, en las siguientes 48 horas al inicio de la atención en urgencias. El mejor modelo considerado, obtiene una AUROC de 0.85, AUPRC de 0.40, con S del 0.80, con un VPP del 0.26, E 0.85. Esto implica que para obtener una sensibilidad considerada adecuada, la tasa de falsos positivos aún es elevada, alineado con lo referido en la literatura previa. Consideramos que el modelo se muestra competitivo por varias razones.

En primer lugar, se ejecuta sobre variables adquiridas en el entorno de urgencias, en el punto de entrada de los pacientes con sospecha de infección al sistema sanitario, y en sus primeras horas de atención. Esto supera los problemas de generalización que sopone trasladar modelos entrenados en circuitos como UCI a urgencias, desde un entorno de alta monitorización y pocos valores perdidos. La decisión de incluir el ingreso en UCI o la mortalidad en la variable de predicción, impuesta por la carencia de algunos parámetros SOFA, garantiza una solución robusta para que el modelo actúe como un sistema de alerta de riesgo de mal pronóstico.

Además, consideramos que en el diseño de muchos estudios previos, en los que se incluyen los pacientes con SOFA >= 2 al inicio, los modelos trabajan sobre un objetivo

de detección, similar al de las escalas como el propio SOFA. Además, introducir en el entrenamiento del modelo como variables predictoras elementos del SOFA, a la vez que el desenlace se calcula por esta escala, propicia un riesgo de "data leakeage" que debe ser evitado. Al tomar para entrenamiento las variables a tiempo 0 h sobre pacientes sin SOFA elevado, para predecir su aumento en las horas posteriores, se disminuye este riesgo.

Precisamente por el tiempo de predicción que se define; consideramos las métricas del modelo comparables a los reportados previamente. Un AUROC de 0.85 está en línea con lo descrito, pero hemos visto como a mayor tiempo de predicción respecto al tiempo de aparición de la sepsis, el rendimiento bajo. Un meta-análisis de estudios con ML, indica que más allá de 12h, el AUROC medio está por debajo de 0.60 (8). Aunque nuestro modelo no trabaja con modelos temporales estrictos y no se define tal "sepsis-onset", sino aparición del evento a lo largo de un periodo, consideramos adecuado su rendimiento para tal intervalo.

Hemos señalado la importancia de otras métricas en el buen rendimiento del modelo, más allá de la AUROC. El compromiso *recall-precision* (S-VPP), en un problema de detección de una clase poco frecuente como es la sepsis (en nuestro análisis, con una prevalencia de 9.9%), resulta más informatico que la AUROC. Por ello, en nuestra decisión sobre el modelo, se estableció la AUPRC como un criterio esencial, para una S mínima aceptable. En nuestro mejor modelo, para conseguir una S del 0.80, el VPP máximo es 0.26. Eso implica 3 falsos positivos por cada verdadero positivo. Sin embargo, esto es lo frecuente en la literatura descrita, reconocido como un problema universal para sepsis y ML. De hecho, una AUPRC de 0.40, para el intervalo de tiempo considerado, es significativamente mejor que algunos de los ejemplos revisados.

Una decisión que nos parece relevante es la de incluir las comorbilidades como parte del entrenamiento del modelo. Algunos modelos operan sólo con constantes vitales y datos de laboratorio. Sin embargo, es conocido que muchas enfermedades previas son factores de riesgo reconocidos para el desarrollo de sepsis, y deben ser consideradas (23).

Respecto a la elección de los métodos de ML utilizados, RF y XGBoost se habían demostrado por la literatura como soluciones válidas. En nuestro caso, fueron claramente superiores al rendimiento de una RNN. Algo que se mostró esencial, fue el ajuste de balanceo o pesos dentro de la configuración del modelo. Una vez realizado este ajuste, los métodos de imputación o de tratamiento del desbalanceo, modificaron las métricas de forma marginal, aunque lo suficiente para que los mejores modelos de RF y XGBoost considerados, fueran los sometidos a undersampling y oversampling x2, respectivamente. El coste computacional de aplicar métodos como Knn o SMOTE, que ralentizan el modelo, no se muestra útil en nuestro. Es llamativo el mal comportamiento de los modelos entrenados con SMOTE, en contradicción con lo descrito en otros casos (20).

