UNIVERSIDAD DE VALLADOLID

FACULTAD DE MEDICINA

ESCUELA DE INGENIERÍAS INDUSTRIALES

TRABAJO DE FIN DE GRADO

GRADO EN INGENIERÍA BIOMÉDICA

Aplicación de algoritmos de machine learning para predecir el riesgo de deterioro en pacientes sépticos

Autora:

D.ª María Gutiérrez Burgos

Tutoras:

Dra. Dª María García Gadañón

Dra. Dª Elena Bustamante Munguira

Valladolid, septiembre 2025.

TÍTULO:	Aplicación de Algoritmos de Machine Learning para Predecir el Riesgo de Deterioro en Pacientes Sépticos
AUTORA:	D.ª María Gutiérrez Burgos
TUTORAS:	Dra. Dª María García Gadañón Dra. Dª Elena Bustamante Munguira
DEPARTAMENTO:	Departamento de Teoría de la Señal y Comunicaciones e Ingeniería Telemática y Departamento de Medicina, Dermatología y Toxicología

TRIBUNAL

PRESIDENTE:	Dra. Dª Elena Bustamante Munguira
SECRETARIO:	Dr. D Javier Gómez Pilar
VOCAL:	Dra. Dª María García Gadañón
SUPLENTE 1:	Dr. D Jesús Poza Crespo
SUPLENTE 2:	Dr. D Carlos Gómez Peña

FECHA:	Septiembre 2025
CALIFICACIÓN:	

A mis padres,
por haber creído en mí siempre.

Agradecimientos

Quisiera expresar mi más sincero agradecimiento a todas las personas que han hecho posible la realización de este Trabajo de Fin de Grado. En primer lugar, agradezco a mis tutoras María García Gadañón y Elena Bustamante Munguira por permitirme realizar este TFG, su orientación, paciencia y ayuda han sido fundamentales para guiarme durante este proceso. Además, quiero agradecer a Javier Gómez Pilar por haberme resuelto muchas dudas y ayudado a resolver varias situaciones a pesar de no tutorizar este trabajo. Asimismo, quiero agradecer a los profesores del Grado en Ingeniería Biomédica por los conocimientos compartidos a lo largo de la carrera, que han sido fundamentales para abordar este proyecto con rigor académico, en especial a Jesús Poza Crespo, quien siempre ha mostrado un remarcable compromiso y fe en nosotros.

Agradezco especialmente a mis amigos, que me han acompañado desde el inicio y sin ellos, esta etapa hubiera sido mucho más dificil, especialmente a mi mejor amiga, por brindarme apoyo y estar ahí siempre.

A mi familia, por estar a mi lado y animarme cuando hay días negros, en especial a mis padres, quienes son lo más importante de mi vida y sin ellos no hubiera sido posible llegar hasta aquí. A mis abuelos por darme todo el cariño que puedo desear y a todas las personas que quiero y que espero que sigan estando a mi lado.

Muchas gracias a todos.

Resumen

La sepsis y el shock séptico constituyen una de las principales causas de mortalidad en las unidades de cuidados intensivos (UCI), con tasas que oscilan entre el 25-30% a nivel mundial. La identificación temprana de pacientes de alto riesgo representa un desafío clínico crítico debido a la necesidad de una rápida actuación ante la afección y la rápida progresión de esta patología. En este contexto, los modelos predictivos basados en machine learning han emergido como herramientas prometedoras para mejorar la estratificación de riesgo y optimizar la toma de decisiones clínicas. En este trabajo, entendemos "riesgo" como la probabilidad de mortalidad temprana (60 días tras el ingreso en UCI), categorizada en dos grupos: riesgo alto (pacientes fallecidos en ese intervalo) y riesgo bajo (pacientes supervivientes). El objetivo principal de este trabajo fue desarrollar y evaluar modelos predictivos de riesgo para pacientes con sepsis ingresados en UCI, comparando diferentes algoritmos de clasificación y técnicas de balanceo de clases, así como analizar la interpretabilidad de los modelos mediante técnicas de inteligencia artificial explicable (XAI).

La metodología empleada se basó en un registro clínico anonimizado de 226 pacientes ingresados en la UCI del Hospital Clínico Universitario de Valladolid durante 2024 y el primer trimestre de 2025, todos con sospecha de sepsis o shock séptico. Tras un exhaustivo proceso de preprocesado que incluyó la limpieza de datos, imputación de valores faltantes mediante *k*NN, y la transformación de variables categóricas mediante *one-hot encoding*, se aplicó la técnica *Fast Correlation Based Filter* (FCBF) para la selección de las 15 variables más relevantes, incluyendo scores clínicos (APACHE II, SOFA), variables fisiológicas (procalcitonina, proteína C reactiva, albúmina, entre otros), factores de riesgo y características del foco infeccioso. Se evaluaron cinco algoritmos de clasificación: Regresión Logística, *Random Forest, XGBoost, LightGBM y Balanced Random Forest*, combinados con cuatro técnicas de balanceo de clases (SMOTE, ADASYN, SMOTEENN y *UnderSampler*) para abordar el desbalance inherente entre pacientes de alto y bajo riesgo. La evaluación se realizó mediante validación cruzada estratificada de 5 particiones, priorizando el *recall* como métrica principal dada la importancia clínica de minimizar falsos negativos. Adicionalmente, se aplicaron técnicas de interpretabilidad SHAP (*SHapley Additive exPlanations*) a los modelos con mejor rendimiento (*XGBoost* y *LightGBM* con *UnderSampler*) para identificar las variables más influyentes en las predicciones.

Los resultados demostraron que la combinación *XGBoost* con *UnderSampler* alcanzó el mejor rendimiento en términos de *recall* (0.69), seguida por *LightGBM con UnderSampler* (0.68) y Regresión Logística con ADASYN (0.65). Los valores de ROC-AUC oscilaron entre 0.68 y 0.76, indicando una capacidad discriminativa de moderada a buena. Es reseñable mencionar que las técnicas de submuestreo mostraron un rendimiento superior respecto a las de sobremuestreo para los algoritmos de *boosting*. Las matrices de confusión mostraron que los modelos mantuvieron un balance adecuado entre la detección de casos de alto riesgo y la especificidad para casos de bajo riesgo. El análisis de interpretabilidad mediante SHAP reveló que las variables más influyentes en las predicciones fueron la dosis de aminas vasopresoras, el score APACHE II, el foco respiratorio, el score SOFA y los niveles de albúmina, resultados coherentes con el conocimiento médico establecido sobre factores pronósticos en sepsis.

En conclusión, este trabajo demuestra la viabilidad y efectividad de los modelos de *machine learning* para la predicción de riesgo en pacientes con sepsis en UCI, con *XGBoost* emergiendo como el algoritmo más prometedor cuando se combina con técnicas de submuestreo. Los modelos desarrollados igualaron el rendimiento de los sistemas de puntuación clínicos tradicionales y mostraron patrones de predicción clínicamente interpretables. Estos hallazgos sugieren que la implementación de herramientas predictivas basadas en inteligencia artificial podría mejorar significativamente la identificación temprana de pacientes de alto riesgo, facilitando intervenciones terapéuticas más oportunas y personalizadas.

Palabras clave

Balanceo de clases, *explainable artificial intelligence*, *machine learning*, predicción de riesgo, sepsis, shock séptico, unidad de vigilancia intensiva.

Abstract

Sepsis and septic shock are among the leading causes of mortality in intensive care units (ICUs), with rates ranging between 25–30% worldwide. Early identification of high-risk patients represents a critical clinical challenge due to the urgent need for prompt intervention and the rapid progression of this condition. In this context, predictive models based on machine learning have emerged as promising tools to improve risk stratification and optimize clinical decision-making. In this work, we define "risk" as the probability of early mortality (within 60 days after ICU admission), categorized into two groups: high risk (patients who died within this interval) and low risk (patients who survived). The main objective of this study was to develop and evaluate predictive risk models for ICU patients with sepsis, comparing different classification algorithms and class balancing techniques, as well as analyzing model interpretability using explainable artificial intelligence (XAI) techniques.

The methodology was based on an anonymized clinical registry of 226 patients admitted to the ICU of the Hospital Clínico Universitario de Valladolid during 2024 and the first quarter of 2025, all with suspected sepsis or septic shock. After an extensive preprocessing phase that included data cleaning, missing value imputation using *k*NN, and categorical variable transformation through one-hot encoding, the Fast Correlation Based Filter (FCBF) technique was applied to select the 15 most relevant variables, including clinical scores (APACHE II, SOFA), physiological variables (procalcitonin, C-reactive protein, albumin, among others), risk factors, and characteristics of the infectious focus. Five classification algorithms were evaluated: Logistic Regression, Random Forest, XGBoost, LightGBM, and Balanced Random Forest, combined with four class balancing techniques (SMOTE, ADASYN, SMOTEENN, and UnderSampler) to address the inherent imbalance between high- and low-risk patients. Model evaluation was performed using 5-fold stratified cross-validation, prioritizing recall as the main metric due to the clinical importance of minimizing false negatives. Additionally, SHAP (SHapley Additive exPlanations) interpretability techniques were applied to the best-performing models (XGBoost and LightGBM with UnderSampler) to identify the most influential variables in the predictions.

The results showed that the combination of XGBoost with UnderSampler achieved the best performance in terms of recall (0.69), followed by LightGBM with UnderSampler (0.68) and Logistic Regression with ADASYN (0.65).

ROC-AUC values ranged between 0.68 and 0.76, indicating moderate to good discriminative ability. Contrary to initial expectations, undersampling techniques outperformed oversampling methods for boosting algorithms. The SHAP-based interpretability analysis revealed that the most influential variables in the predictions were vasopressor dose, APACHE II score, respiratory focus, SOFA score, and albumin levels, findings consistent with established medical knowledge on prognostic factors in sepsis. Confusion matrices showed that the models maintained an adequate balance between detecting high-risk cases and preserving specificity for low-risk cases.

In conclusion, this work demonstrates the feasibility and effectiveness of machine learning models for risk prediction in ICU patients with sepsis, with XGBoost emerging as the most promising algorithm when combined with undersampling techniques. The developed models matched the performance of traditional clinical scoring systems and exhibited clinically interpretable prediction patterns. These findings suggest that the implementation of AI-based predictive tools could significantly improve the early identification of high-risk patients, enabling more timely and personalized therapeutic interventions.

Keywords

Class balancing, explainable artificial intelligence, intensive care unit, Machine learning, risk prediction, sepsis, septic shock.

Índice general

Índice de figuras

Índice de tablas

Glosario de abreviaturas y acrónimos

ACCP: American College of Chest Physicians (Sociedad Americana de Médicos del Tórax)

ADASYN: Adaptative Synthetic Sampling (Muestreo sintético adaptativo)

APACHE II: Acute Physiology and Chronic Health Evaluation (Evaluación de la Fisiología Aguda y de la Salud Crónica)

AUC ROC: Area bajo la Curva Receiver Operating Characteristics

BRC: B Cell Receptor (célula receptora tipo B)

CPM: Ciclos por minuto

DM2: Diabetes Mellitus tipo 2

EHR: Historia Clínica Electrónica (Electronic Health Record)

EPOC: Enfermedad Pulmonar Obstructiva Crónica

HTA: Hipertensión arterial

FC: Frecuencia cardíaca

FiO2: Fracción Inspirada de Oxígeno

FR: Frecuencia respiratoria

IA: Inteligencia Artificial

LAK: Lymphokine Activated Killer (asesinas activadas por linfoquinas)

LPM: latidos por minuto

MIMIC: Medical Information Mart for Intensive Care (Base de Datos de Información Médica para Cuidados Intensivos)

ML: Machine Learning (aprendizaje automático)

NK: Natural Killer (células asesinas)

PAM: Presión Arterial Media

PAMP: Patrones Moleculares Asociados a Patógenos

PaO2: Presión Parcial de Oxígeno

PAS: Presión Arterial Sistólica

PCR: Proteína C Reactiva

PTC: Procalcitonina

RF: Random Forest (bosque aleatorio)

SHAP: SHapley Additive exPlanations (Explicaciones Aditivas de Shapley)

SMOTE: Synthetic Minority Over-sampling Technique (Técnica de Sobremuestreo Sintético de la Minoría)

SOFA: Sequential Organ Failure Assestment (Evaluación Secuencial de la Insuficiencia Orgánica)

SRIS: Síndrome de Respuesta Inflamatoria Sistémica

TCR: T Cell Receptor (célula receptora tipo T)

VIH: Virus de la Inmunodeficiencia Humana

Capítulo 1. Introducción y contenidos teóricos

1.1 Introducción y motivación

El presente Trabajo de Fin de Grado (TFG) titulado "Aplicación de Algoritmos de *Machine Learning* para Predecir el Riesgo de Deterioro en Pacientes Sépticos", propone el análisis de un problema clínico de gran relevancia en la medicina, como es la sepsis y el shock séptico, situaciones que comprometen seriamente la supervivencia del paciente en la Unidad de Cuidados Intensivos (UCI). Este tema no solo tiene una notable pertinencia en el ámbito médico, sino también en el área de la Ingeniería Biomédica, ya que se persigue la aplicación de técnicas computacionales y redes neuronales que permitan caracterizar el estado físico del paciente con el fin de vaticinar su empeoramiento.

La sepsis es una disfunción orgánica potencialmente mortal causada por una respuesta desregulada del organismo frente a una infección, mientras que el shock séptico constituye su forma más grave, caracterizada por una alteración circulatoria y metabólica que conduce a una elevada mortalidad. Ambas condiciones representan uno de los principales retos en las UCI, con tasas de fallecimiento que oscilan entre el 25–30% a nivel mundial, a pesar de los avances en el tratamiento y monitorización de los pacientes críticos (Fleuren et al., 2020; Rhee et al., 2019).

La motivación principal de este trabajo es disponer de métodos que complementen el juicio clínico y ayuden a los profesionales sanitarios a predecir de forma temprana el deterioro de los pacientes. En este contexto, los sistemas basados en técnicas de inteligencia artificial y *machine learning* son una alternativa prometedora, debido a la posibilidad de analizar grandes volúmenes de datos clínicos e identificar patrones que tal vez se pasen por alto a simple vista (Nikravangolsefid et al., 2024). La importancia de este trabajo reside en el impacto potencial que puede tener la mejora en la atención médica intensiva, facilitando una pronta adecuación del tratamiento al paciente antes de que se produzca una evolución negativa de su estado de salud. El desarrollo de modelos capaces de anticipar el deterioro físico de los pacientes sépticos podría contribuir no solo a salvar vidas, sino también a optimizar los recursos disponibles en las UCI y avanzar en el conocimiento de los mecanismos fisiopatológicos subyacentes a esta enfermedad.

1.2 El sistema inmunitario

El sistema inmunitario es una red de órganos linfoides, células, factores humorales y citoquinas que son capaces de diferenciar los cuerpos extraños u ofensivos de los inofensivos (Parkin, 2001). Su función principal es la defensa del huésped frente a una agresión u organismo invasor y mantener la homeostasis dentro del cuerpo. No obstante, en la sepsis y el shock séptico

esta respuesta se desregula, generando una inflamación excesiva que puede llevar a disfunción orgánica y alta mortalidad. La respuesta inmune es la respuesta global y coordinada de todas las células y moléculas ante los patógenos y células malignas, en conjunto, los antígenos. Los seres humanos poseemos dos subtipos de sistemas inmunitarios, el sistema inmune innato y el sistema inmune adquirido, cada uno con sus respuestas típicas y fisiología propia (Sanz, 2017):

Sistema inmune innato: Constituye una primera defensa frente a patógenos mayoritariamente inespecífica ya que el mismo mecanismo actúa frente a diferentes agentes eliminando también la memoria inmunológica. De este modo, siempre actúa de la misma manera y desencadena una respuesta similar y de forma rápida, incluso cuando la misma infección se repite. Su activación es fundamental para la intervención de su contraparte adaptativa, situación posterior que hará frente al agente invasor (Cutuli et al., 2008).
Se compone de barreras (físicas, químicas o biológicas), factores solubles (citoquinas y microbiota) y células:

- Células fagocíticas:
  - Neutrófilos: Son las células blancas sanguíneas más abundantes en humanos, y son las primeras en acudir a los puntos de inflamación aguda para responder a la agresión mediante la fagocitosis de patógenos y/o la liberación de factores antimicrobianos.
  - Eosinófilos: Son responsables de muchas funciones proinflamatorias en las que se requiere una hipersensibilidad inmediata (alergias), y son particularmente efectivos en matar bacterias como *E. Coli*, además de poseer una gran toxicidad para las células tumorales gracias a su proteína básica (Galeana et al., 2003).
  - Basófilos: Reaccionan liberando histamina y constituyen un elemento protector frente a agentes malignos pluricelulares (Sanz, 2017).
  - Macrófagos: En estado inactivo, patrullan los tejidos, manteniendo una altísima capacidad fagocítica. Son capaces de reconocer patrones moleculares asociados a patógenos (PAMP) y liberar grandes cantidades de citoquinas iniciadoras de la inflamación y el reclutamiento celular. Su función principal es destruir patógenos, fagocitándolos. Además, tiene un papel fundamental en la reparación de tejidos, remodelación vascular y estimulación de células madre (Sanz, 2017).
  - Células dendríticas: Conforman una pequeña proporción de células circundantes, sin embargo, su función principal es enlazar el sistema inmune innato con el adaptativo. Además, detectan

patógenos y activan los linfocitos T para dirigirlos hacia su diferenciación (Sanz, 2017).

- Células asesinas:
  - NK (*Natural Killer*): Este subtipo celular se compone de células innatas citotóxicas producidas en la médula ósea que producen citoquinas y quimiocinas inflamatorias (Wolf, 2023). Provocan la destrucción de gérmenes y células cancerosas mediante citólisis en lugar de fagocitosis (Caligiuri et al., 2008).
  - LAK (*Lymphokine Activated Killer*): Son células asesinas activadas por linfoquinas, que se activan de forma artificial en laboratorio. Su función es aumentar su capacidad citolítica y su especificidad de membrana (Grimm, 1986).

Sistema inmune adaptativo: Este sistema es el encargado de eliminar los tipos específicos de gérmenes causantes de la infección cuando el sistema innato falla. Es un mecanismo más lento que el anterior, puesto que primero debe reconocer el patógeno. Una de sus ventajas es que tiene memoria inmunológica, y una infección repetida será combatida más rápidamente en situaciones próximas, y puede tener menos efecto en el organismo (Wang et al., 2024). El sistema inmune adaptativo se compone de:

- Linfocitos T: Son responsables de la inmunidad celular, se originan a partir de precursores en la médula ósea y maduran en el timo (Díaz Martín, 2013). Sus principales funciones son erradicar las infecciones producidas por microbios intracelulares y activar otras células como macrófagos y linfocitos B. Presentan receptores TCR (*T-cell receptor*) en su membrana, y mediante ellos, son capaces de identificar el antígeno de forma específica, con la ayuda de las moléculas de histocompatibilidad.
- Linfocitos B: Son responsables de la inmunidad humoral, defienden al huésped de gérmenes por medio de la secreción de anticuerpos o inmunoglobulinas que reconocen a los antígenos (Martín et al, 2013). Tienen receptores de tipo BCR (*B-cell receptor*) y funciones inmunoreguladoras, secretan citoquinas e influencian la acción de otras células, neogénesis de tejidos linfoides y reparación de tejidos dañados por la inflamación.
- Anticuerpos: Son proteínas cuya estructura contiene un epítopo que reconoce al antígeno específico. Estos epítopos se unen a su antígeno y lo marca para que sea reconocido por otros elementos del sistema inmunitario. Impiden que los patógenos entren en las células, estimulan la eliminación de un germen por los macrófagos

y pueden desencadenar su destrucción activando la vía del complemento.

Sistema del complemento: este sistema es de vital relevancia en el contexto de las infecciones, puesto que es el componente fundamental de la respuesta inmunitaria defensiva frente a un agente hostil (Conigliaro et al., 2019). Está compuesto por más de 50 proteínas unidas a las membranas y es el centinela que detecta y elimina patógenos cuando estos atraviesan las barreras protectoras del individuo. El mecanismo de iniciación es la detección del patrón molecular asociado al patógeno (PAMP), lo que activa una de las tres vías de activación principales, la vía clásica, la vía de la lectina y la vía alternativa (West et al, 2018). Todas las vías culminan en la formación del complejo de ataque de membrana, que provocará la muerte lítica del germen, facilitando la defensa contra las infecciones. Un fallo en el sistema de complemento deja al huésped expuesto ante agresiones, causando una desregulación en la defensa y una respuesta inflamatoria alterada.

La inflamación es un proceso biológico constituido por una serie de fenómenos moleculares, celulares y vasculares con la finalidad de defender a un individuo frente a agresiones físicas, químicas o biológicas (González et al., 2010). La respuesta inflamatoria es inicialmente inespecífica, inmediata. En primer lugar, se focaliza la respuesta limitando la zona de lucha contra el agente invasor. El foco inflamatorio atrae a las células inmunes de los tejidos cercanos, facilitando su movilización por la alteración de la permeabilidad vascular, lo que ocasionará calor y rubor, e incluso un edema por el acúmulo de estas células.

Las fases de la inflamación son (Guarino et al., 2025):

- Liberación y efecto de mediadores: sintetizados por el mastocito. Producen alteraciones vasculares y efectos quimiotácticos que favorecen la llegada de las células defensivas al foco inflamatorio.
- Regulación del proceso inflamatorio: Mediado por inhibidores y activadores trabajando en conjunto.
- Reparación: Recuperación de tejidos dañados por el agente agresor.

En ocasiones, la inflamación aguda local falla y se produce una disfunción no modulada. Esto puede desencadenar una reacción generalizada en todo el cuerpo que, si no se controla, puede provocar el síndrome de respuesta inflamatoria generalizada (SRIS), mediada por una liberación masiva de citoquinas, macrófagos, plaquetas y factores de crecimiento, la cual conduce al fracaso funcional de los diferentes órganos y sistemas, poniendo en peligro la vida del individuo (García Barreno et al., 2008). Cuando el síndrome de respuesta inflamatoria generalizada es causado por una infección vírica,

bacteriana o fúngica, se habla de sepsis, la cual es el motivo de investigación de este trabajo. Esto se ve reflejado en la figura 1.

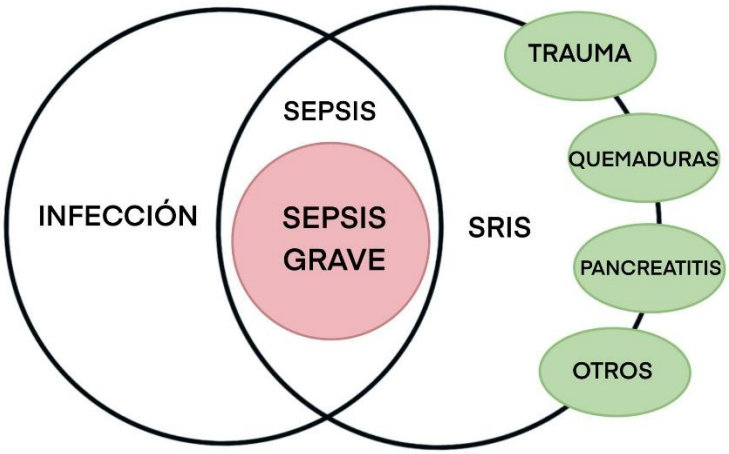


Figura 1. Correlación entre infección, sepsis y SRIS. Figura adaptada de Villao et al (2019).

1.3 Introducción a la sepsis

El estado de sepsis es el resultado de una respuesta inadecuada del huésped a una infección que ocasiona disfunción en uno o más órganos. Esto quiere decir que, en la mayoría de los individuos, la respuesta inflamatoria es adaptativa, y contribuye a controlar la infección. Sin embargo, en la sepsis se produce un desequilibrio entre los mecanismos proinflamatorios y antiinflamatorios, ocasionando una liberación masiva de mediadores y creando una cadena de eventos que producen daño tisular (sepsis grave) o una situación de hipotensión refractaria (shock séptico) (Chiscano-Camón, 2022).

En Estados Unidos, las hospitalizaciones relacionadas con sepsis han ascendido a aproximadamente 2.5 millones en 2021, evidenciando un incremento progresivo en su carga hospitalaria (Agency for Healthcare Research and Quality [AHRQ], 2024). Es la principal causa de muerte en las UCI no cardiacas (Font et al., 2020). En España, cifras recientes estiman una incidencia de entre 100 y 150 casos de sepsis por cada 100.000 habitantes/año (≈ 50.000-75.000 casos al año), con una mortalidad anual cercana a los 20.000 fallecimientos (*Gaceta Médica*, 2023).

1.3.1 Definiciones de sepsis

Aunque el término sepsis no fue acuñado hasta 1991, ya en el siglo XIX se teorizaba que la disfunción orgánica no sucedía internamente, sino como consecuencia de un factor externo, como organismos o gérmenes. Esta revisión incluye tres definiciones oficiales (Ackerman et al., 2021).

La *American College of Chest Physicians* (ACCP) y la *Society of Critical Care Medicine*, acordaron un primer acercamiento a lo que llamaron Sepsis 1,

además de sepsis grave, shock séptico y síndrome de respuesta inflamatoria sistémica (Bone et al., 2992). En esa época, sepsis 1 se consideró como la manifestación de dos o más criterios contenidos en el SRIS, considerados en la tabla 1. Esta definición se ajustó en 2001, para incluir la disfunción orgánica, pero no fue hasta 2016 que se desarrolló la nueva definición (sepsis 3), y se propuso el estado de sepsis como la "disfunción orgánica potencialmente mortal causada por una respuesta no regulada del organismo ante una infección", junto con unos nuevos criterios clínicos contemplados en la tabla 1 (Singer et al., 2016).

Criterios	Sepsis 1 (1991)	Sepsis 2 (2001)	Sepsis 3 (2016)
Parámetros generales	Temperatura >38ºC o <36ºC. FC >90 lpm FR >20 cpm PaCO2 <32 mmHg	Temperatura >38ºC o <36ºC. FC >90 lpm FR >30 cpm Estado mental alterado PAM <70 mmHg, PaO2/FiO2 <300.	Temperatura >38ºC o <36ºC Infección documentada o sospechada con aumento de >2 puntos SOFA. Puntuación GCS <15, PAS <100 mmHg, (PAM <65 mmHg shock) FR >22 lpm.
Parámetros inflamatorios	Glóbulos blancos >12.000/mm3 o <4.000/mm3.	Glóbulos blancos >12.000/mm3 o <4.000/mm3, plaquetas <100.000/µL	Glóbulos blancos >12.000/mm3 o <4.000/mm3, plaquetas <100.000/µL
Parámetros bioquímicos		Glucosa >110 mg/dL, PTC <2 SD por encima del valor normal, creatinina >0,5 mg/dL, bilirrubina >4 mg/dL.	Lactato >18 mg/dL glucosa >110 mg/dL, PTC <2 SD por encima del valor normal, creatinina >0,5 mg/dL, bilirrubina >4 mg/dL.

Abreviaturas: PaO2, presión parcial de oxígeno; FiO2, fracción inspirada de oxígeno; PAM, presión arterial media; PAS, presión arterial sistólica; PTC, procalcitonina; FR, frecuencia respiratoria; FC, frecuencia cardiaca; SOFA, *Secuential Organ Failure Assessment*.

Tabla 1. Criterios de definición de sepsis en 1991, 2001 y 2016.

1.3.2 Fisiopatología de la sepsis

La fisiopatología de la sepsis es un proceso complejo que involucra diferentes partes del sistema inmune. Cuando un germen o microorganismo (virus, bacterias, hongos), entra en contacto con un individuo y provoca una infección, se desencadena una respuesta del huésped en la que diversos mecanismos proinflamatorios y antiinflamatorios tratan de contrarrestar la acción del patógeno y eliminarlo, para la recuperación de los tejidos y órganos afectados (Hotchkiss et al., 2013).

La sepsis comienza cuando un germen es detectado y reconocido por células del sistema inmune, principalmente los macrófagos (Raymond et al., 2017). Un grupo importante de receptores presentes en estas células son los *toll-like*, como el TLR-4 y TLR-2, además del CD-14 y el MD-2 (activador de los TLR). Los TLR desempeñan un papel crítico en la activación de la respuesta inflamatoria. Cuando la membrana de un germen libera PAMPs, estas son captadas por los TLR, quienes dan la orden de liberar las citoquinas y mediadores que participan en la inflamación, tanto proinflamatorios como antiinflamatorios (Gómez et al., 2015). En condiciones normales, la acción inflamatoria es compensada y contrarrestada por la antiinflamatoria, con citoquinas como la IL-10 y el TGF-β. Durante la sepsis, la respuesta inflamatoria no está regulada y se descontrola, eventualmente causando inmunosupresión. El endotelio tiene un papel importante en la sepsis, pues su actividad reguladora es notable durante la inflamación. En condiciones normales se encarga de controlar la coagulación y la fibrinolisis, la permeabilidad capilar, el tono vascular y la adhesión y migración de las células inmunitarias. Durante la sepsis, la actividad del endotelio se altera significativamente, aumentando mucho la permeabilidad patológica a nivel de los grandes vasos. Esto hace que se active la cascada de la coagulación, lo que a su vez impulsa la actividad proinflamatoria. En estas condiciones, los factores anticoagulantes dejan de funcionar como signo de respuesta al estrés (Ince et al., 2016). Durante la fase temprana de la sepsis, se produce una vasodilatación por un aumento de la producción de óxido nítrico en la circulación general, junto con una vasoconstricción en la microcirculación. El resultado es la hipotensión característica de la sepsis a nivel sistémico, y la hipoperfusión tisular debido a la microcirculación. Cuando disminuye, por tanto, el flujo sanguíneo a los tejidos, se producen despolarizaciones en las membranas celulares, causando la acumulación de calcio intracelular y más producción de óxido nítrico. Todo esto induce más inflamación y disfunción mitocondrial, lo que inevitablemente acaba causando necrosis y apoptosis celular (Singer et al., 2016).

Durante la sepsis, hay un incremento de la activación de las adhesinas del endotelio, causando la atracción y adhesión de leucocitos al área afectada y amplificando la respuesta inflamatoria (Raia & Zafrani, 2022). Esto culmina en un estado hiperdinámico caracterizado por un aumento de la frecuencia cardíaca y disminución de las resistencias arteriales sistémicas, consecuentemente aumentando el gasto cardiaco para suplir la demanda de oxígeno. La presencia de un estado hipermetabólico también es notable, porque en condiciones de estrés, el organismo tiende a aumentar la producción de glucosa, pero si este estado persiste, la proteólisis dañará los músculos, eventualmente gastando las proteínas de los músculos de órganos vitales como el corazón o los músculos respiratorios (Angus, 2013). Los diferentes mecanismos de acción de la sepsis se recogen en la figura 2.

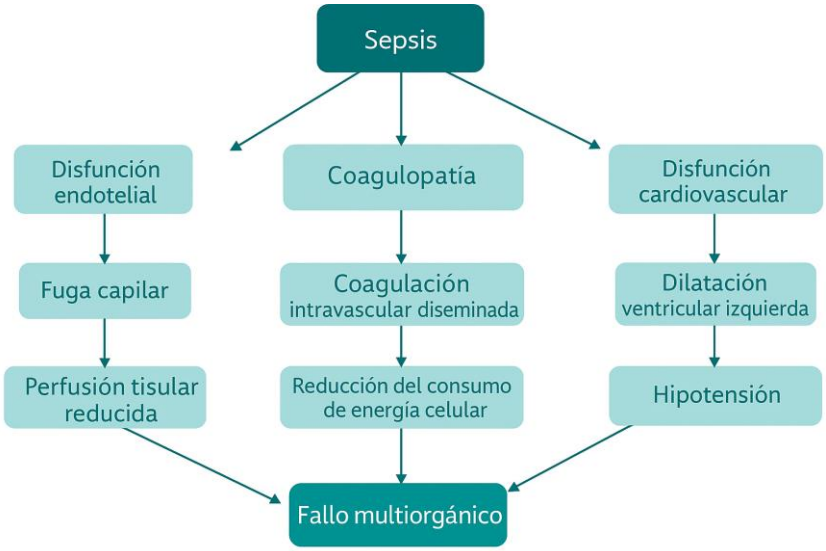


Figura 2. Mecanismos de la fisiopatología de la sepsis. Figura adaptada de Evans T, (2018).

1.3.3 Factores de riesgo y manifestaciones clínicas

La septicemia es el resultado de una infección, sin embargo, es común que haya una causa subyacente o una condición que predisponga a esa condición (Rudd et al., 2020). Los factores de riesgo para desarrollar sepsis incluyen fragilidad, estado inmunocomprometido (debido a tratamientos como quimioterapia, los cuales producen neutropenia), procedimientos quirúrgicos recientes, supervivientes de sepsis anteriores, bacteriemias de origen desconocido y comorbilidades como cáncer, enfermedad renal, enfermedad pulmonar, enfermedad hepática, VIH y diabetes. Antecedentes de patologías cerebrovasculares, dislipemia, obesidad, hábito de fumar, enfermedad pulmonar obstructiva crónica (EPOC), diálisis, tratamiento con corticoides, alcoholismo, etc. también se han recogido como factores de riesgo. El riesgo de contraer sepsis aumenta con la edad, las personas mayores de 65 años tienen más posibilidades de desarrollarla. Además, hay un mayor riesgo en

neonatos y en mujeres embarazadas (Bladon et al., 2024). Las poblaciones de áreas más desfavorecidas o en situación vulnerable también tienen más posibilidades de padecer este estado (Fleischmann-Struzek et al., 2020). Diversos estudios han demostrado que existe una predisposición genética al desarrollo de la sepsis (Chávez et al, 2013). Algunos pacientes poseen unas características tales como mutaciones (polimorfismos) en ciertas citoquinas y mediadores de la inflamación, como el TLR-4, lo que les hace tener una mayor o menor susceptibilidad a la enfermedad, influyendo en la tasa de mortalidad. En el contexto etiológico, se pueden distinguir dos tipos de sepsis: nosocomial y comunitaria.

La sepsis nosocomial se define como aquella que ocurre cuando un paciente ingresado en el hospital contrae una infección, es decir, no son la causa del ingreso ni estaban en periodo de incubación en el momento de internamiento del paciente (Mestanza, 2017). Este tipo es menos común que las comunitarias, sin embargo, se asocian con mayor mortalidad. Son comunes las infecciones urinarias debido al sondaje vesical o catéteres, las respiratorias debido a la intubación orotraqueal, la nutrición parenteral, el estado de coma, los fármacos y el abuso de antibióticos. Existe un grupo de microorganismos que son frecuentes en las infecciones. Los gérmenes que más se han documentado son las bacterias Gram negativas (63%), seguido de Gram positivas (19%) y hongos (17%). *Escherichia Coli* es el patógeno más frecuente (25%), seguido de Pseudomonas *Aeruginosa* (12%), *Enterococcus faecalis* (11%), *cándida albicans* (9%) y *Klebsiella Pneumoniae* (Martín, 2018). La sepsis comunitaria es aquella que se produce fuera del hospital. Dentro del alto campo de infecciones, destacan por su frecuencia las respiratorias (neumonía), las infecciones del sistema nervioso central y las infecciones del tracto urinario. Los organismos que más se aíslan en este contexto son *Streptococcus Pneumoniae*, seguido de la *Legionella Pneumophila*, *Hamophilus Influenzae*, *Staphylococcus Aureus*, meningococo y *Escherichia Coli* (Singer et al., 2016; Rudd et al., 2020).

Las manifestaciones clínicas propias de la sepsis son inespecíficas y variables entre individuos. Las más frecuentes son insuficiencia respiratoria (en casos graves síndrome de distrés respiratorio agudo), fiebre, desorientación y confusión, mialgias, oliguria, lesiones cutáneas, diseminación hematógena y zonas de necrosis (Sánchez-Conrado, 2018). Estos síntomas y signos pueden ser pistas muy orientativas de la magnitud de un proceso infeccioso La acidosis láctica (causante de la hipoperfusión, frialdad, taquipnea y oliguria), junto con hipoglucemia e hipotensión suelen manifestar la presencia de insuficiencia suprarrenal subyacente. Como consecuencia del daño miocárdico, puede producirse una disminución de las resistencias periféricas, aumento de la frecuencia cardíaca y del gasto cardiaco, así como disminución

de la fracción de eyección (Bellani et al., 2016; Beesley et al., 2018). Por lo general, las bacteriemias y otros procesos infecciosos tienen un foco o una fuente inicial, sin embargo, hay algunos casos en los que no hay una fuente identificable, y el síndrome séptico es causado por una inmunosupresión que puede hacer que la flora del paciente se altere, por ejemplo, en el aparato gastrointestinal (Langius et al., 2024).

1.3.4 Diferencias entre sepsis y shock séptico

La sepsis se define como una disfunción orgánica potencialmente mortal causada por una respuesta desregulada del huésped a una infección (Evans et al., 2021). El shock séptico, en cambio, constituye la complicación más grave dentro del espectro de la sepsis y se caracteriza por una hipotensión persistente que requiere vasopresores para mantener una presión arterial media ≥ 65 mmHg, junto con un lactato sérico elevado a pesar de una adecuada reanimación con fluidos (Singer et al., 2016). Esta distinción es de remarcada importancia, ya que el shock séptico implica un estado distributivo de hipoperfusión más profundo y se asocia con una mortalidad significativamente mayor. Cuando la hipotensión inducida por sepsis persiste refractaria al tratamiento inicial, se produce shock séptico, que se diferencia de otros tipos de shock por la excesiva producción de mediadores proinflamatorios. Estos aumentan la permeabilidad capilar y reducen las resistencias vasculares sistémicas, lo que conlleva disminución tanto de la precarga como de la poscarga debido al compromiso del retorno venoso (Habimana et al., 2020; Evans et al., 2021). Inicialmente, la reducción del volumen sistólico es compensada por un incremento en la frecuencia cardíaca, dando lugar a un estado hiperdinámico. Si esta situación progresa, el paciente puede evolucionar hacia un shock no compensado con extremidades frías, llenado capilar lento y pulsos filiformes, conocido como shock frío. En fases avanzadas, la hipoperfusión mantenida puede provocar disfunción multiorgánica irreversible y muerte (Mahapatra & Heffner, 2023).

En términos de mortalidad, la sepsis presenta tasas que oscilan entre un 15–25%, mientras que el shock séptico puede alcanzar cifras de entre un 35–55% según la cohorte estudiada (Shankar-Hari et al., 2016; Evans et al., 2021). Esta diferencia subraya la importancia del diagnóstico temprano y el inicio rápido de medidas terapéuticas, con el fin de evitar la progresión desde sepsis hacia shock séptico.

1.4 Diagnóstico y monitorización de pacientes sépticos

La transición clínica de la infección a la sepsis se denomina "tiempo cero". La evaluación inicial del paciente debe empezar con la elaboración de una historia clínica de forma detallada y cuidadosa, prestando especial atención a

la evaluación de cualquier trastorno predisponente, así como de los tratamientos aplicados para dichas afecciones (Arora et al., 2023). Existen numerosos marcadores en los pacientes que pueden indicar un estado de infección y la subsecuente sepsis, los cuales se analizarán a continuación.

1.4.1 Biomarcadores y parámetros relevantes

Es cierto que la sepsis se define en base a alteraciones fisiológicas inespecíficas como los cambios en la temperatura corporal o en la frecuencia cardíaca. La etiología incierta complica el inicio de la terapia antimicrobiana, retrasando el tratamiento, lo que puede influir en la mortalidad de determinados grupos de pacientes con sepsis grave. Los cultivos microbiológicos permiten distinguir la sepsis de afecciones no infecciosas, pero este método carece de sensibilidad y especificidad, además del retraso en la obtención de los resultados por parte del laboratorio. Es por esto por lo que, se han definido una serie de biomarcadores específicos de la septicemia (Pierrakos, 2010).

Se define como biomarcador a aquella molécula medible en una muestra biológica de forma objetiva, sistemática y precisa, cuyos niveles se constituyen en indicadores de que un proceso es normal o patológico, y sirven para monitorizar la respuesta al tratamiento (Julián-Jiménez, 2014). En el presente trabajo se han considerado como biomarcadores de mayor relevancia la procalcitonina (PCT), la proteína C reactiva (PCR), el lactato y la albúmina. Asimismo, se han analizado el recuento de leucocitos, neutrófilos y linfocitos como marcadores indicativos de infección. A continuación, se definen estos marcadores.

- Procalcitonina: Este marcador es la prohormona de la calcitonina, una hormona que se encarga de regular de forma inhibitoria los niveles de calcio en sangre (Srinivasan et al., 2020). Se encuentra en las células C tiroideas y en las endocrinas pulmonares, sin embargo, todos los tejidos del cuerpo tienen el potencial de elaborar PCT, y está compuesta por 116 aminoácidos. Su relación con la sepsis ha sido ampliamente estudiada. Durante esta situación, su producción aumenta considerablemente, pudiendo alcanzar niveles cientos de veces superiores a los normales, por tanto, es un marcador de empeoramiento. Los valores séricos de PCT se correlacionan con la gravedad de la sepsis; disminuyen con su mejoría y aumentan con el declive clínico. (Becker et al., 2010). Como se ha explicado anteriormente, durante la sepsis se produce una hipercalcemia, lo que indirectamente activa la síntesis de calcitonina a partir de la procalcitonina para intentar regular el calcio. Además de su acción

principal, esta hormona aumenta la expresión de los marcadores CD16 y CD14, incrementa las citoquinas derivadas de leucocitos y aumenta el óxido nítrico, entre otras cosas. En la tabla 2 se recoge la interpretación de los resultados de los niveles. Es importante destacar que esta prueba no tiene la sensibilidad suficiente para distinguir entre SIRS y sepsis, por lo que es necesario complementar el diagnóstico con otros parámetros (Zaki et al., 2024).

Resultado PCT (ng/ml)	Nivel	Comentario
PCT <0.5	Sepsis poco probable	Niveles más bajos no descartan infección. Bajo riesgo.
PCT > 0.5 y < 2	Sepsis posible	Riesgo moderado
PCT > 2 y < 10	Sepsis muy probable	Alto riesgo de progresión a sepsis grave.
PCT > 10	SRIS importante debida a sepsis grave o shock séptico	Alto riesgo de progresión a shock séptico.

Abreviaturas: PCT, procalcitonina; SRIS, síndrome de respuesta inflamatoria sistémica.

Tabla 2. Escala de niveles de elevación de PCT. Adaptada de (Pitarch, 2008).

Todos estos factores indican a que la PCT es el principal biomarcador para contribuir a detectar o descartar la sepsis de forma temprana.