Los módulos de explicabilidad son la clave para que los resultados de un modelo de ML, pueda ser aplicado a la práctica, y el clínico confíe en las predicciones del modelo. El análisis global de la importancia de las variables en ML, es otra forma de generar conocimiento científico. En nuestro caso, coincidiendo con lo descrito previamente, variables como la edad, la función renal, la saturación de oxígeno, tuvieron un peso relevante en la mayoría de análisis de importancia de las variables. Es llamativa la dominancia del recuento de plaquetas como factor relevante en todos los análisis de nuestros modelos. Siendo un marcador clásico de desarrollo de sepsis (24), y que incluso forma parte del SOFA, en ninguno de los ejemplos con análisis SHAP revisados, fue tan determinante como en nuestro caso (15,18).

Por último, en este trabajo proponemos la utilidad de implementar una herramienta de análisis individual de la predicción. Esta información, permite actuar sobre aspectos concretos reversibles que están materializando el riesgo (como identificar si el riesgo deriva de parámetros respiratorios, renales, hematológicos, etc.), y es un ejemplo de cómo las técnicas de ML pueden suponer una mejora en la atención médica.

Las limitaciones de nuestro estudio vienen dadas por el diseño retrospectivo y la ausencia de algunos datos para las que hubo que recurrir a técnicas de derivación, sustitución, o imputación. El dataset final contenía variables con alto porcentaje de valores perdidos, aunque esto es reflejo de la atención hospitalaria real del manejo en urgencias. Además, se trabajó con una predicción a tiempo inicial y desarrollo de eventos a lo largo de un intervalo de tiempo, y no sobre predicciones actualizadas a tiempo real o tendencias temporales, que podrían ser más útiles en una estimación dinámica del riesgo. Vistos los problemas de generalización de modelos previos, sería necesario evaluar las métricas de rendimiento de nuestro modelo sobre otra población, para asegurar su utilidad.

6 Conclusión

Las técnicas de ML prometen ser útiles para la predicción de sepsis, pero se enfrentan aún a ciertos problemas para su aplicación generalizada. Nuestro modelo se enfoca hacia la evaluación del riesgo desfavorable desde la atención inicial en urgencias, con una capacidad predictiva competitiva que tiene en cuenta los problemas reconocidos para una tarea de estas características. Los módulos de explicabilidad, especialmente los que analizan una predicción individualizada, serán una herramienta imprescindible para una implantación clínica confiable.

7 Secciones finales.

7.1 Apéndices.

Se incluye aparte el Anexo I, en el que figuran:.

1. Rangos definidos para las variables.
2. Variables seleccionadas para el entrenamiento.

3.1 Rendimiento de modelos RF balanceados, filtrados por Sensibilidad > 0.80.

3.2 Rendimiento de modelos XGB balanceados, filtrados por Sensibilidad > 0.80.

4.1 Permutation importance de modelo RF elegido.

4.2 Permutation importance de modelo XGB elegido.

8 AGRADECIMIENTOS.

Al autor del presente documento le gustaría expresar su agradecimiento a los miembros del HUF que han participado como investigadores colaboradores; en especial a los miembros del departamento de Sistemas de Información, por su ayuda en la conceptualización y extracción de datos, haciendo viable el desarrollo del proyecto.

REFERENCIAS

1. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 23 de febrero de 2016;315(8):801-10.

2. Sepsis. World Health Organisation [Internet]. [citado 7 de julio de 2025]. Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/sepsis

3. Álvaro-Meca A, Jiménez-Sousa MA, Micheloud D, Sánchez-Lopez A, Heredia-Rodríguez M, Tamayo E, et al. Epidemiological trends of sepsis in the twenty-first century (2000-2013): an analysis of incidence, mortality, and associated costs in Spain. Popul Health Metr. 12 de febrero de 2018;16(1):4.