- Lactato (ácido láctico): El lactato se descubrió por primera vez en sangre en el siglo XIX y desde siempre se ha usado como predictor del pronóstico en enfermedades críticas (Nguyen, 2011). Es el producto de desecho del metabolismo anaerobio de los mamíferos. Cuando hay suficiente oxígeno, la ruta metabólica clásica es la de la respiración celular a través de las mitocondrias, lo cual produce abundantes moléculas de adenosín trifosfato (ATP) por glucosa, lo que se traduce en mucha energía utilizable. Sin embargo, cuando el oxígeno no está disponible, las mitocondrias no pueden utilizar la nicotinamida adenina dinucleótido hidruro (NADH) (producto de la glucosa), quedando la fermentación como única opción metabólica, cuyo producto es el lactato que debe ser excretado para no acumularse en los tejidos (Rabinowitz, 2020). Los niveles séricos de lactato se interpretan como un marcador de hipoxia tisular, por tanto, la medición a menudo se incorpora en el manejo clínico de enfermedades críticas, particularmente en casos de sepsis grave y shock séptico, que como se concluyó anteriormente, tienen un papel determinante en el estado de

hipoperfusión de los tejidos. En este contexto, los niveles elevados de ácido láctico denotan la gravedad de la enfermedad y son un predictor de mortalidad. Una elevación del lactato de > 2 mmol/L se considera patológica (Chertoff et al., 2015).

- Proteína C reactiva (PCR): Es una proteína plasmática sintetizada en el hígado que forma parte del sistema inmunitario innato y que participa en la respuesta sistémica a la inflamación junto con los reactantes de fase aguda. Su concentración se eleva notablemente en plasma durante los procesos de inflamación aguda, por tanto, es un potencial marcador de sepsis en la práctica clínica (Wu et al., 2017). En condiciones normales, la PCR actúa como una proteína defensiva que ayuda a neutralizar agentes externos y facilita la eliminación de restos celulares. Durante la sepsis se produce una elevación mantenida de los niveles de PCR como una señal de desequilibrio metabólico, lo que es un predictor de mortalidad. Unos niveles de PCR < 8 ml/L se asocian con bajo riesgo de septicemia, y unos niveles de PCR > 20 mg/L en pacientes con clínica compatible con sepsis suelen desvelar un origen bacteriano en vez de vírico (Markanday et al., 2008).
- Albúmina: Es la proteína más abundante en el plasma sanguíneo y es producida por el hígado. Su concentración normal en sangre oscila entre 3.5 y 5 g/dL, y es fundamental para el mantenimiento de la presión oncótica y la distribución de los líquidos corporales (Erdogan et al., 2021). Esto quiere decir que previene la filtración de fluidos al espacio extravascular. Las causas más comunes de una disminución de los niveles de albúmina son las enfermedades hepáticas como el fallo hepático, fallo renal crónico, síndrome nefrótico o malnutrición. Como se ha visto antes, la sepsis se asocia con un incremento en la permeabilidad de los capilares, y si no hay suficiente albúmina para retener los líquidos, la sepsis puede agravarse y poner en riesgo la supervivencia del paciente (Wiedermann, 2021).

1.4.2 La escala SOFA

Para valorar la gravedad clínica de los pacientes con sospecha de sepsis, se emplea la Evaluación Secuencial del Fallo Orgánico (*Secuential Organ Failure Assessment*, SOFA). Esta herramienta permite evaluar el grado de disfunción orgánica a través de seis sistemas fisiológicos: respiratorio, cardiovascular, hepático, renal, neurológico y de coagulación. Cada parámetro se puntúa de 0 a 4, en función del nivel de alteración observado, lo que proporciona una estimación objetiva del estado del paciente y su evolución. El incremento de la puntuación total en dos puntos o más se asocia con un mayor riesgo de morbimortalidad, lo que convierte a esta escala en un elemento clave

para la estratificación pronóstica y la toma de decisiones terapéuticas en unidades de cuidados críticos (Marik et al., 2017; Wang et al., 2023). Para la sepsis, a través de la Campaña de Supervivencia a la Sepsis, se sugirieron instrucciones y directrices claras en cuanto a la modificación de las puntuaciones SOFA, descritas en la tabla 3 (Palencia Herrejón, 2013).

Sistema	**Puntuación**				
	0	1	2	3	4
Respiración					
PaO2/FiO2 (mmHg)	≥ 400	< 400	< 300	< 200 con asistencia respiratoria	< 100 con asistencia respiratoria
Coagulación					
Plaquetas x10^3/microL	≥ 150	< 150	< 100	< 50	< 20
Hígado					
Bilirrubina (mg/dL)	<1.2	1.2 - 1.9	2.0 - 5.9	6.0 - 11.9	> 12.0
Cardiovascular	PAM ≥ 70 mmHg	PAM< 70 mmHg	Dopamina< 5 o dobutamina	Dopamina 5.1-15 o epinefrina ≤ 0,1	Dopamina >15 o epinefrina >0.1
Sistema nervioso central					
Escala de coma de Glasgow	15	13 - 14	10 - 12	6 - 9	< 6
Renal					
Creatinina (mg/dL)	< 1.2	1.2 - 1.9	2.0 - 3.4	3.5 - 4.9	> 5.0
Producción de orina				< 500	< 200

Abreviaturas: FiO2, fracción inspirada de oxígeno; PAM, presión arterial media; PaO2, presión parcial de oxígeno.

Tabla 3. Puntuación de evaluación de insuficiencia orgánica secuencial relacionada con sepsis (Singer et al, 2016).

1.4.3 La escala Apache II

Otra herramienta ampliamente utilizada en la evaluación de pacientes críticos es la Evaluación de Fisiología Aguda y Salud Crónica (*Acute Physiology and Chronic Health Evaluation II, APACHE II*), desarrollada con el objetivo de estimar la gravedad de la enfermedad y predecir la probabilidad de mortalidad hospitalaria en pacientes ingresados en unidades de cuidados intensivos (UCI) (Cerro et al., 2014). Esta escala se basa en la medición de 12 variables fisiológicas (temperatura corporal, presión arterial media, frecuencia cardíaca, frecuencia respiratoria, oxigenación arterial, pH arterial, sodio sérico, potasio sérico, creatinina sérica, hematocrito, recuento de leucocitos y nivel de conciencia), obtenidas durante las primeras 24 horas de ingreso, a las que se suman puntos en función de la edad y la presencia de comorbilidades crónicas. El resultado final, expresado como una puntuación global, permite estratificar el riesgo y orientar decisiones terapéuticas. A diferencia de la escala SOFA, el sistema APACHE II no está diseñado específicamente para la sepsis, pero su uso sigue siendo frecuente en estudios clínicos y en la práctica asistencial por su capacidad pronóstica robusta y validada, de ahí su utilización en este trabajo como marcador de gravedad de los pacientes (Yang et al., 2024).

1.4.5 Pruebas diagnósticas

En el abordaje diagnóstico de la sepsis, las pruebas de laboratorio juegan un papel fundamental para la identificación etiológica del agente causal y la orientación terapéutica. La elección de las muestras biológicas y de las técnicas microbiológicas específicas debe ajustarse al posible foco infeccioso, con el fin de optimizar el rendimiento diagnóstico. A continuación, se detallan las recomendaciones principales para la recogida y análisis de muestras según el foco sospechado de infección (Guzmán González, 2019):

- Foco respiratorio: Cultivo de esputo, cultivo y tinción de Gram de líquido pleural, hemocultivo.
- Foco abdominal: Hemocultivo, pruebas de imagen para descartar colecciones, tinción de Gram y cultivo de material purulento.
- Foco urológico: Hemocultivo, urocultivo, tinción de Gram y cultivo de material purulento.
- Foco de piel y partes blandas: Hemocultivo, tinción de Gram y cultivo de muestras de tejido, aspiración de secreciones de úlceras, biopsias.
- Foco de dispositivos intravasculares: Hemocultivo.
- Foco sistema nervioso central: Hemocultivo, tinción de Gram, cultivo y determinación antigénica de líquido cefalorraquídeo, tinción de Gram, cultivo y *Ziehl* de material obtenido de punciones o abscesos.

1.4.6 Tratamiento

El manejo inicial de la sepsis incluye maniobras básicas de reanimación (administración de fluidos intravenosos, uso de vasoactivos/inótropos, transfusión de glóbulos rojos y soporte ventilatorio), para restablecer una entrega adecuada de oxígeno a los tejidos, así como la administración de antibióticos y el control del foco infeccioso (Bruhn et al., 2011). Estas medidas iniciales deben estar dirigidas hacia la normalización de la perfusión en los tejidos. Según las directrices publicadas por la *Surviving Sepsis Campaign*, el protocolo recomendado incluye los siguientes pasos: manejo hidroelectrolítico, administración de antibióticos, soporte hemodinámico y manejo respiratorio. Es clave una monitorización adecuada durante el estado de sepsis o shock séptico del paciente. La monitorización hemodinámica básica debe incluir la presión arterial invasiva, el trazado electrocardiográfico y la oximetría de pulso. Antes de que se disponga de los resultados de laboratorio y antibiogramas, debe iniciarse el tratamiento antibiótico, antivírico o antifúngico empírico adecuado. El control de la fuente es crucial, puede incluir el drenaje de un absceso, desbridamiento del tejido necrótico infectado, extracción de un dispositivo infectado, etc. El control del foco se debe implementar tan pronto como sea posible, incluso inmediatamente después de la reanimación inicial (Evans et al., 2021).

1.5 Aprendizaje automático en relación con la sepsis

En los últimos años, la inteligencia artificial (IA) ha emergido como una herramienta disruptiva en el ámbito de las ciencias de la salud, con un potencial significativo para transformar el diagnóstico, el pronóstico y el tratamiento de múltiples patologías. No existe una única definición de inteligencia artificial. Se podría valorar como una rama de la informática que se ocupa del desarrollo de sistemas capaces de realizar tareas que, tradicionalmente, requieren de la inteligencia humana. Estas tareas incluyen el razonamiento, el aprendizaje, la toma de decisiones, el reconocimiento de patrones, el procesamiento del lenguaje natural y la percepción visual, entre otras (Jiang et al., 2022).

En el ámbito técnico, la IA se sustenta en múltiples disciplinas como las matemáticas, la estadística, la lógica computacional y las ciencias cognitivas. Su evolución ha sido posible gracias al desarrollo de algoritmos complejos y al acceso a grandes volúmenes de datos (*big data*), así como al aumento en la capacidad de procesamiento de los sistemas informáticos (Batko et al., 2022). Entre otras aplicaciones, la IA tiene el potencial de reducir errores médicos mediante la integración y análisis de grandes volúmenes de datos clínicos, lo que permite una interpretación más precisa y oportuna de la información

médica. Además, la automatización de tareas repetitivas puede aliviar la carga administrativa de los profesionales de la salud, permitiéndoles dedicar más tiempo a la atención directa del paciente y fortaleciendo así la relación médico-paciente (Hamet et al., 2017).

El aprendizaje automático, o *machine learning (ML)* es un subcampo de la IA. Un algoritmo de *machine learning* es un proceso computacional que utiliza datos de entrada para lograr una tarea deseada sin estar literalmente programado para producir un resultado particular (Li et al., 2015). Estos algoritmos se adaptan automáticamente su arquitectura mediante la repetición (experiencia) para mejorar cada vez más en el logro de la tarea deseada. El término "Maquinaria Inteligente" fue acuñado en 1950, y desde entonces los informáticos se han dedicado al avance del aprendizaje automático (Shinde, 2018).

Como se ha incidido anteriormente, la sepsis es una afección con una evolución clínica rápida y heterogénea, con una gran cantidad de factores involucrados. Para su diagnóstico, se debe tener en cuenta una serie de signos y síntomas, además de los biomarcadores analizados en apartados previos. En este contexto, el uso de técnicas de *machine learning* resulta especialmente pertinente, ya que permite analizar grandes volúmenes de datos clínicos en tiempo real y detectar patrones sutiles que pueden pasar desapercibidos mediante el juicio clínico convencional. A diferencia de los métodos estadísticos tradicionales, los algoritmos de aprendizaje automático pueden adaptarse a la complejidad y variabilidad de la respuesta inflamatoria sistémica, generando modelos predictivos personalizados que anticipan la aparición de la sepsis con horas de antelación. Esto se traduce en una mejora sustancial en la toma de decisiones clínicas, al facilitar intervenciones tempranas y reducir tanto la mortalidad como las complicaciones asociadas (Moor et al., 2021). En este contexto, este trabajo se enmarca en esta línea de investigación, aplicando modelos supervisados sobre bases clínicas hospitalarias para desarrollar un sistema predictivo de riesgo en pacientes con sepsis, con el fin de apoyar la toma de decisiones clínicas e intervenir de manera temprana.

1.6 Hipótesis

En el presente TFG, se plantea una hipótesis clara en relación con el análisis predictivo del deterioro de los pacientes sépticos, a través de herramientas de *machine learning*. Se teoriza que *es posible prever el deterioro de los pacientes recientemente ingresados en la UCI por sepsis y shock séptico, con el uso de algoritmos de machine learning*. Esto implica que existen marcadores y predictores de sepsis que se pueden utilizar para determinar su

relación con el deterioro clínico utilizando herramientas de aprendizaje supervisado y así adaptar el tratamiento a pacientes críticos, reduciendo las tasas de mortalidad. Esta hipótesis se sustenta en la literatura científica existente y servirán como base para la realización de experimentos y análisis en el marco del presente Trabajo de Fin de Grado.

1.7 Objetivos

En base a la hipótesis previamente descrita, el objetivo principal de este TFG es desarrollar y evaluar un método basado en *machine learning* para predecir el riesgo de desarrollar sepsis o shock séptico en pacientes ingresados en la Unidad de Vigilancia Intensiva del Hospital Clínico Universitario de Valladolid. Todo ello con el propósito de aplicar los algoritmos a futuros pacientes para, de una forma temprana, saber quiénes corren más riesgo y por tanto necesitan una focalización y vigilancia más personalizada en su tratamiento. Para conseguir este objetivo general, se han descrito estos objetivos específicos:

1. Recoger una base de datos de registros de pacientes con variables estandarizadas relevantes en el contexto de la sepsis.
2. Realizar una revisión bibliográfica de las técnicas de aprendizaje automático, que serán utilizadas para lograr el cometido.
3. Desarrollar modelos de *machine learning* para predecir el riesgo de deterioro en pacientes ingresados en la UCI.
4. Entrenar y validar los modelos desarrollados empleando un subconjunto de los datos recogidos.
5. Comparar los diferentes modelos desarrollados empleando un subconjunto de los datos recogidos.

1.8 Descripción del documento

El presente TFG se encuentra estructurado en 6 capítulos:

Capítulo 1. Introducción: En este actual capítulo, se ha presentado una introducción teórica de los aspectos fundamentales abordados en el trabajo, como pilares principales, la sepsis y la necesidad de aplicar la inteligencia artificial en el contexto clínico.

Capítulo 2. Revisión de modelos predictivos asociados a la sepsis: En este capítulo se realiza una revisión bibliográfica de las técnicas computacionales empleadas en la elaboración del algoritmo y su relación con la clasificación y predicción de parámetros clínicos. Posteriormente, se analizará cual es objetivamente la metodología más adecuada para este caso.

Capítulo 3. Materiales y métodos: En este capítulo se describen detalladamente la metodología seguida para la elaboración del trabajo. Comenzará con una breve descripción de la base de datos recopilada, y se dará una explicación detallada de los pasos seguidos para su acondicionamiento y preprocesado. A continuación, se describirá el modelo y se entrenará el conjunto de entrenamiento, para después probar el algoritmo en el conjunto de validación y prueba.

Capítulo 4: Resultados: En este capítulo se presentan los resultados obtenidos mediante la implementación del algoritmo desarrollado en este trabajo.

Capítulo 5: Discusión: En este capítulo se contempla la interpretación de los resultados y las aplicaciones más interesantes en este campo. Por último, se compararán los resultados obtenidos con resultados de otros estudios.

Capítulo 6: Conclusiones y líneas futuras: En este capítulo, se mostrarán las conclusiones obtenidas en este trabajo y se comentarán posibles líneas futuras que podrían ser interesantes para dar continuidad a esta investigación.

Capítulo 2. Revisión de modelos predictivos asociados a la sepsis.

2.1 Introducción

Los modelos de *Machine Learning* en el ento rno sanitario han experimentado un auge en su utilización tras la pandemia de COVID-19 (Carobene et al., 2022). La modelización predictiva en la sepsis representa un campo de investigación que abarca el desarrollo de herramientas matemáticas y computacionales capaces de anticipar el diagnóstico, la progresión o los desenlaces de esta condición clínica. Estos modelos constituyen un puente entre los datos clínicos, biológicos y fisiológicos del paciente y las decisiones médicas, transformando variables observables en probabilidades que orienten la práctica asistencial.

Los modelos predictivos aplicados a la sepsis han mostrado un rendimiento prometedor en la práctica clínica. Diversos estudios han evidenciado la superioridad de los enfoques basados en *machine learning*, en particular los modelos de árboles de decisión, frente a métodos estadísticos convencionales. La precisión de la predicción depende en gran medida tanto del tipo de modelo utilizado como de las características del conjunto de datos empleado, factores que condicionan su capacidad para identificar patrones complejos en pacientes críticos (Yadgarov et al., 2024). Estos modelos han emergido como herramientas fundamentales en el abordaje clínico de la sepsis, respondiendo a la necesidad crítica de identificación temprana y estratificación de riesgo en una patología caracterizada por su rápida progresión y elevada mortalidad. A diferencia de otras condiciones médicas donde la progresión es más predecible, la sepsis presenta una ventana terapéutica estrecha donde cada hora de retraso en el tratamiento adecuado incrementa significativamente la mortalidad (Kumar et al., 2006; Huang et al., 2024). Esta realidad clínica ha impulsado el desarrollo de diversos sistemas predictivos que han evolucionado desde simples escalas de puntuación clínica hasta sofisticados algoritmos de inteligencia artificial, capaces de integrar y analizar simultáneamente múltiples variables biológicas, clínicas y fisiológicas (Gao et al., 2024).

Los sistemas de aprendizaje automático supervisado aprovechan los datos clínicos de los registros electrónicos (EHR) y señales fisiológicas para generar modelos predictivos de riesgo que pueden superar a las escalas tradicionales de riesgo (como SOFA, qSOFA o escalas de alerta temprana) (Fleuren et al., 2020), y han demostrado un rendimiento general superior sobre estos sistemas de puntuación comúnmente utilizados en el momento del inicio de la sepsis, utilizando signos vitales, resultados de pruebas de laboratorio y variabilidad demográfica (Camacho-Cogollo et al., 2022).

Por ejemplo, Taylor et al. (2016) entrenaron un modelo *Random Forest* con más de 500 variables de registros de urgencias, alcanzando un área bajo la

curva ROC (AUC-ROC) de 0.86 para predecir la mortalidad intrahospitalaria en sepsis, un rendimiento superior al de las herramientas clínicas convencionales. De forma similar, Desautels et al. (2016) desarrollaron un clasificador basado en *gradient boosting* para la detección temprana de sepsis en UCI, obteniendo un AUC-ROC de 0.83 y demostrando la utilidad de los modelos de aprendizaje automático en entornos críticos. Otros estudios, como el de Nemati et al. (2018), aplicaron *machine learning* en grandes cohortes de pacientes críticos y alcanzaron valores de AUC comparables (0.80–0.85), mostrando además la capacidad de estos algoritmos para generar predicciones en tiempo real. Estos resultados individuales se enmarcan en la tendencia descrita por revisiones sistemáticas recientes, que reportan una discriminación generalmente buena a excelente (AUC 0.70–0.90) en cohortes retrospectivas, aunque con una notable heterogeneidad metodológica entre trabajos (Kausch et al., 2021).

2.2 Tipos de datos y características usadas

Los modelos de riesgo en sepsis emplean datos clínicos de diversa naturaleza. Predomina el uso de variables procedentes de historias clínicas electrónicas: edad, comorbilidades, resultados de laboratorio (lactato, leucocitos, creatinina, etc.), y valores vitales (frecuencia cardiaca, TA, respiratoria, temperatura) (Yang et al., 2023).

En algunos estudios se usan datos longitudinales (series temporales de constantes vitales o laboratorios) que permiten predecir eventos por adelantado (p. ej. 3–6 horas antes del deterioro) (Spoto et al., 2022). Otros modelos simplifican la entrada a variables clínicas básicas: por ejemplo, Cheng et al. analizaron únicamente la dinámica de signos vitales (presión arterial sistólica, presión arterial diastólica, frecuencia cardiaca, frecuencia respiratoria, temperatura) junto con el puntaje SOFA con resultados prometedores en mortalidad intrahospitalaria (Cheng et al., 2022). La calidad del dato (completitud, frecuencia de medición) es crucial; modelos como los de Giannini et al. (2019) requirieron definiciones precisas de sepsis (cultivo positivo + lactato elevado o hipotensión) para entrenar su *Random Forest.* Además, la heterogeneidad en la definición de sepsis y en la adquisición de datos es un reto metodológico destacado.

2.3. Bases de datos utilizadas en investigación (MIMIC)

La base de datos *MIMIC (Medical Information Mart for Intensive Care),* ha sido fundamental en el desarrollo de modelos predictivos de sepsis. El estudio de cohorte retrospectivo de Liu et al. (2025) utilizó datos de la base de datos *MIMIC-IV*, desarrollando dos modelos: Modelo 1 basado únicamente en signos vitales, y Modelo 2 incorporando signos vitales, características

demográficas, historial médico y quejas principales. *MIMIC* proporciona datos de alta calidad de pacientes de cuidados intensivos, incluyendo información demográfica, signos vitales, resultados de laboratorio, medicamentos y procedimientos. Esta riqueza de datos ha permitido el desarrollo de modelos más sofisticados y la validación robusta de algoritmos predictivos (Johnson et al., 2023).

2.4 Algoritmos supervisados empleados

Entre los métodos supervisados más usados se encuentran la regresión logística, árboles de decisión, bosques aleatorios (*Random Forest*), máquinas de soporte vectorial (SVM) y métodos de ensamble (p. ej. XGBoost). Estos modelos aprenden a clasificar pacientes sépticos de alto riesgo (p. ej. *shock* séptico, muerte) versus bajo riesgo (Nemati et al., 2018).

Los algoritmos basados en árboles de decisión han mostrado un rendimiento excepcional en la predicción de sepsis debido a su capacidad para manejar variables categóricas y numéricas simultáneamente, así como por su interpretabilidad inherente. *Random Forest*, en particular, ha demostrado ser uno de los algoritmos más efectivos para este propósito (Gao et al., 2024). Esta superioridad se debe en parte a la capacidad del algoritmo para reducir el sobreajuste mediante la combinación de múltiples árboles de decisión débiles. Por ejemplo, bosques aleatorios han sido muy populares debido a su buena precisión y manejo de datos heterogéneos. Taylor et al. (2016) entrenó un modelo de *Random Forest* con 5.278 casos de sepsis del servicio de urgencias, empleando 500 variables EHR con el objetivo de predecir la mortalidad intrahospitalaria, y obtuvo AUC=0.86 (1,056 pacientes en validación). Giannini et al. (2019) desarrolló un clasificador *Random Forest* en 162 mil admisiones hospitalarias, obteniendo una especificidad del 98% pero con una sensibilidad limitada (26%), lo que restringe su utilidad en la detección temprana. Más recientemente, Cheng et al. (2022) usaron solo cambios en los signos vitales para predecir mortalidad séptica dentro de 48 h. Compararon *Random Forest* con redes neuronales (CNN/LSTM) y observaron que RF alcanzó AUC≈0.770 (con 6 horas de anticipación), resultado similar a la CNN. Estos trabajos ponen de relieve el potencial de los modelos de ensamble, pero también evidencian dos limitaciones que aborda este TFG: la pérdida de sensibilidad al aplicarse en poblaciones clínicas desbalanceadas y la escasa exploración de técnicas de balanceo avanzado (p. ej., ADASYN o submuestreo) en cohortes de sepsis. En nuestro estudio, se combinan estos algoritmos con métodos de balanceo para evaluar si es posible mejorar la detección temprana sin sacrificar especificidad.

Otras técnicas basadas en árboles incluyen modelos *Boosted* (p. ej. XGBoost) que combinan múltiples árboles débiles; estudios han reportado que XGBoost frecuentemente alcanza la mejor precisión en entrenamiento, como indica una revisión donde *XGBoost* alcanzó *accuracies* de hasta 0.97 en el conjunto de entrenamiento, lo que lo hace particularmente efectivo en la predicción de sepsis. Su capacidad para manejar datos desbalanceados y su robustez frente a valores atípicos lo han convertido en una opción popular en estudios recientes (Zhang et al., 2023). En la literatura, es habitual combinar modelos predictivos de alto rendimiento con técnicas de interpretabilidad para identificar los factores que impulsan las predicciones. Por ejemplo, Zhou et al. (2024) aplicaron SHAP (*SHapley Additive exPlanations*) para descomponer la contribución de cada variable en modelos de predicción temprana de sepsis utilizando 2.385 pacientes, lo que facilita la interpretación clínica de las predicciones. Liang et al. (2022) usaron *XGBoost* en una cohorte china de UCI con 5.189 pacientes, reportando un AUC de 0.91 para predicción de *shock* séptico a 12 horas, demostrando la aplicabilidad del modelo en poblaciones no occidentales.

Aunque menos común que *XGBoost, LightGBM* ha sido implementado en varios estudios con resultados competitivos. Liu et al. (2021) compararon *XGBoost, LightGBM, Random Forest* y *Logistic Regression* en la base *MIMIC-III* (n=23.106). *LightGBM* logró un AUC de 0.89 para la predicción de sepsis dentro de las primeras 24 horas, siendo el modelo más eficiente en tiempo de entrenamiento. En otro estudio, Yao et al. (2023) combinaron *LightGBM* con técnicas de selección de características y *SHAP values* para aumentar la interpretabilidad del modelo aplicado a 1.245 pacientes. No obstante, la literatura muestra que la mayor eficiencia de LightGBM suele implicar menor estabilidad de resultados en datasets pequeños (Ke et al., 2017).

Aunque los modelos más avanzados han ganado popularidad, la regresión logística (LR) sigue siendo un algoritmo fundamental en la predicción de sepsis debido a su simplicidad, interpretabilidad y eficacia demostrada (Zhang et al., 2024). La regresión logística multivariable ha demostrado ser particularmente útil para identificar factores de riesgo específicos y cuantificar su contribución individual al riesgo de sepsis. Esto la convierte en una herramienta valiosa tanto para la predicción como para la comprensión de los mecanismos subyacentes de la enfermedad (Li et al., 2022). La regresión logística se usa con frecuencia como comparador; por ejemplo, Taylor et al. (2016) comparó un *Random Forest* con un modelo LR y obtuvo AUC=0.86 frente a 0.76 (LR). Nemati et al. (2018) compararon LR con otros modelos en datos de UCI y hallaron que, aunque LR tenía buen rendimiento (AUC 0.82), *XGBoost* y RF lo superaban sistemáticamente (AUC > 0.9). Mao

et al. (2020) desarrollaron un modelo LR en el contexto de atención primaria en China, obteniendo AUC=0.76. A pesar de ser inferior a XGBoost (AUC=0.88), se justificaba por su facilidad de uso clínico.

Las llamadas máquinas de soporte vectorial (SVM) también se han aplicado, con resultados variables. En general, los análisis muestran que los mejores resultados suelen provenir de métodos de ensamble avanzados o redes (aunque en este capítulo nos enfocamos en ML clásico), mientras que modelos simples como LDA o pequeñas redes neuronales obtienen menor desempeño. (Klaush et al, 2020). En el estudio de Zhang et al. (2023) se analizaron 104 modelos de predicción de mortalidad séptica: incluyeron NB, *Random Forest*, ANN, *XGBoost, Logistic Regression*, DT, *k*NN, SVM y LASSO, confirmando la variedad de algoritmos probados en la literatura.

La métrica de referencia más usada es el área bajo la curva ROC (AUC-ROC), que resume sensibilidad y especificidad a distintos umbrales (Fleuren et al., 2020). En la mayoría de los estudios revisados (>80%) se reporta el AUC como indicador principal de discriminación. Adicionalmente se suelen indicar medidas como *sensibilidad*, *especificidad*, valores predictivos y métricas derivadas (F1, precisión). En tareas de supervivencia (riesgo de muerte a cierto tiempo) se usa a veces el *C-index* (índice de concordancia). Por ejemplo, un metaanálisis reciente calculó la mediana del *C-index* de modelos ML (aprox. 0.80–0.85) y encontró que el ML supera a las escalas clásicas de mortalidad (Zhang et al., 2023). También se emplean curvas ROC específicas (p. ej. AUC-PRC) cuando el problema es muy desbalanceado. Es fundamental además validar los modelos, idealmente con datos externos independientes. Sin embargo, la mayoría de los estudios reportan solo validación interna (split train/test o validación cruzada, la cual será la usada en este trabajo). La revisión de Kausch et al. (2021) indica que pocos modelos han sido probados en entornos reales, y solo dos estudios observaron mejoría de resultados clínicos tras su implementación. Flores et al. (2020) advierten que la heterogeneidad en definiciones de sepsis impide agrupar resultados; por ello propusieron que las comparaciones se hagan por *etapa clínica* (UCI, sala, ED). Según su análisis, el AUC óptimo alcanzado varía: p.ej. 0.96–0.98 en hospitalizaciones generales, 0.87–0.97 en urgencias y 0.68–0.99 en UCI.

2.5 Aplicación de balanceadores

El desbalance entre pacientes sépticos graves (minoría) y no sépticos (mayoría), es un reto común en los modelos de predicción basados en ML. Para evitar que los modelos aprendan sesgos hacia la clase mayoritaria, se aplican estrategias de re-muestreo como *SMOTE, SMOTEENN, ADASYN* o submuestreo aleatorio.

SMOTE ha sido ampliamente utilizado en el ámbito clínico para afrontar el desbalanceo de clases. En un estudio reciente con cohortes de sepsis en la base MIMIC-IV, Gao et al. (2024) aplicaron SMOTE para equilibrar la proporción entre pacientes supervivientes y no supervivientes en UCI. El modelo Random Forest entrenado tras el balanceo alcanzó un AUROC=0.94 ± 0.01, mostrando mayor robustez y estabilidad frente a los resultados obtenidos con datos sin balancear. De forma complementaria, Islam et al. (2025) emplearon SMOTE junto con técnicas de modelado generativo (COPULA) en el reto PhysioNet Challenge, que buscaba detectar casos de sepsis temprana. Aunque COPULA mostró el mejor rendimiento global (F1=93–94%), el uso de SMOTE también contribuyó a mejorar de forma significativa la sensibilidad, lo que subraya su utilidad en la detección de la clase minoritaria.

SMOTEENN, que combina sobremuestreo sintético con limpieza de instancias conflictivas, ha mostrado resultados prometedores en dominios médicos generales. Kumar et al. (2023) compararon siete técnicas de balanceo en distintos clasificadores (SMOTE, ADASYN, SMOTEENN, SVM-SMOTE, SMOTETOMEK, *undersampling* y *oversampling*) en un escenario orientado a la predicción de diagnósticos clínicos a partir de datos desbalanceados de salud pública, concluyendo que SMOTEENN proporcionaba un mejor equilibrio entre precisión y sensibilidad. Aunque no se han publicado todavía estudios específicos en sepsis, la evidencia disponible sugiere que esta técnica podría resultar eficaz en escenarios clínicos con gran ruido en los datos.

ADASYN es un balanceador poco utilizado hasta la fecha en sepsis. En la revisión general de desequilibrio clase en ámbito médico, ADASYN aparece como técnica adaptativa con mejoría en clasificación en contextos multicategoría, pero es una propuesta interesante debido a su superioridad de clasificación en *datasets* con ruido, lo que podría ofrecer resultados prometedores con respecto al desbalanceo de clases. (Salmi et al., 2024). En modelado de predicción de complicaciones quirúrgicas, se aplicó ADASYN junto a SMOTE y diversas variantes de *undersampling*. Al-Ahmari et al. (2025) aplicaron ADASYN en la predicción de complicaciones postquirúrgicas. Su objetivo era anticipar eventos adversos en pacientes con bajo número de casos positivos (< 5%). Los modelos entrenados con ADASYN lograron mejor *F1-score* y mejor calibración en comparación con SMOTE, mostrando el potencial de la técnica para situaciones con escasez crítica de ejemplos positivos. En este TFG se pretende abordar el uso de este balanceador en el contexto de predicción de riesgo en sepsis, proporcionando una nueva variante en la investigación gracias a su aportación.

Un estudio reciente diseñó un enfoque de *down-sampling* (submuestreo) personalizado combinado con ventanas deslizantes dinámicas sobre los datos

del *PhysioNet* y FHC para detectar sepsis en tiempo real, entrenando un modelo *XGBoost* (Wu et al., 2024). Con submuestreo se logró AUC de 0.98 en una de las bases (FHC), con sensibilidad del 88 % y especificidad del 95 %. En un estudio más general de datasets quirúrgicos muy desbalanceados para predecir mortalidad postoperatoria, se compararon técnicas como *random undersampling*, *near-miss undersampling* y SMOTE. El *undersampling* proporcionó resultados robustos, incluso superando SMOTE en términos de calibración y generalización, especialmente cuando se combina con modelos robustos como RF o *XGBoost* (Chen et al., 2024).

A partir del análisis de los estudios anteriores, se puede observar que hay mucha variabilidad entre los diferentes estudios. Por una parte, se observa falta de estandarización: distintas definiciones de sepsis (Sepsis-2 vs Sepsis-3), varía la población (ED vs UCI), y la prevalencia de eventos críticos es distinta en cada cohorte, lo que dificulta comparaciones. En la revisión de Fleuren et al. (2016), se destaca que muchos estudios presentan alto riesgo de sesgo (*bias*), especialmente en el análisis estadístico (tamaño muestral insuficiente, manejo de datos faltantes, ausencia de intervalos de confianza). De hecho, se observó una correlación inversa: los estudios con mayor sesgo suelen reportar mejores AUC (por posibles sobreajustes). Además, casi todos los modelos publicados se entrenaron con poblaciones restringidas: p.ej. muchos usan la base pública *MIMIC* con una gran cantidad de pacientes que, sin embargo, solo provienen de Estados Unidos y no de otras partes del mundo. Esto limita la generalización: un modelo que rinde bien en EE. UU. puede fallar en otro sistema de salud con pacientes distintos. Otro punto crítico es la publicación de detalles, como el método y lenguaje de programación, lo que dificulta replicar o validar externamente los modelos. Por estas razones, se considera que, aunque el desempeño reportado de ML es alto, la calidad metodológica aún debe mejorar: se requieren normas de estandarizar el informe (p.ej. TRIPOD) y validaciones externas rigurosas. En la tabla 4 se resumen las características del conjunto de estudios revisado.

Estudio (Autor, año)	Muestra y datos	Balanceo de clases	Modelo principal	Desempeño destacado
Taylor et al. (2016)	ED EE. UU, n=5.278	No	Random forest	Mortalidad intrahosp. AUC = 0.86
Desautels et al. (2016)	MIMIC-II, n=17.000	No	Random Forest	AUC = 0.91
Nemati et al. (2018)	UCI multicéntrica	No	Logistic Regression	AUC = 0.82
Giannini et al. (2019)	Hospital de Filadelfia, EE. UU (n=162.212)	No	Random Forest	Choque séptico. Sens=26%, Esp=98%
Mao et al. (2020)	Atención primaria China (n≈4.500)	No	Logistic Regression	AUC = 0.76
Liu et al. (2021)	MIMIC-III (n=23.106)	No	LightGBM	AUC = 0.89
Cheng et al. (2022)	ED Taiwan, n=193.646	No	Random Forest vs CNN	RF AUC ≈ 0.77, CNN AUC ≈ 0.84
Liang et al. (2022)	UCI China (n=5.189)	No	XGBoost	Shock séptico AUC = 0.91
Yao et al. (2023)	Clínicas múltiples (n=1.245)	No	LightGBM	AUC = 0.87, explicabilidad con SHAP
Hassanzadeh (2023)	Hospital de Besat, Iran	SMOTE	Random Forest, XGBoost	Más robusto tras balanceo en cuanto a la accuracy
Zhao et al. (2024)	PhysioNet Challenge 2019, n=40.336	SMOTE	XGBOOST	F1 ≈ 93–94 % con COPULA; SMOTE mejora recall notablemente
Zhou et al. (2024)	Hospital público de Anhui, n=2.385 (364 sépticos)	SMOTE	Random Forest	AUC=0.87, F1=0.77 tras SMOTE; buena explicabilidad SHAP
Gao et al. (2024)	UCI MIMIC-IV n=17.429	SMOTE	Random Forest	AUROC=0.94 ± 0.01; modelo más robusto tras balanceo
Wu et al. (2024).	Datos PhysioNet n=40.336	Submuestreo	XGBoost	AUC ≈ 0.89–0.98 (sens 84 %, esp 95 %)
Al-Ahmari et al. (2025)	Datos de hospitales de Arabia Saudí, n=64.793	SMOTE + ADASYN + submuestreo	Random Forest, LR, SGB	RF con SMOTE mostró el mejor desempeño MCC=0.2

Tabla 4. Comparación entre los estudios más relevantes acerca de predicción de mortalidad en sepsis con métodos de machine learning.

2.6 Herramientas de explicabilidad (XAI)

La explicabilidad de los modelos de *machine learning* se ha vuelto crucial en aplicaciones médicas, donde la comprensión de las decisiones del modelo es tan importante como su precisión. El análisis SHAP ha sido aplicado junto con múltiples modelos algorítmicos para explorar los principales factores de riesgo para la predicción precisa de la sepsis (Zhou et al., 2024).

2.6.1 SHAP (SHapley Additive exPlanations)

SHAP proporciona valores de contribución para cada característica en cada predicción individual, permitiendo a los clínicos entender qué factores específicos llevaron a una predicción de alto riesgo para un paciente particular. Esta capacidad es especialmente valiosa en la sepsis, donde la heterogeneidad de la presentación clínica requiere un enfoque personalizado (Chen et al., 2022).

Numerosos estudios han hecho uso de la inteligencia artificial explicable o XAI, en concreto de los algoritmos de SHAP, por ejemplo, Shumilov et al. (2025) entrenaron modelos con los datos de *MIMIC-III* aplicando SMOTE para balancear clases y *XGBoost* como modelo principal. Utilizaron SHAP para interpretar la importancia de las 35 variables seleccionadas. El modelo con *Random Forest* y balanceo SMOTE obtuvo AUC = 0.97, precisión 0.93 y *recall* 0.91. SHAP permitió identificar el impacto de cada variable en cada predicción individualmente. En un estudio publicado en el *European Journal of Medical Research*, Li et al. (2024) desarrollaron un modelo *XGBoost* para mortalidad intra-UCI y aplicaron SHAP para evaluar la importancia global (edad, potasio, NLR, ventilación invasiva, albúmina) y comparación con importancia tradicional de *XGBoost*, mostrando que SHAP captura interacciones no lineales ignoradas por métodos clásicos. Algunos autores, como Hu et al, (2022), realizaron un estudio donde entrenaron *XGBoost* en pacientes con sepsis en UCI y usaron SHAP *values* para explicar predicciones individuales. Las variables más relevantes fueron la escala de coma de Glasgow, nitrógeno ureico en sangre, frecuencia respiratoria, volumen miccional y edad. Otros estudios, como el de He et al, (2024), crearon un *XGBoost* explicable para mortalidad séptica junto con *Logistic Regression* con *5-fold cross validation* para mostrar los efectos del índice de masa corporal, nitrógeno ureico en sangre, edad y volumen miccional. Implementaron la explicación con SHAP para interpretar el modelo óptimo, además de proporcionar tres casos clínicos representativos para servir como ejemplo para futuros análisis.

Además de SHAP, otras técnicas como *LIME* (*Local Interpretable Model-agnostic Explanations*) y los gráficos de importancia de características han

sido utilizados para hacer los modelos de sepsis más interpretables. Estas herramientas permiten a los médicos confiar en las predicciones del modelo y comprender los mecanismos subyacentes de la toma de decisiones (Arrieta et al., 2020; Zhou et al., 2024; Liu et al., 2025; Zhu et al., 2025).

2.6.2 Barreras para la implementación

A pesar de los avances técnicos, persisten barreras significativas para la implementación clínica amplia de modelos predictivos de sepsis. Estas incluyen la necesidad de validación prospectiva, integración con sistemas de información clínica, entrenamiento del personal y establecimiento de protocolos de respuesta a las alertas del modelo (Johnson et al., 2023; Shickel et al., 2022). La aceptación por parte de los clínicos también requiere que los modelos no solo sean precisos sino también interpretables y que generen un número manejable de falsas alarmas para evitar la fatiga de alertas (Sendak et al., 2020).

Capítulo 3: Materiales y métodos

3.1 Introducción

En este capítulo se realizará una descripción detallada de la base de datos empleada para este TFG. A continuación, se explican las distintas etapas del método propuesto, desde la limpieza y preprocesado hasta la imputación de lo diferentes modelos de *machine learning*.

3.2 Descripción de la base de datos

La base de datos utilizada en este trabajo corresponde a un registro clínico anonimizado compuesto por 226 pacientes ingresados en la unidad de vigilancia intensiva del Hospital Clínico Universitario de Valladolid, recogidos en el trascurso del año 2024, y el primer trimestre del año 2025.

En esta base está incluido cada paciente etiquetado con el número de historia clínica, y sus respectivas variables clínicas, demográficas, analíticas o microbiológicas relevante para el estudio del riesgo en el contexto clínico analizado. Entre las variables presentes se incluyen datos numéricos (como edad, constantes vitales, parámetros analíticos), categóricos (sexo, motivo de ingreso, factores de riesgo, gérmenes, foco infeccioso), y variables que contribuyeron a definir la variable objetivo, como la evolución de los pacientes durante su estancia, y en los posteriores 60 días. La calidad de la base de datos fue cuidadosamente revisada, identificando y gestionando valores ausentes, inconsistencias en los nombres de las variables y heterogeneidad en los valores de las etiquetas categóricas, aspectos fundamentales para asegurar la robustez de los análisis posteriores.

Para esta investigación, se definieron criterios tanto de inclusión como de exclusión, con el objetivo de delimitar adecuadamente la muestra y garantizar la validez de los hallazgos.

Criterios de inclusión:

- Pacientes mayores de 18 años.
- Pacientes con ingreso en la UCI durante el periodo de estudio.
- Diagnóstico o sospecha clínica de sepsis o shock séptico en el momento del ingreso o durante las primeras 24 horas posteriores.
- Disponibilidad de variables clínicas y analíticas necesarias para el modelado predictivo en las bases de datos (Jimena e ICA).

Criterios de exclusión:

- Edad menor a 18 años, excluyendo de este modo a los pacientes pediátricos.
- Pacientes con un tiempo de estancia en la UCI menor de 24 horas.

- Pacientes sin sospecha de sepsis o shock séptico en el momento de ingreso y/o sin sospecha de sepsis o shock séptico en las horas posteriores al ingreso.
- Sin infección presente en ningún momento de la estancia en la UCI.

Estos criterios de exclusión se han establecido con el fin de reducir la heterogeneidad de la población y descartar afecciones que no tengan que ver con el estado infeccioso de interés.