4. Liu VX, Fielding-Singh V, Greene JD, Baker JM, Iwashyna TJ, Bhattacharya J, et al. The Timing of Early Antibiotics and Hospital Mortality in Sepsis. Am J Respir Crit Care Med. 1 de octubre de 2017;196(7):856-63.

5. De Grooth HJ, Postema J, Loer SA, Parienti JJ, Oudemans-van Straaten HM, Girbes AR. Unexplained mortality differences between septic shock trials: a systematic analysis of population characteristics and control-group mortality rates. Intensive Care Med. marzo de 2018;44(3):311-22.

6. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med. julio de 1996;22(7):707-10.

7. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA. 3 de abril de 2018;319(13):1317.

8. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med. 2020;46(3):383-400.

9. Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, et al. A computational approach to early sepsis detection. Comput Biol Med. julio de 2016;74:69-73.

10. Scherpf M, Gräßer F, Malberg H, Zaunseder S. Predicting sepsis with a recurrent neural network using the MIMIC III database. Comput Biol Med. octubre de 2019;113:103395.

11. Persson I, Östling A, Arlbrandt M, Söderberg J, Becedas D. A Machine Learning Sepsis Prediction Algorithm for Intended Intensive Care Unit Use (NAVOY Sepsis): Proof-of-Concept Study. JMIR Form Res. 30 de septiembre de 2021;5(9):e28000.

12. Rosnati M, Fortuin V. MGP-AttTCN: An interpretable machine learning model for the prediction of sepsis. Olier I, editor. PLOS ONE. 7 de mayo de 2021;16(5):e0251248.

13. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. Crit Care Med. abril de 2018;46(4):547-53.

14. Zhou L, Shao M, Wang C, Wang Y. An early sepsis prediction model utilizing machine learning and unbalanced data processing in a clinical context. Prev Med Rep. septiembre de 2024;45:102841.

15. Bhargava A, López-Espina C, Schmalz L, Khan S, Watson GL, Urdiales D, et al. FDA-Authorized AI/ML Tool for Sepsis Prediction: Development and Validation. NEJM AI [Internet]. 27 de noviembre de 2024 [citado 7 de julio de 2025];1(12). Disponible en: https://ai.nejm.org/doi/10.1056/AIoa2400867

16. Liu X, Li M, Liu X, Luo Y, Yang D, Ouyang H, et al. Clinical validation and optimization of machine learning models for early prediction of sepsis. Front Med [Internet]. 5 de febrero de 2025 [citado 7 de julio de 2025];12. Disponible en: https://www.frontiersin.org/articles/10.3389/fmed.2025.1521660/full

17. Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open. enero de 2018;8(1):e017833.

18. Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun [Internet]. 31 de julio de 2020 [citado 7 de julio de 2025];11(1). Disponible en: https://www.nature.com/articles/s41467-020-17431-x

19. Wang Z, Wang W, Sun C, Li J, Xie S, Xu J, et al. A methodological systematic review of validation and performance of sepsis real-time prediction models. Npj Digit Med. 7 de abril de 2025;8(1):190.

20. Goh KH, Wang L, Yeow AYK, Poh H, Li K, Yeow JJL, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun [Internet]. 29 de enero de 2021 [citado 7 de julio de 2025];12(1). Disponible en: https://www.nature.com/articles/s41467-021-20910-4

21. Shashikumar SP, Wardi G, Malhotra A, Nemati S. Artificial intelligence sepsis prediction algorithm learns to say "I don't know". Npj Digit Med [Internet]. 9 de septiembre de 2021 [citado 7 de julio de 2025];4(1). Disponible en: https://www.nature.com/articles/s41746-021-00504-6

22. Burdick H, Pino E, Gabel-Comeau D, McCoy A, Gu C, Roberts J, et al. Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: a prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals. BMJ Health Care Inform. abril de 2020;27(1):e100109.

23. Shibata J, Osawa I, Ito H, Soeno S, Hara K, Sonoo T, et al. Risk factors of sepsis among patients with qSOFA<2 in the emergency department. Am J Emerg Med. diciembre de 2021;50:699-706.

24. Cox D. Sepsis - it is all about the platelets. Front Immunol. 2023;14:1210219.