3.2.1 Datos Clínicos y demográficos

Dentro de la base de datos utilizada en este trabajo, se encuentran tanto variables clínicas como sociodemográficas. Las variables proporcionadas fueron la edad, el sexo, los factores de riesgo del informe de ingreso, el motivo de ingreso en la unidad, la tensión arterial media, la frecuencia cardíaca, la frecuencia respiratoria, la proteína C reactiva, la dosis de aminas (adrenalina/noradrenalina), la necesidad o no de ventilación mecánica invasiva, el foco infeccioso, los niveles de procalcitonina al ingreso y 24 horas después, los niveles de lactato al ingreso y 24 horas después, el porcentaje de linfocitos y leucocitos al ingreso, el nivel de albúmina, el germen presente en la infección, la puntuación de la escala APACHE II, la puntuación de la escala SOFA, el exitus (mortalidad) durante la estancia, y el exitus a los 60 días posteriores al ingreso. Las variables utilizadas en este estudio fueron extraídas de las bases de datos Jimena e ICA, las cuales recopilan información clínica de pacientes ingresados en diferentes áreas hospitalarias, incluyendo unidades de cuidados intensivos en Castilla y León. La selección de estas variables se realizó mediante acceso a la historia clínica individual de cada paciente, con el fin de verificar la presencia de infecciones relevantes para el objetivo del estudio

No todas las variables recogidas en la base de datos se utilizaron finalmente en el estudio. Se eliminaron aquellas variables donde había más de un 25% de datos faltantes, siguiendo las recomendaciones de expertos en el manejo de datos faltantes en investigaciones clínicas (Lee et al., 2021). En las tablas consecutivas se puede ver un análisis detallado del contenido de la base de datos. La tabla 5 contiene el número de pacientes en el *dataset* separados por su clasificación en nivel de riesgo alto y bajo.

Número de pacientes	226
Pacientes de riesgo ALTO	78 (35%)
Pacientes de riesgo BAJO	148 (66%)

Tabla 5. Conjunto y porcentaje de pacientes totales y divididos por grupo de riesgo (alto y bajo).

Al observar el porcentaje de pacientes de alto y bajo riesgo, es evidente que existe un grado de desbalanceo, es decir, no hay un 50% de pacientes de riesgo alto y un 50% de pacientes de riesgo bajo, sino que los pacientes de riesgo alto tan solo son un 35% del total, mientras que la clase mayoritaria es el riesgo bajo, con un 66% del total. Por lo tanto, en este TFG es necesario abordar este desbalanceo entre ambas clases para intentar alcanzar las máximas métricas de rendimiento en los modelos que posteriormente se pondrán a prueba. En las tablas 6 y 7 se refleja un análisis estadístico de las variables obtenidas del laboratorio y de los scores de gravedad de los pacientes.

Variable	Media	Desviación típica	Mediana	Mínimo	Máximo
Edad	67.65	13.3	71	18	93
TAM	80.72	17.44	79.00	41.00	153.00
FC	92.28	22.71	89.00	46.00	167.00
PCR	188.88	127.58	180.00	2.00	667.00
Procalcitonina ingreso	18.56	37.13	3.23	0.01	354.00
Procalcitonina 24h	24.16	69.54	3.31	0.03	781.00
Leucocitos	13.46	8.49	11.57	0.11	60.60
Linfocitos	11.28	13.68	7.00	0.07	91.00
Albúmina	2.95	0.58	2.90	1.30	4.50
Aminas	0.14	0.25	0.06	0.00	2.00

Tabla 6. Análisis estadístico de las variables de laboratorio del dataset.

Score	Media	Desviación estándar	Mediana	Mínimo	Máximo
APACHE II	22.19	8.43	22.0	4.0	49.0
SOFA	6.28	3.45	6.0	0.0	16.0

Tabla 7. Análisis estadístico de los scores del dataset.

En la tabla 8 se muestra el número de pacientes que requirieron ventilación mecánica invasiva.

Ventilación mecánica invasiva	(número de pacientes) *N*	%
No	123	54.4
Sí	103	45.6

Tabla 8. Número y porcentaje de pacientes con ventilación mecánica invasiva.

En la tabla 9 se refleja el número y porcentaje de pacientes con cada factor de riesgo de la base de datos.

Factor de riesgo	**(número de pacientes) *N***	**%**
HTA	103	45.6
DM2	60	26.5
Tabaquismo	30	13.3
Cáncer	26	11.5
Otros	40	17.7

Abreviaturas: HTA: Hipertensión arterial; DM2: Diabetes mellitus tipo 2.

Tabla 9. Número y porcentaje de pacientes con los factores de riesgo representativos de la cohorte.

En la tabla 10 se muestran los focos de infección junto con el número y porcentaje de pacientes de cada uno.

Foco	**(número de pacientes) *N***	**%**
Respiratorio	101	44.7
Urinario	58	25.7
Abdominal	33	14.6
Biliar	10	4.4
Otros*	—	—

*Tabla 10. Número y porcentaje de pacientes con los focos de infección representativos de la cohorte. *Otros: bacteriemia, cardíaco, catéter, desconocido, genital, óseo, piel, sistémico, SNC, ótico (menos frecuentes, <3%).*

En la tabla 11 se incluyen los gérmenes más relevantes de la base de datos junto al número y porcentaje de pacientes que fueron afectados.

Germen	**(número de pacientes) *N***	**%**
Escherichia coli	41	18.1
Staphylococcus aureus	13	5.8
Klebsiella pneumoniae	17	7.5
Candidas	9	4
Pseudomona aeruginosa	9	4
Cultivos negativos	52	23
Otros*	31	13.17

*Tabla 11. Número y porcentaje de pacientes con los gérmenes representativos de la cohorte. *Otros incluye: adenovirus, enterococcus, gripe A, neumococo, proteus, rinovirus, S. epidermidis, S. pneumoniae, SARS-CoV-2/3, etc. (cada uno <5%).*

3.3 Preprocesado de los datos

En este trabajo, el preprocesado se compuso de las siguientes etapas: limpieza de la base de datos, imputación de valores faltantes, transformación de variables categóricas y binarización de la variable objetivo.

3.3.1 Limpieza de la base de datos

Primero se realizó la normalización de los nombres de las variables para estandarizar lo máximo posible cualquier base de datos sin que sea necesario adaptarla manualmente al algoritmo.

Las variables con un alto porcentaje de datos faltantes (más del 25%) se eliminaron del análisis final (Lee et al., 2021). Estas incluyen el lactato y la frecuencia respiratoria, ya que mantenerlas comprometería seriamente la calidad de los resultados. Además, se revisaron y corrigieron los errores en la transcripción de los datos, como por ejemplo inconsistencia en el formato de las variables.

Uno de los retos del *dataset* era la presencia de variables multietiqueta. Algunas variables, como las que indicaban los factores de riesgo, el foco de infección o el germen detectado en la analítica presentaban múltiples categorías y se habían codificado con una nomenclatura variable. Para garantizar la coherencia de los datos, se aplicó un proceso de limpieza y normalización mediante diccionarios de equivalencias, en los cuales se mapearon sinónimos, variantes textuales y abreviaturas hacia una representación unificada. Este enfoque se ha señalado como una estrategia efectiva para la normalización de conceptos clínicos y la reducción de ruido en datos categóricos (Kraljevic et al., 2021). Posteriormente, se recurrió a la técnica de binarización multietiqueta para transformar las categorías en variables binarias, utilizando esquemas de codificación y el transformador *MultiLabelBinarizer* de *scikit-learn*. Estas técnicas permiten que variables que contienen muchas categorías binaricen cada una de ellas mediante la creación de variables *dummy*. Así se permite que los algoritmos de aprendizaje automático procesen las variables categóricas de forma eficiente, evitando sesgos derivados de una codificación ordinal arbitraria (Potdar et al., 2017; Seger, 2024).

Asimismo, variables categóricas simples como el sexo y el motivo de ingreso fueron transformadas mediante *one-hot encoding*, técnica ampliamente utilizada en el preprocesamiento de datos para aprendizaje automático. Este procedimiento genera una columna por cada categoría posible, codificándola como binaria (0/1), lo que facilita la interpretación de los modelos sin introducir relaciones inexistentes entre categorías (Khurana & Anand, 2023).

3.3.1.1 Imputación de valores faltantes

Para la imputación de valores ausentes en variables numéricas se utilizó el algoritmo *k*NN *Imputer* (Imputación por los *k* vecinos más cercanos), que estima los valores faltantes de una variable numérica a partir de los valores de las observaciones numéricas más similares (Scikit-learn, 2021). En este TFG se han utilizado k=5. Formalmente, para una variable (x_j) con valor faltante en la observación *(i)*, el valor imputado es la media de los valores de (x_j) en las *(k)* observaciones más cercanas a *(i),* según una distancia (en este trabajo se usó la Euclídea) definida sobre las variables no faltantes (Scikit-learn, 2021). Para las variables categóricas donde fue necesario, se imputó utilizando la moda, es decir, el valor más frecuente en la columna correspondiente. El uso de *k*NN está ampliamente justificado en estudios clínicos recientes, pues presenta la ventaja de aprovechar la información de observaciones similares, mejorando la consistencia estadística de los datos imputados (Wang et al., 2024). Tras esto, fueron 73 el número de variables seleccionadas. Finalmente, se comprobó que todas las variables tenían 226 valores no nulos (es decir, sin datos faltantes tras el preprocesado), y también el tipo de cada variable (*int64* para enteros, *float64* para decimales, *bool* para variables binarias y *object* para texto). Esto confirmó que la limpieza y la imputación de valores fueron correctas.

Para simplificar el problema de clasificación y facilitar la interpretación del modelo, la variable objetivo RIESGO se binarizó como sigue: $\text{RIESGO}_{\text{binario}} = 0$ si RIESGO = 'BAJO', 1 si RIESGO = 'ALTO'.

Además, otras variables diana como las que representan el exitus (fallecimiento) se recodificaron a valores binarios (0 para "NO", 1 para "SI") para su análisis posterior. Finalmente, a la variable objetivo se le asignó el valor *(y)*.

Binarizar la variable objetivo simplifica el problema, mejora la interpretabilidad y es una práctica común cuando el objetivo es diferenciar dos estados clínicos claros (supervivencia vs no supervivencia, bajo vs alto riesgo). Los modelos de machine learning, especialmente los clasificadores, trabajan de forma más eficiente con codificación binaria y facilita la aplicación de métricas estándar como *accuracy, precision* o *recall*. (Chicco & Tötsch, 2022).

3.4 Análisis exploratorio de datos

Se llevó a cabo un análisis exploratorio de datos (*exploratory data analisis, EDA*) mediante histogramas, diagramas de caja (*boxplots)* y *violinplots*, con el fin de comprender la estructura del conjunto de datos. Este tipo de

visualizaciones gráficas es fundamental para identificar la forma de distribución, la dispersión, valores atípicos y multimodalidades en contextos biomédicos antes de emprender análisis más complejos o aplicar modelos predictivos (Antoine & Demeester, 2018). La distribución de edad se muestra en el histograma de la figura 3.

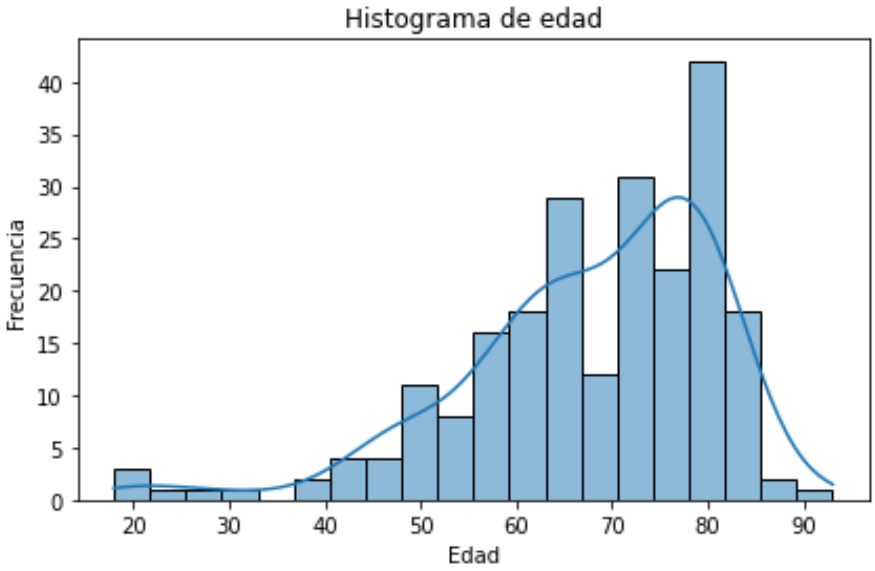


Figura 3: Histograma de la distribución de edad de los pacientes.

La detección de valores atípicos (*outliers*) se realizó mediante *violinplots* (figuras 4 y 5) ya que son especialmente útiles mostrando valores distanciados de los comunes y la densidad de datos en cada valor debido a la forma de la gráfica. Fueron centrados especialmente en variables de laboratorio (albúmina, proteína C reactiva, procalcitonina al ingreso y a las 24h, leucocitos, dosis de aminas, frecuencia cardiaca y tensión arterial media) y scores clínicos (SOFA, APACHE II), ya que estas pueden tener un impacto significativo en la interpretación clínica y el rendimiento de los modelos. También se exploraron diferencias por subgrupos relevantes, como la distribución de sexo (variable estable a lo largo del tiempo y por tanto muy representativa de cada paciente) según el riesgo con un histograma (figura 6).

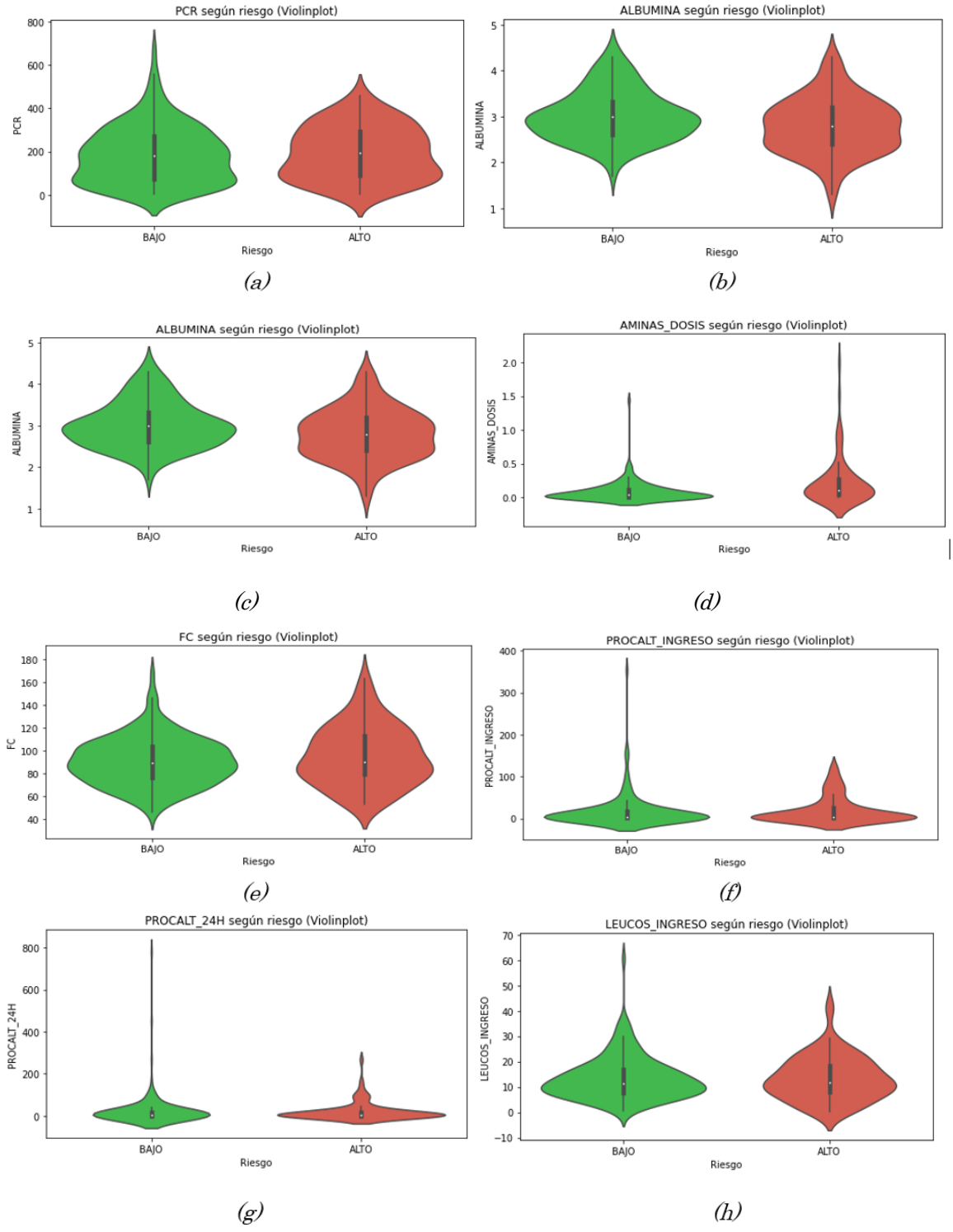
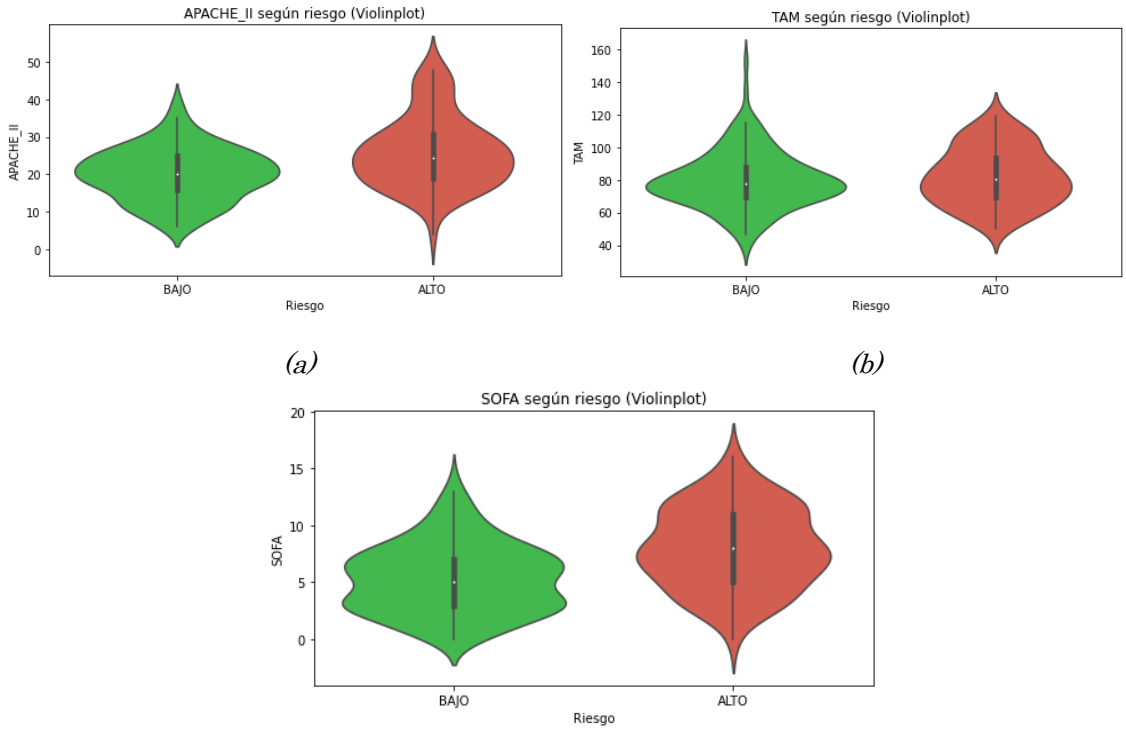


Figura 4. Violinplots de detección de distribución de datos separados en grupos de riesgo alto y bajo. (a): Distribución de la variable "albúmina"; (b): Distribución de la variable "proteína C reactiva"; (c): Distribución de la variable "leucocitos al ingreso"; (d): Distribución de la variable "dosis de aminas"; (e): Distribución de la variable "procalcitonina al ingreso"; (f): Distribución de la variable "procalcitonina a las 24h"; (g): Distribución de la variable "frecuencia cardíaca"; (h): Distribución de la variable "tensión arterial media";



(a) *(b)*

(c)

Figura 5. Violinplots de detección de distribución de datos separados en grupos de riesgo alto y bajo. (a): Distribución de la variable "escala APACHE II"; (b): Distribución de la variable "tensión arterial media"; (c): Distribución de la variable "escala SOFA".

Lo que se pretende con este análisis es entender las variables de las que disponemos. Al mostrar visualmente su relación con los dos tipos de riesgo, podemos señalar las diferencias de cada grupo y los distintos valores que toman las variables.

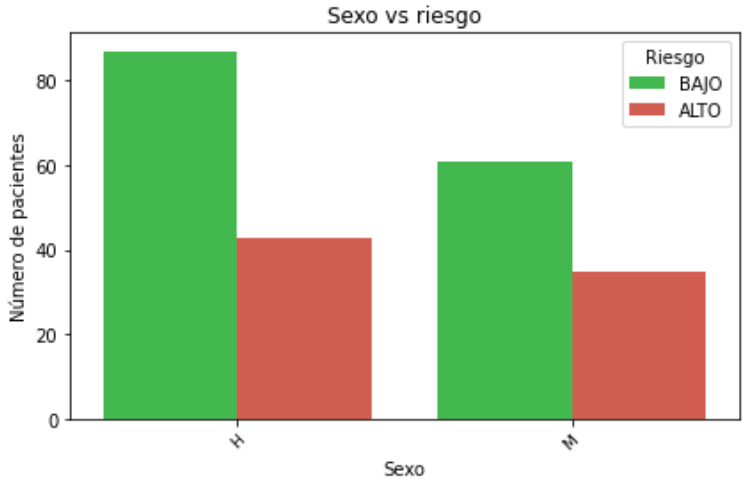


Figura 6. Histograma de comparación del sexo con el riesgo en hombres (izquierda) y mujeres (derecha).

Adicionalmente, y para mostrar una exploración más profunda que nos permita realizar una discusión sobre la naturaleza de los casos, se realizaron distintos análisis exploratorios visuales y estadísticos para describir la distribución de las variables y su relación con el riesgo. Se exploraron especialmente las distribuciones de motivo de ingreso, factores de riesgo, gérmenes, focos infecciosos, sexo y edad en función del nivel de riesgo, identificando patrones relevantes y posibles asociaciones, todas representadas en las figuras a continuación. La figura 7 muestra la distribución del foco de infección, la figura 8 muestra la distribución del motivo de ingreso, la figura 9 muestra la distribución del germen causante de la infección y la figura 10 muestra la distribución de los factores de riesgo, todas ellas con respecto a los dos niveles de riesgo.

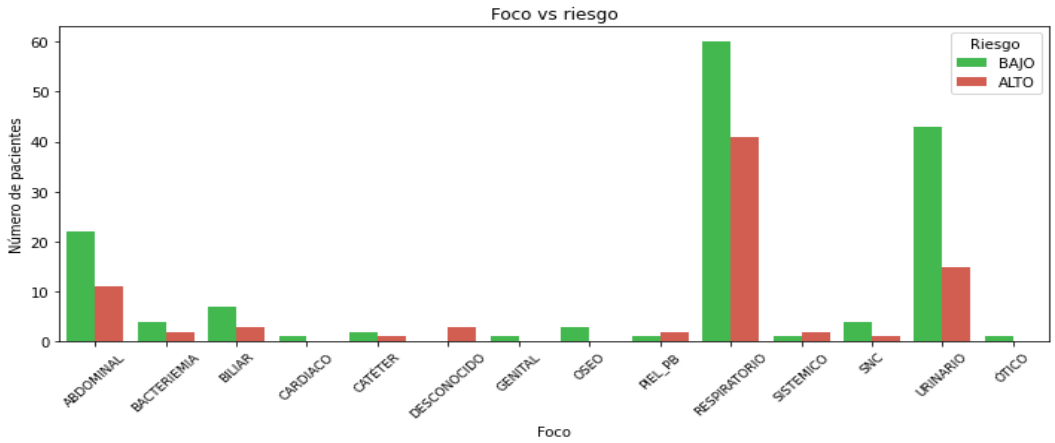


Figura 7: Histogramas de distribución del foco de infección en relación con el riesgo

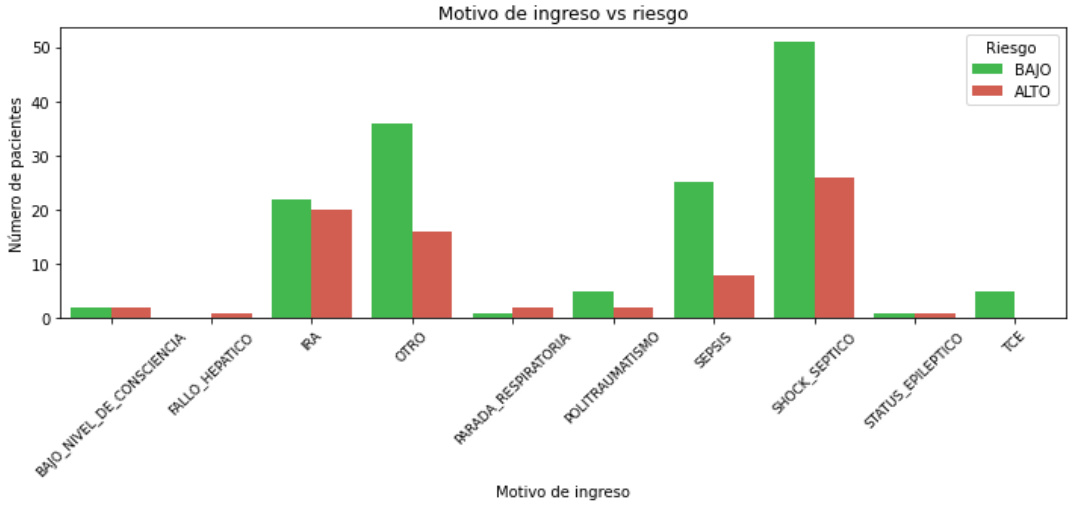


Figura 8: Histogramas de distribución del motivo de ingreso en relación con el riesgo

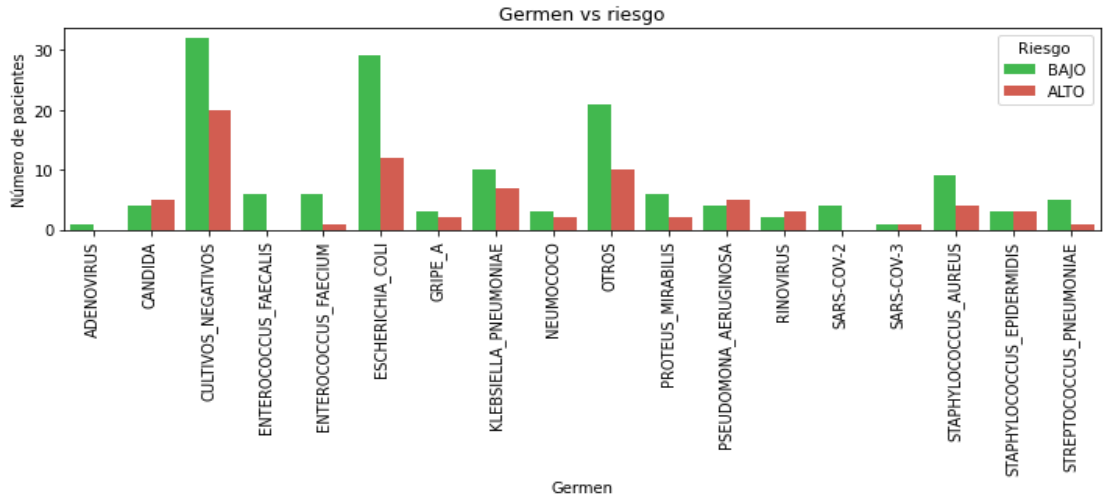


Figura 9: Histogramas de distribución del germen en relación con el riesgo

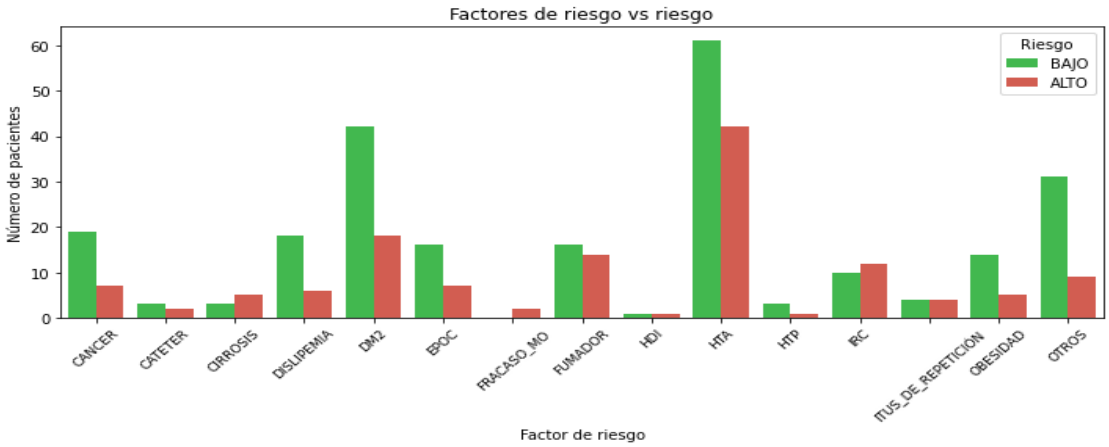


Figura 10: Histogramas de distribución de los factores de riesgo en relación con el riesgo

3.5 Selección de características

Para la selección de características se empleó el método *Fast Correlation-Based Filter* (FCBF), originalmente descrito por Yu y Liu (2003), cuyo criterio de relevancia se basa en la *Symmetrical Uncertainty (SU)*, una medida de dependencia informacional entre variables. El algoritmo FCBF ha sido ampliamente utilizado en contextos biomédicos recientes para reducir dimensionalidad manteniendo interpretabilidad (Li et al., 2018; Jović et al., 2020). Sin embargo, el cálculo de SU requiere variables discretas, por lo que, para *datasets* clínicos continuos como el usado en este trabajo, suele ser necesario aplicar discretización previa. Dado que este proceso puede introducir artefactos y pérdida de información, en el presente trabajo se implementó una adaptación práctica de FCBF sustituyendo SU por correlación de Pearson, ampliamente aceptada para medir asociaciones lineales entre variables continuas en biomedicina (Mukaka, 2012; Akoglu,

2018). Esta decisión metodológica permitió mantener la filosofía del filtro (seleccionar variables relevantes y eliminar redundancia), evitando al mismo tiempo transformaciones adicionales en los datos.

En consecuencia, las variables fueron ordenadas según su correlación absoluta con la variable objetivo, eliminando aquellas con baja relevancia (correlación < 0.08). Posteriormente, se depuraron variables redundantes mediante la correlación mutua (>0.8), y finalmente se seleccionaron las 15 variables más informativas de las 73 originales del conjunto de entrenamiento para el modelado con el fin de evitar y/o reducir el posible sobreajuste y garantizar que cada variable incluida tenga una contribución significativa al rendimiento del modelo (Remeseiro, 2019). Las 15 características más relevantes seleccionadas fueron: la escala SOFA, la escala APACHE II, la dosis de aminas, la albúmina, los focos de infección desconocido, respiratorio y urinario, la insuficiencia respiratoria crónica, el fracaso multiorgánico, la cirrosis, la hipertensión arterial y otros como factores de riesgo, la insuficiencia respiratoria aguda y los traumatismos como motivo de ingreso, y el germen *enterococcus faecalis.*

Finalmente, las variables numéricas seleccionadas se normalizaron mediante una transformación de estandarización, que consiste en centrar cada variable en torno a su media y escalarla en función de su desviación típica, de manera que cada variable resultante tiene media 0 y desviación típica 1. Esto es muy útil en algoritmos de *machine learning* que dependen de la magnitud de las características, ya que evita que variables con rangos mayores dominen la estimación de los modelos y mejora la convergencia en métodos iterativos, como regresión logística o algoritmos de *boosting* (Jantos et al., 2025).

3.6 Definición de balanceadores

Dado el anteriormente comentado desbalanceo de clases (35% pacientes de riesgo alto y 66% pacientes de riesgo bajo), se implementaron distintas estrategias de resampleo, para intentar corregirlo y aumentar la robustez de los algoritmos:

- SMOTE (*Synthetic Minority Oversampling Technique*): Genera sintéticamente nuevas instancias de la clase minoritaria interpolando linealmente entre sus vecinos más cercanos del espacio de características. Para cada muestra minoritaria, encuentra sus k vecinos más cercanos del mismo tipo, después selecciona aleatoriamente uno de ellos y genera una nueva muestra sintética de acuerdo con la siguiente ecuación (Chawla et al., 2002).

$$x_{new} = x_i + \lambda * (x_j - x_i) \quad (1)$$

Donde λ es un número aleatorio uniforme. Entre sus ventajas están la expansión de las regiones de decisión, y además se evita la duplicación exacta. Entre sus limitaciones están la posibilidad de generar ruido por solapamiento entre clases (Kim et al., 2025).

- ADASYN (*Adaptive Synthetic Sampling*): Similar a SMOTE, pero adapta el número de instancias sintéticas generadas en función de la dificultad local (es decir, la distribución de las muestras de cada clase) mediante un índice de densidad *(r)* para clasificar la clase minoritaria. Un índice de densidad alto indica regiones con más muestras mayoritarias, lo que se traduce en más muestras sintéticas (He et al., 2008).

- SMOTEENN (SMOTE + *Edited Nearest Neighbours*): Combinación que aplica SMOTE y luego elimina patrones difíciles de clasificar mediante la técnica *Edited Nearest Neighbours* (Kim et al., 2025). En este caso, las variables se eliminan si se cumple la siguiente condición:

$$\left|\{j \in NN_k(x_i): y_j \neq y_i\}\right| > \frac{k}{2} \quad (2)$$

  Este algoritmo reduce el ruido introducido por SMOTE, pero es posible que elimine tanto ejemplos sintéticos como naturales problemáticos.

- *RandomUnderSampler*: Reduce el número de ejemplos de la clase mayoritaria mediante submuestreo aleatorio. Esta estrategia se seleccionó por su simplicidad y por evitar el sesgo que puede introducir la generación sintética de instancias (como ocurre en SMOTE o ADASYN), permitiendo un entrenamiento más balanceado y reduciendo el tiempo computacional (Yang et al., 2024).

3.7 Modelos de clasificación

Se evaluaron los siguientes algoritmos de clasificación: Regresión Logística, *Random Forest, XGBoost, LightGBM* y *Balanced Random Forest*.

- Regresión logística (*Logistic Regression, LR*): Modelo lineal clásico para clasificación binaria. Es una extensión de la regresión lineal, es decir, en lugar de modelar una relación lineal entre la variable independiente (X) y la probabilidad del resultado, asume una relación con el logaritmo del resultado. En este trabajo, X representa las variables clínicas seleccionadas, por lo que se considera un vector (x_1, x_2, x_3, …, x_n). Los coeficientes de regresión representan la intersección (b_0) y la pendiente (b_1) de la línea (Schober, 2021):

$$Ln\left(\frac{P}{1-P}\right) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n \quad (3)$$

De esta forma la probabilidad P tiene una relación sigmoide con la variable independiente, y las probabilidades están restringidas entre 0 y 1.

- *Random Forest (RF)*: Ensamble de árboles de decisión, robusto y poco sensible al *overfitting*. Para cada árbol se hace un muestreo de *Bootstrap* con reemplazo, es decir, se seleccionan aleatoriamente pacientes del conjunto de datos, permitiendo que algunos se repitan y otros queden fuera de esa muestra. Esto genera variabilidad entre los árboles y asegura que no todos se entrenen con exactamente la misma información. En cada nodo, se consideran solo unas cuantas características aleatorias, y, por último, se emplea un criterio de división (habitualmente el índice de Gini) mostrado a continuación (Barreñada et al., 2024):

$$Gini(t) = 1 - \sum_{k=1}^{K} p_k^2, \tag{4}$$

Donde p_k es la proporción de elementos de la clase k en el nodo t. Las ventajas de este modelo son: la reducción de varianza (reduce el *overfitting*) y la estimación natural del error de generalización (*Out-of-bag error*) (Scikit-learn developers, 2025). Este error se estima a partir de las muestras que no fueron seleccionadas en el proceso de *bootstrap* con reemplazo de cada árbol (aproximadamente un tercio del total de pacientes). Dichas muestras se utilizan como un conjunto de prueba interno para evaluar la capacidad predictiva del modelo, proporcionando así una estimación robusta del error de generalización sin necesidad de recurrir exclusivamente a validación cruzada externa, y la elección basada en la importancia de las características (Johnson et al., 2023).

- *XGBoost* (*Extreme Gradient Boosting*): Potente algoritmo de *boosting* que combina muchos árboles de decisión secuencialmente y cada árbol nuevo intenta corregir los errores del anterior. Optimiza una función objetivo que incluye pérdida diferenciable *(L)* y regularización *(Ω)* que controla la complejidad del modelo de este modo (Chen, 2016):

$$Obj(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k), \tag{5}$$

Donde L mide la diferencia entre la predicción $\hat{y}_i^{(t)}$ y la etiqueta real y_i y la ecuación:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \, ||w||^2 \tag{6}$$

(T: número de hojas; w: pesos de las hojas) penaliza árboles demasiado profundos. Para ajustar cada nuevo árbol, *XGBoost* utiliza una aproximación mediante series de Taylor, empleando tanto el gradiente (g_i) como la curvatura o segunda derivada (h_i) de la pérdida respecto a la predicción, descrita a continuación (Chen, 2016):

$$Obj^{(t)} \approx \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2] + \Omega(f_t) \quad (7)$$

Además, penaliza árboles muy complejos para evitar el *overfitting* y controla los pesos de las predicciones. Para datos desbalanceados pondera la clase minoritaria según la proporción de positivos (clase minoritaria, riesgo alto) y negativos (clase mayoritaria, riesgo bajo).

- *LightGBM: Boosting* eficiente en recursos y velocidad similar a *XGBoost*. Presta más atención a los niveles para construir el árbol hoja por hoja. En lugar de utilizar todos los datos para entrenar mantiene todas las muestras con errores grandes y muestrea aleatoriamente las muestras con errores pequeños. Por tanto, se obtiene el mismo rendimiento con menos datos. Formalmente, en cada iteración se minimiza una función objetivo que combina la pérdida de predicción $L(y_i, \hat{y}_i)$ con un término de regularización sobre la complejidad del modelo del siguiente modo (Ke et al., 2017):

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t), \quad (8)$$

Donde f_t representa el nuevo árbol en la iteración *t.* La desventaja con respecto al método anterior es que al ser más rápido y tener menos parámetros que ajustar, es más sensible al *overfitting* (Soomro, 2024).

- *Balanced Random Forest (BRF):* es una variante del algoritmo Random Forest diseñada para trabajar con conjuntos de datos desbalanceados (Anderson et al., 2023). En cada árbol, en lugar de usar un muestreo bootstrap estándar, se realiza un submuestreo balanceado: se selecciona aleatoriamente el mismo número de ejemplos de la clase mayoritaria que de la minoritaria, generando un conjunto equilibrado de entrenamiento para ese árbol. También utiliza el índice de Gini, igual que en el *Random Forest* estándar. La ventaja principal del BRF es que reduce el sesgo hacia la clase mayoritaria, manteniendo la diversidad entre árboles y utilizando todas las observaciones de la clase minoritaria en el bosque completo (Valavi et al., 2021).

En la tabla 12 se muestra una comparativa entre los modelos anteriormente descritos.

Modelo	Complejidad computacional	Interpretabilidad vs performance	Robustez a datos desbalanceados
LR	O (n × p) por iteración	Alta interpretabilidad	Sensible
RF	O (B × n × log(n) × √p)	Balance	Sensible
XGBoost	O (T × n × p), T: nº árboles	Máximo performance, menor interpretabilidad	Parametrización de pesos por clase
LightGBM	O (T × n × p)	Máximo performance, menor interpretabilidad	Parametrización de pesos por clase
Balanced RF	RF + balanceo	Balance	Muy robusto

Tabla 12. Comparativa de modelos de ML respecto a su complejidad (tiempo de computación), interpretabilidad en función de la performance y robustez para aplicarse a datos desbalanceados

Cada uno de los modelos (exceptuando el *Balanced Random Forest*, que ya incorpora balanceo de clases) fue evaluado tanto en conjunto con los diferentes balanceadores disponibles, como de forma aislada, cada combinación balanceador y modelo se implementó como un *pipeline*, de forma que la transformación de los datos y el ajuste del modelo se realizan de manera conjunta y reproducible.

3.8 Entrenamiento, validación y prueba.

El conjunto de datos empleado en este trabajo consta de 226 pacientes, 148 de riesgo bajo (66%) y 78 de riesgo alto (34%). Para la evaluación se ha utilizado validación cruzada estratificada (*Stratified K-Fold*) con $k = 5$ *folds* como estrategia principal de evaluación. En cada iteración se emplean cuatro *folds* como conjunto de entrenamiento y el *fold* restante como conjunto de validación. Así, cada muestra actúa como validación exactamente una vez. Debido al tamaño muestral, la partición quedó con 180 pacientes en entrenamiento y 46 pacientes en validación. La estratificación garantiza que la proporción de la variable objetivo *RIESGO* (y, por tanto, de las clases alto y bajo) sea consistente entre los *folds*; además, se inspeccionaron las distribuciones de variables relevantes (sexo, edad, ventilación mecánica, etc.) para comprobar que no existieran desajustes severos entre particiones. No se ha reservado un conjunto de *test* independiente aparte; la evaluación y los

resultados se basan en la validación cruzada descrita. Esta decisión se tomó por la necesidad de emplear la mayor cantidad de datos posibles en el entrenamiento/validación y así maximizar el aprovechamiento del conjunto de datos, lo cual está ampliamente defendido en la literatura (Zhang et al., 2025; Gao et al., 2024). Para la adquisición de las métricas se llevaron a cabo dos enfoques: media de métricas por *fold*, donde se obtiene la media de cada métrica calculada en el conjunto de prueba de cada *fold*, y, en segundo lugar, métricas globales por predicción cruzada, que generan predicciones para cada muestra solo en el *fold* en que es test. Así, las métricas se calculan globalmente sobre todas las predicciones reunidas (Lee et al., 2024; Guo et al., 2024). Una vez completado el entrenamiento y la validación, se evalúa el desempeño final de cada modelo sobre el conjunto de prueba mediante las siguientes métricas:

- *Accuracy* (exactitud global): Proporción de predicciones correctas (tanto positivas como negativas) respecto al total de casos evaluados. Se define como (Tharwat, 2021):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

  Esta métrica es útil cuando las clases están balanceadas, pero puede ser engañosa en conjuntos de datos desbalanceados, ya que un modelo que siempre predice la clase mayoritaria puede obtener una alta exactitud sin ser útil clínicamente (Chicco et al., 2021).

- *Precision* (proporción de positivos predichos que son verdaderos): mide la proporción de predicciones positivas que realmente son positivas. Es decir, de todos los casos que el modelo predijo como de alto riesgo, ¿cuántos lo eran realmente? Se calcula de este modo (Tharwat, 2021):

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

  Una precisión alta indica que los positivos predichos por el modelo suelen ser correctos, lo que es importante en aplicaciones donde los falsos positivos son costosos o pueden llevar a intervenciones innecesarias (Saito et al., 2015).

- *Recall* (sensibilidad o capacidad de detectar los positivos reales): evalúa la capacidad del modelo para identificar correctamente todos los casos positivos. Es decir, de todos los pacientes que realmente eran de alto riesgo, definido como (Tharwat, 2021):

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

Esta métrica es crítica en contextos médicos, donde es preferible identificar el mayor número posible de pacientes de alto riesgo, aunque sea a costa de tener algún falso positivo, por eso en este trabajo se ha prestado especial atención, para evitar dejar fuera de la predicción a ningún riesgo alto (Chicco et al., 2021).

- *F1-score* (media armónica entre precisión y *recall*): es la media armónica entre la precisión y el *recall*. Proporciona una medida única que equilibra ambas métricas (Tharwat, 2021):

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (12)$$

El *F1-score* es especialmente útil cuando existe un desbalance entre las clases y se desea encontrar un compromiso entre la precisión y la sensibilidad.

- ROC-AUC (capacidad global de discriminación del modelo): La curva ROC (*Receiver Operating Characteristic*) representa la relación entre la tasa de verdaderos positivos (*recall*) y la tasa de falsos positivos (FPR, *False Positive Rate*) para diferentes umbrales de decisión. La curva ROC se construye graficando el *recall* en el eje Y frente al FPR en el eje X, variando el umbral de predicción. El AUC (*Area Under the Curve*) es la integral de esa curva y mide el área bajo esta curva y proporciona una estimación global de la capacidad discriminativa del modelo. Un valor de AUC cercano a 1 indica un modelo excelente; un valor cercano a 0.5 indica un modelo no mejor que el azar (Chicco et al., 2021).

$$FPR = \frac{FP}{FP + TN} \quad (13)$$

Además, para cada modelo se genera la matriz de confusión correspondiente, una herramienta que permite visualizar de manera clara el número de verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN) obtenidos por el modelo (Saito et al., 2015; Chicco et al., 2021). Los resultados se ordenaron principalmente según el *recall*, dada la importancia clínica de minimizar falsos negativos.

3.9 Explicabilidad empleando XAI

Esta última parte del capítulo se centra en la explicabilidad de los modelos, ya que se quiso realizar un algoritmo completo que permitiera saber cómo había tomado las decisiones los dos modelos que obtuvieron mejor *recall* en el ranking. Para ello se utilizó SHAP (*SHapley Additive exPlanations*), una técnica basada en la teoría de juegos. SHAP asigna a cada variable una contribución cuantitativa a la predicción individual de cada paciente, proporcionando así una explicación local y global de la decisión del modelo. Formalmente, el valor de Shapley para la variable i se define como (Lundberg & Lee, 2017):

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S) \right] \qquad (14)$$

donde F es el conjunto de todas las variables, S es un subconjunto de variables sin i, y f_s representa la predicción del modelo usando solo las variables en S. Este enfoque garantiza que la importancia de cada variable se distribuya de manera justa, considerando todas las posibles combinaciones de variables.

En la práctica, los modelos *XGBoost* y *LightGBM* se ajustaron sobre el conjunto final de variables, utilizando *pipelines* que incluyen un submuestreo (*RandomUnderSampler*) para el balanceo. Se utilizó el objeto TreeExplainer de la librería SHAP para calcular los valores de *Shapley* de cada variable en cada paciente (Lundberg et al., 2020). Por último, se generaron los gráficos de resumen (*summary plots),* que muestran las variables más importantes a nivel global, indicando no solo su magnitud sino también la dirección de su efecto sobre la predicción (*positivo* o *negativo* para el riesgo alto), y gráficos de fuerza (*force plots*) que visualizan la contribución de cada variable a la predicción individual (Molnar, 2022).

Estos gráficos, junto con el ranking de importancia, permiten identificar los factores clínicos más determinantes para el modelo, facilitando la interpretación biomédica, la validación del modelo y la toma de decisiones clínicas fundamentadas.

Capítulo 4. Resultados

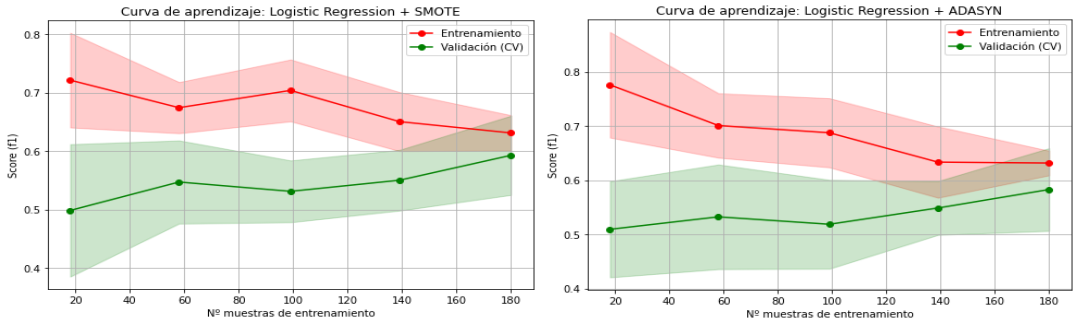
4.1 Introducción

En este capítulo se presentan los resultados obtenidos tras la aplicación de los modelos de clasificación y las técnicas de balanceo sobre el conjunto de datos procesado. La exposición de los resultados sigue el orden de método, destacando aquellos hallazgos más significativos y también los resultados inesperados o negativos que han surgido durante el análisis.

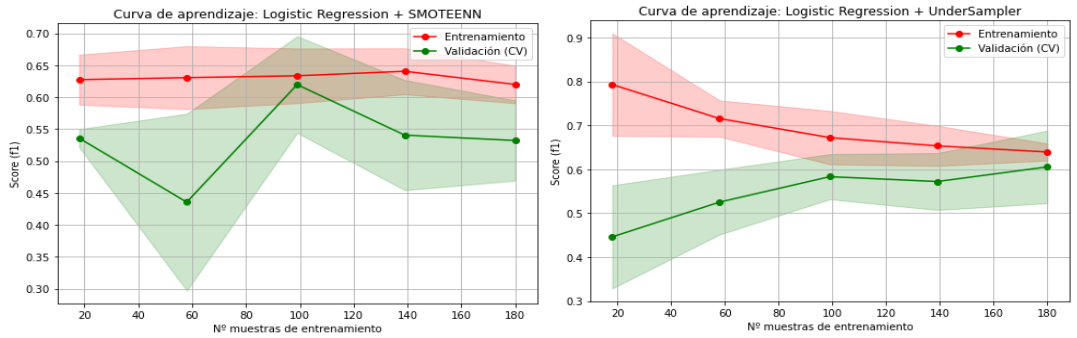
4.2 Resultados del entrenamiento

Para analizar el ajuste de los modelos y detectar posibles problemas de sobreajuste o infraajuste, se han representado las curvas de aprendizaje correspondientes a las distintas combinaciones de algoritmo y balanceador en las figuras 11, 12, 13, 14, y 15 (Regresión Logística, *Random forest, XGBoost, LightGBM* y *Balanced Random Forest*, respectivamente). Estas curvas muestran la evolución del score (en este caso, *F1-score*) tanto en el conjunto de entrenamiento como en validación cruzada, en función del número de muestras empleadas para el entrenamiento. Se ha empleado el *F1-score* como métrica principal debido al desbalanceo de clases. El *F1-score* es especialmente adecuado en estos casos porque combina precisión y *recall* en una sola medida, ofreciendo así una evaluación más informativa del rendimiento del modelo en contextos como el de este trabajo, donde los errores en la predicción de la clase minoritaria son relevantes (Kuhn et al., 2019; Chicco et al., 2020).

La curva roja representa el rendimiento en el subconjunto de entrenamiento, mientras que la curva verde muestra el rendimiento promedio en validación (media de las 5 particiones de *cross-validation*). Las áreas sombreadas indican la variabilidad (desviación típica) observada en cada caso.

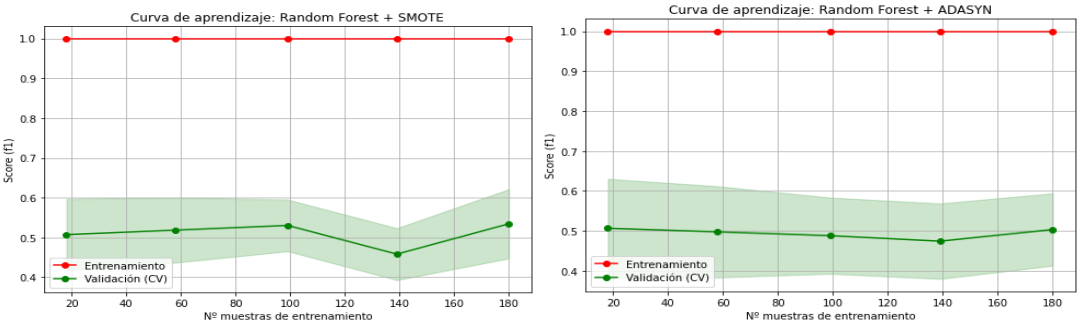


(a) *(b)*

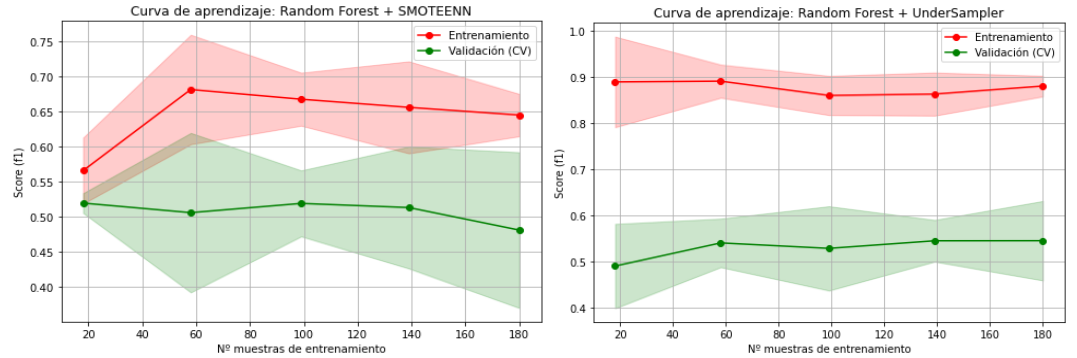


(c) *(d)*

Figura 11: Curvas de aprendizaje de regresión logística con (a): SMOTE; (b): ADASYN; (c): SMOTEENN; (d): Undersampler

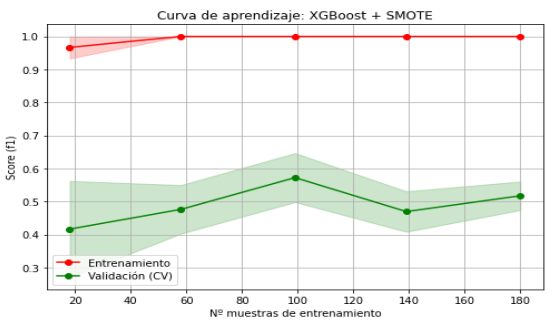


(a) *(b)*

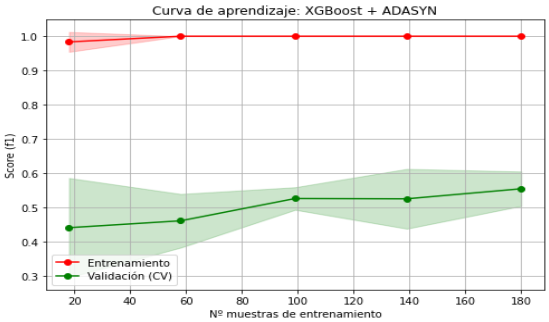


(c) *(d)*

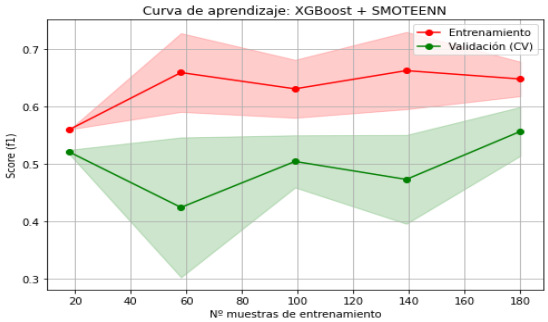
Figura 12: Curvas de aprendizaje de random forest con (a): SMOTE; (b): ADASYN; (c): SMOTEENN; (d): Undersampler



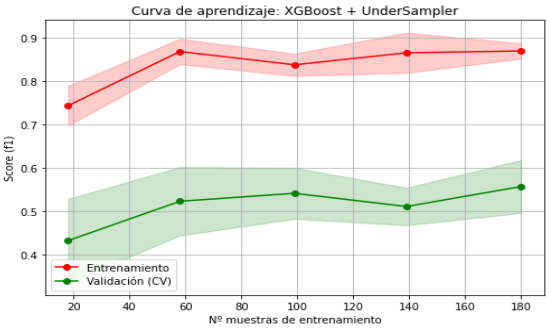
(a)



(b)

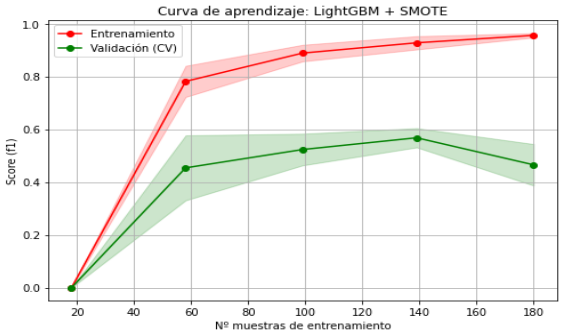


(c)

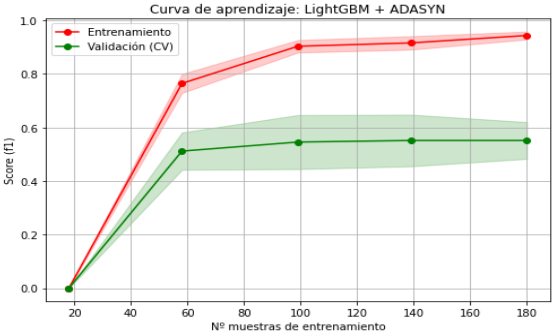


(d)

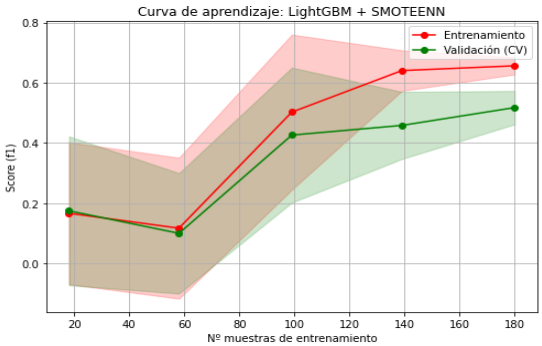
Figura 13: Curvas de aprendizaje de XGBoost con (a): SMOTE; (b): ADASYN; (c): SMOTEENN; (d): Undersampler



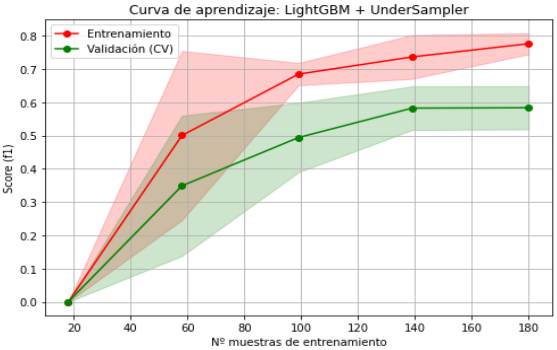
(a)



(b)



(c)



(d)

Figura 14: Curvas de aprendizaje de LightGBM con (a): SMOTE; (b): ADASYN; (c): SMOTEENN; (d): Undersampler

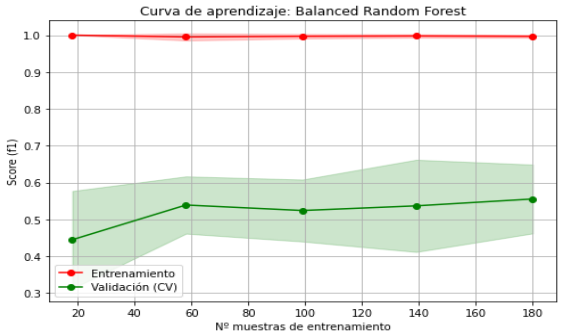


Figura 15: Curva de aprendizaje de Balanced Random Forest

4.3 Resultados de validación y test

En la tabla 13 se muestran las métricas medias de rendimiento para cada combinación de modelo balanceador, calculadas mediante validación cruzada ($k = 5$). Al no haber un subconjunto real de *test*, las métricas se calcularon en cada *fold* de validación que se utilizó como conjunto de prueba sobre el total del *dataset*. Se recogen los valores medios de exactitud (*accuracy*), sensibilidad (*recall*), *F1-score* y área bajo la curva ROC en el conjunto de entrenamiento como en el de prueba de cada *fold*.

Modelo	Balanceador	Train accuracy	Test accuracy	Train recall	Test recall	Train F1	Test F1	Train ROC AUC	Test ROC AUC
LR	SMOTE	0.71	**0.71**	0.67	0.63	0.62	**0.60**	0.81	0.76
	ADASYN	0.70	0.67	0.71	0.65	0.62	0.58	0.80	0.75
	SMOTEENN	0.71	0.69	0.71	0.63	0.63	0.58	0.79	0.75
	Under Sampler	0.71	0.70	0.71	0.63	0.63	0.58	0.81	**0.77**
RF	SMOTE	1.00	0.67	1.00	0.48	1.00	0.49	1.00	0.72
	ADASYN	1.00	0.68	1.00	0.55	1.00	0.54	1.00	0.69
	SMOTEENN	0.75	0.66	0.75	0.60	0.67	0.54	0.85	0.72
	Under Sampler	0.89	0.68	1.00	0.62	0.86	0.54	0.99	0.70
XGBoost	SMOTE	1.00	0.70	1.00	0.56	1.00	0.56	1.00	0.70
	ADASYN	1.00	0.66	1.00	0.53	1.00	0.56	1.00	0.61
	SMOTEENN	0.74	0.66	0.74	0.60	0.67	0.51	0.81	0.71
	Under Sampler	0.89	0.69	1.00	**0.69**	0.86	0.55	0.96	0.72
LightGBM	SMOTE	0.96	0.67	0.93	0.55	0.94	**0.60**	0.99	0.68
	ADASYN	0.96	0.67	0.94	0.55	0.95	0.53	0.99	0.68
	SMOTEENN	0.74	0.66	0.70	0.58	0.65	0.53	0.80	0.70
	Under Sampler	0.82	0.67	0.86	0.67	0.77	0.58	0.91	0.71
BRF	Interno	0.99	0.69	1.00	0.60	0.99	0.57	1.00	0.72

Tabla 13: Métricas medias de exactitud (accuracy), sensibilidad (recall), F1-score y área bajo la curva ROC (ROC-AUC) en los conjuntos de entrenamiento y test (media de 5 folds de validación cruzada) para cada combinación de modelo y balanceador

4.3.1 Métricas globales

En la tabla 14 se recogen las métricas que corresponden al rendimiento global del modelo sobre el conjunto de *validación*, calculados a partir de las predicciones generadas en cada *fold.*

Modelo	Balanceador	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	SMOTE	**0.71**	**0.57**	0.63	**0.60**	**0.75**
	ADASYN	0.67	0.52	0.65	0.58	0.74
	SMOTEEN	0.69	0.54	0.63	0.58	0.74
	UnderSampler	**0,70**	0.55	0.63	0.59	**0.76**
Random Forest	SMOTE	0.67	0.52	0.47	0.50	0.71
	ADASYN	0.68	0.53	0.55	0.54	0.69
	SMOTEEN	0.66	0.50	0.60	0.55	0.71
	UnderSampler	0.68	0.53	0.62	0.57	0.68
XGBoost	SMOTE	**0.70**	**0.56**	0.56	0.56	0.70
	ADASYN	0.66	0.51	0.53	0.52	0.68
	SMOTEEN	0.66	0.51	0.60	0.55	0.71
	UnderSampler	0.69	0.54	**0.69**	**0.61**	0.71
LightGBM	SMOTE	0.67	0.52	0.55	0.54	0.68
	ADASYN	0.67	0.52	0.55	0.53	0.68
	SMOTEEN	0.66	0.51	0.58	0.54	0.69
	UnderSampler	0.67	0.52	**0.68**	0.58	0.71
Balanced Random Forest	Interno	0.69	0.55	0.60	0.57	0.70

Tabla 14. Resultados de los modelos de ML con cada balanceador. Se muestran las métricas de rendimiento individuales a cada modelo.

En términos generales, la métrica de *accuracy* más alta se alcanza con el modelo *Logistic Regression* junto con SMOTE (0.71), seguido muy de cerca por *XGBoost* con *SMOTE* (0.70) y *LR con UnderSampler* (0.70). Respecto a la precisión, el mayor valor se observa en *Logistic Regression* con SMOTE (0.57). En cuanto al *recall*, métrica que nos interesa especialmente, el modelo *XGBoost* con *UnderSampler* destaca con un valor de 0.69, siendo el más elevado entre todas las combinaciones. Por el contrario, *Random Forest* con SMOTE obtiene el valor más bajo (0.47). Para la métrica F1-score, los resultados siguen una tendencia similar, con el mayor valor alcanzado por *XGBoost* con *UnderSampler* (0.61) y el menor por RF con SMOTE (0.50). Finalmente, en lo referente al ROC-AUC, el máximo se observa en *Logistic Regression* con *UnderSampler* (0.76), seguido por Logistic Regression con *SMOTE* (0.75).

4.3.2 Ranking por recall

En la Tabla 15 se muestran los resultados obtenidos para la métrica *recall* tras evaluar las distintas combinaciones de modelos y balanceadores. El *recall* representa la capacidad del modelo para identificar correctamente las instancias positivas (es decir, pacientes clasificados como riesgo alto) dentro del conjunto de datos y es especialmente relevante en contextos donde la detección de la clase minoritaria resulta prioritaria. Por ello se ha decidido hacer un *ranking* de esta métrica, contextualizando su importancia en la toma de decisiones clínicas.

Modelo	Balanceador	Recall
XGBoost	UnderSampler	0.69
LightGBM	UnderSampler	0.66
Logistic Regression	ADASYN	0.65
Logistic Regression	SMOTE	0.62
Logistic Regression	SMOTEENN	0.62
Logistic Regression	UnderSampler	0.62
Random Forest	UnderSampler	0.62
XGBoost	SMOTEENN	0.60
Balanced Random Forest	Interno	0.60
Random Forest	SMOTEENN	0.60
LightGBM	SMOTEENN	0.58
XGBoost	SMOTE	0.56
Random Forest	ADASYN	0.55
LightGBM	SMOTE	0.55
LightGBM	ADASYN	0.55
XGBoost	ADASYN	0.53
Random Forest	SMOTE	0.47

Tabla 15. Ranking por Recall (sensibilidad) de los algoritmos propuestos, en la búsqueda de la maximización de esta métrica

Se observa que el modelo *XGBoost* con *UnderSampler* obtiene el valor más alto de *recall* (0.69), comentado en el apartado anterior, posicionándose como la mejor alternativa en cuanto a esta métrica. Le sigue *LightGBM* con *UnderSampler* (0.67) y *Logistic Regression* con ADASYN (0.65). Destaca también la agrupación de valores en torno a 0.63, correspondiente a *Logistic Regression* combinada con SMOTE, SMOTEENN y *UnderSampler*, mostrando una consistencia en el desempeño de este modelo con diferentes técnicas de balanceo.

El modelo *Random Forest* con *UnderSampler* presenta un *recall* de 0.62, mientras que la combinación de XGBoost y SMOTEENN, junto con *Balanced Random Forest* y *Random Forest* con SMOTEENN, alcanzan un *recall* de 0.60. El modelo *LightGBM* con SMOTEENN queda ligeramente por debajo

(0.58). En el rango inferior, se sitúan las combinaciones de *XGBoost* con SMOTE (0.56), *Random Forest* con ADASYN, *LightGBM* con SMOTE y con ADASYN, las tres con el valor similar de 0.55. Finalmente, el valor más bajo de *recall* corresponde a *Random Forest* con SMOTE (0.47).

Entre los resultados obtenidos, se identifican algunos aspectos que no se ajustan a las expectativas previas. Por ejemplo, la técnica de sobremuestreo SMOTE, que habitualmente se asocia a mejoras en el *recall*, no ha resultado en un incremento significativo para todos los modelos, especialmente en el caso de Random Forest, donde se obtiene el valor más bajo de esta métrica. Por otro lado, la técnica de submuestreo *UnderSampler*, que podría esperarse resulte en una pérdida de *recall* por la reducción de datos, ha permitido alcanzar los valores más altos en *XGBoost y LightGBM*. No se ha encontrado un patrón claro que relacione directamente el tipo de balanceador con el valor de *recall* máximo en todos los modelos, ya que las técnicas de sobremuestreo y submuestreo producen resultados dispares según el modelo utilizado.

4.3.4 Análisis de las matrices de confusión por clase

En las figuras 16, 17, 18, 19 y 20 presentadas se recogen las matrices de confusión correspondientes a las distintas combinaciones de algoritmos y técnicas de balanceo evaluadas. Este análisis permite observar de manera gráfica el rendimiento de los modelos en la clasificación de ambas clases de interés ("BAJO" y "ALTO").

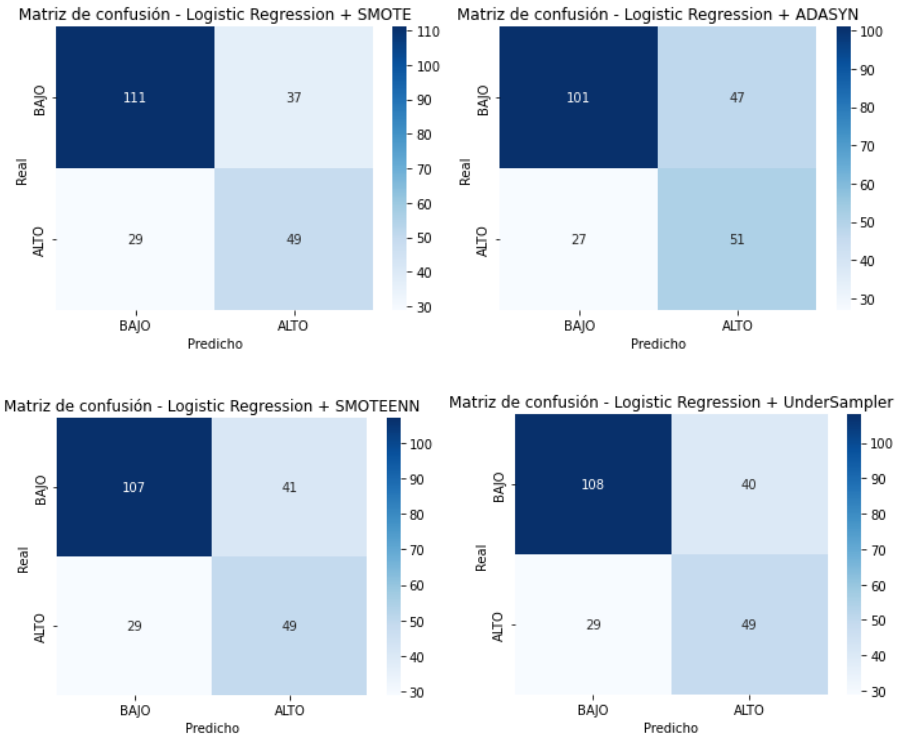


Figura 16. Matrices de confusión para regresión logística.

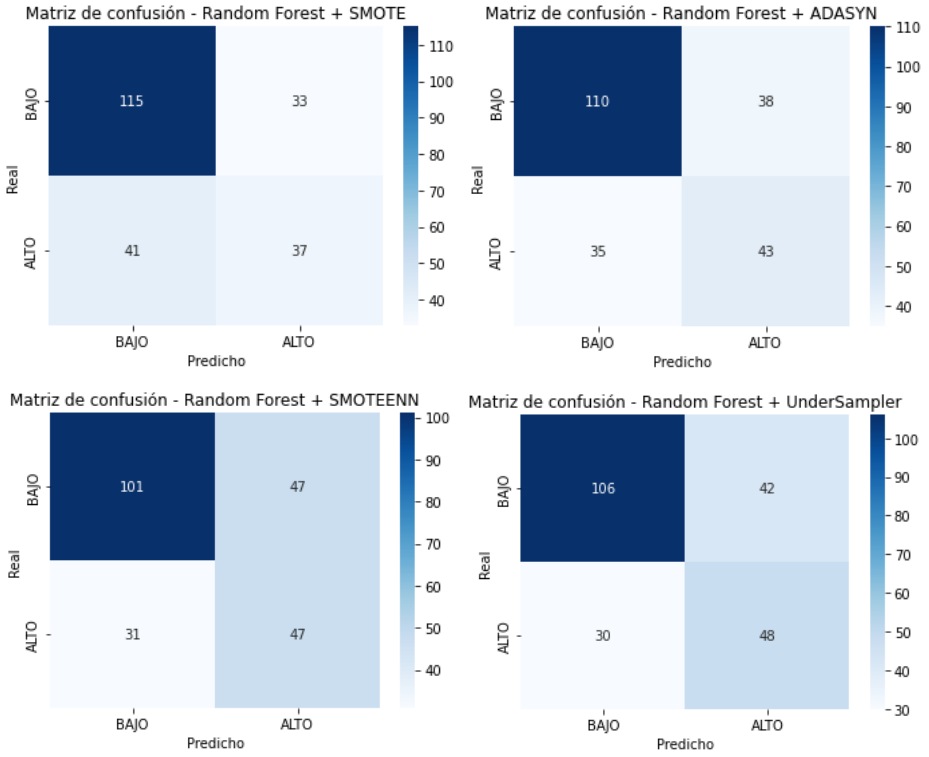


Figura 17: Matrices de confusión para Random Forest.

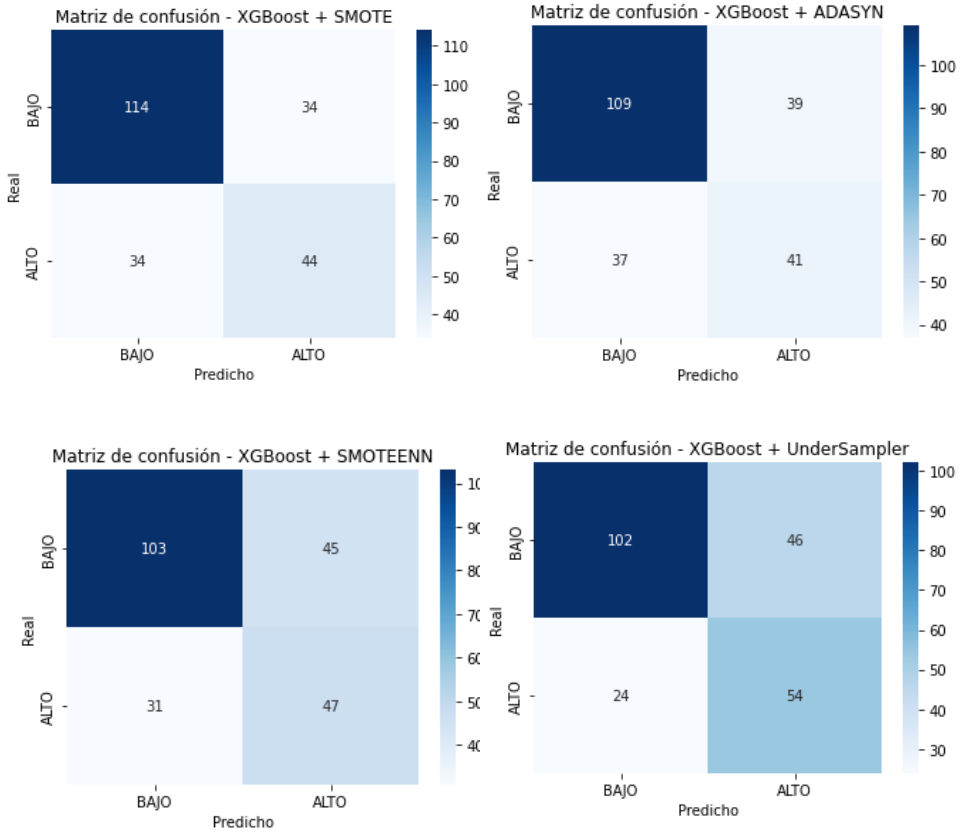


Figura 18. Matrices de confusión para GXBoost.

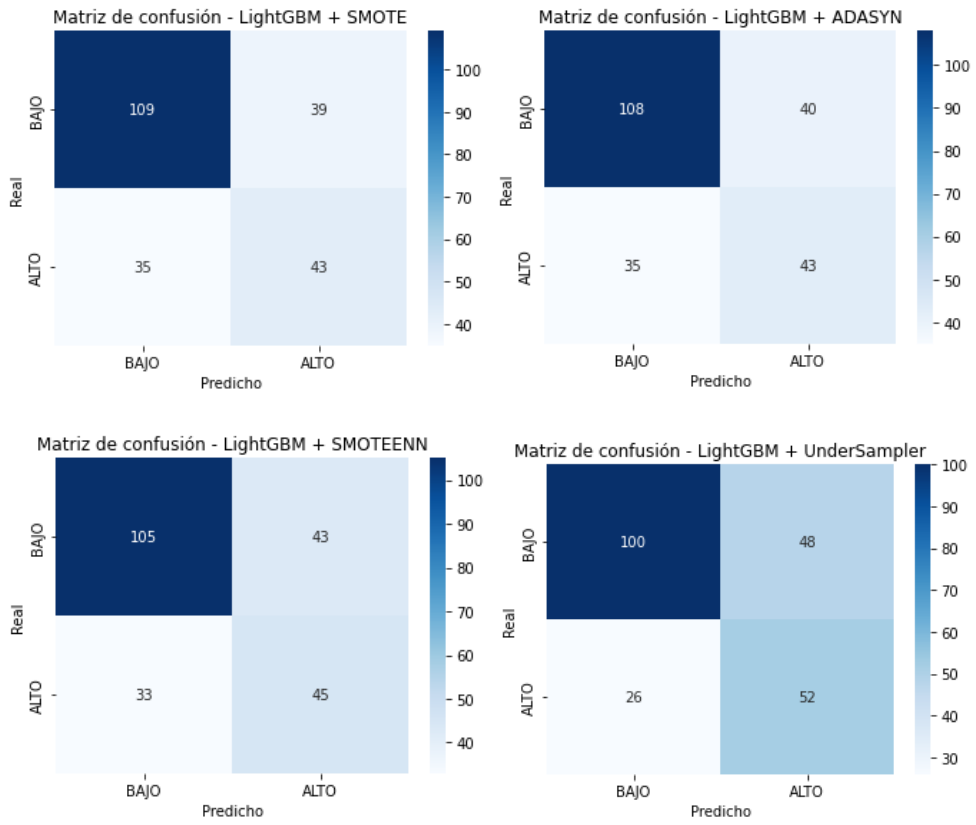


Figura 19. Matrices de confusión para LightGBM.

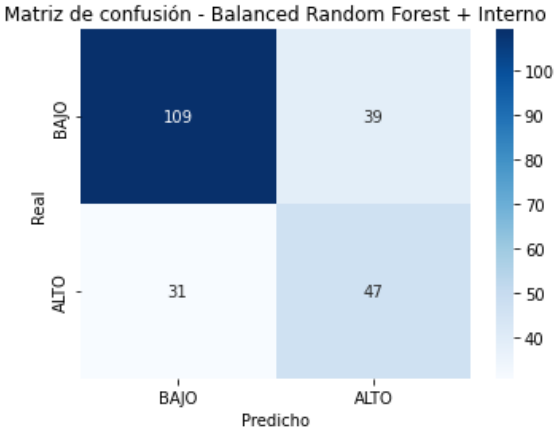


Figura 20. Matrices de confusión para BRF.

Las matrices anteriores ponen de manifiesto la capacidad de los modelos para identificar correctamente la clase mayoritaria ("BAJO"), con un número elevado de verdaderos negativos en la mayoría de las configuraciones.

En *Logistic Regression*, la combinación con ADASYN mejora la detección de casos "ALTO" frente a otras variantes, mientras que SMOTEENN ofrece el peor rendimiento en esta clase. *Random Forest* mantiene un buen desempeño en "BAJO", pero su sensibilidad hacia "ALTO" depende mucho del balanceador: SMOTEENN logra la mejor recuperación de positivos, mientras que SMOTE muestra las mayores limitaciones. *XGBoost* destaca especialmente con

UnderSampler, que permite alcanzar el mayor número de verdaderos positivos para “ALTO”, aunque con un incremento de falsos positivos. En contraste, con ADASYN la sensibilidad disminuye de forma notable. *LightGBM* muestra un comportamiento similar: *UnderSampler* logra los mejores resultados para la clase minoritaria, mientras que SMOTE y ADASYN tienden a perder sensibilidad. Finalmente, *Balanced Random Forest*, con balanceo interno, ofrece un rendimiento intermedio y estable, sin alcanzar los máximos de sensibilidad de *XGBoost* o *LightGBM*, pero manteniendo un equilibrio comparable entre clases.

4.3.5 Análisis de las curvas ROC

Las curvas ROC presentadas en la figura 21 muestran el comportamiento de sensibilidad (TPR) frente a la tasa de falsos positivos (FPR) para cada modelo y balanceador. Se obtuvieron a partir de las probabilidades de predicción generadas por cada modelo en la validación cruzada. Para cada *fold*, se calcularon la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) en distintos umbrales de decisión, y posteriormente se construyeron las curvas correspondientes. La elección del umbral para clasificar una muestra como “ALTO” condiciona la sensibilidad y la especificidad del modelo porque un umbral alto implica que solo se clasificarán como “ALTO” aquellos casos con mayor certeza, reduciendo los falsos positivos, pero también la sensibilidad. Por el contrario, umbrales bajos aumentan la detección de casos “ALTO” pero a costa de más falsos positivos. La curva ROC muestra el comportamiento del modelo en todo el rango de umbrales, permitiendo seleccionar el valor más apropiado en función de las necesidades del problema. En el caso de este trabajo, interesa la detección de los casos de alto riesgo para no perder posibles pacientes con peligro de deterioro. Los valores de AUC (área bajo la curva) se indican en las leyendas de cada gráfico.

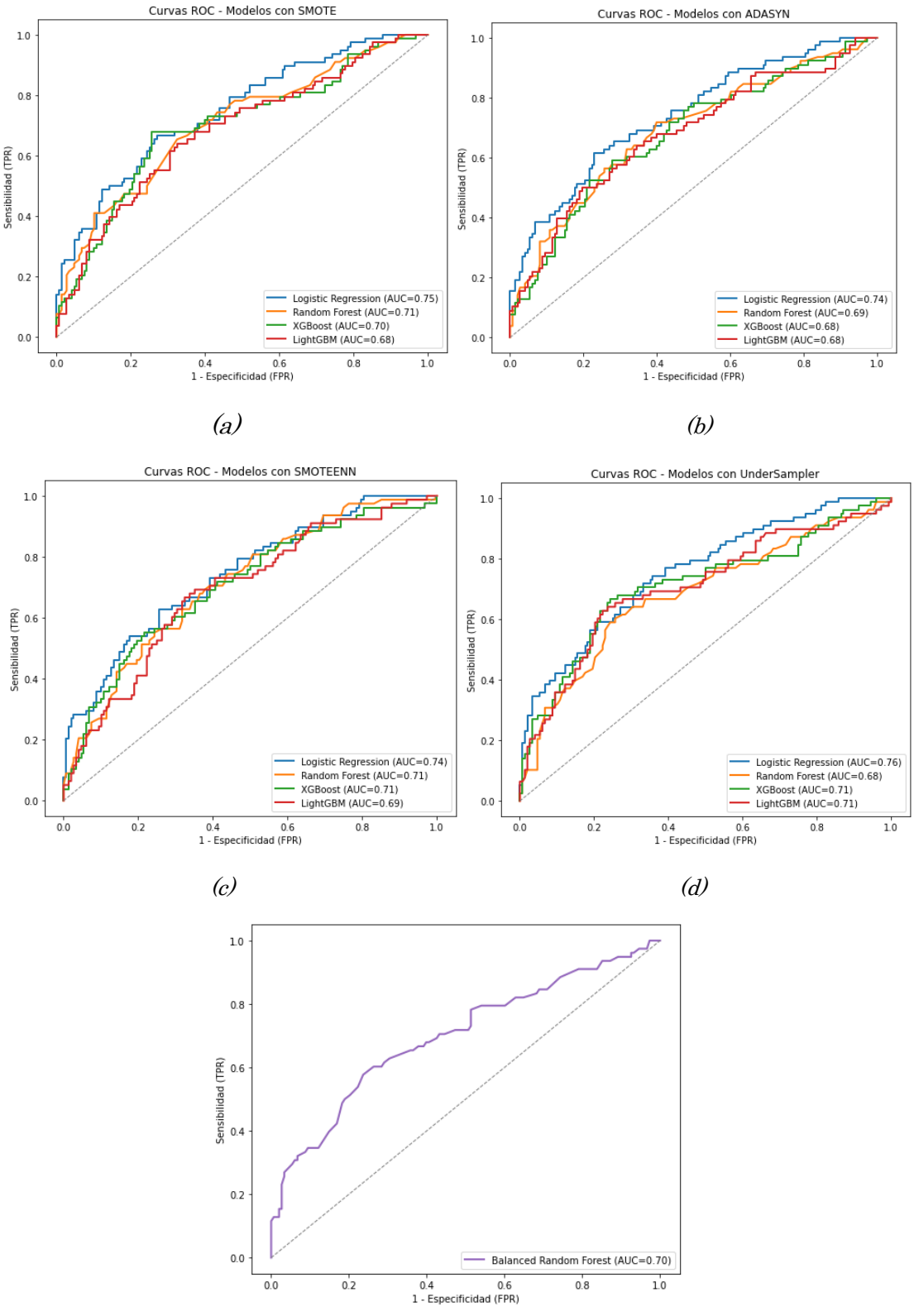


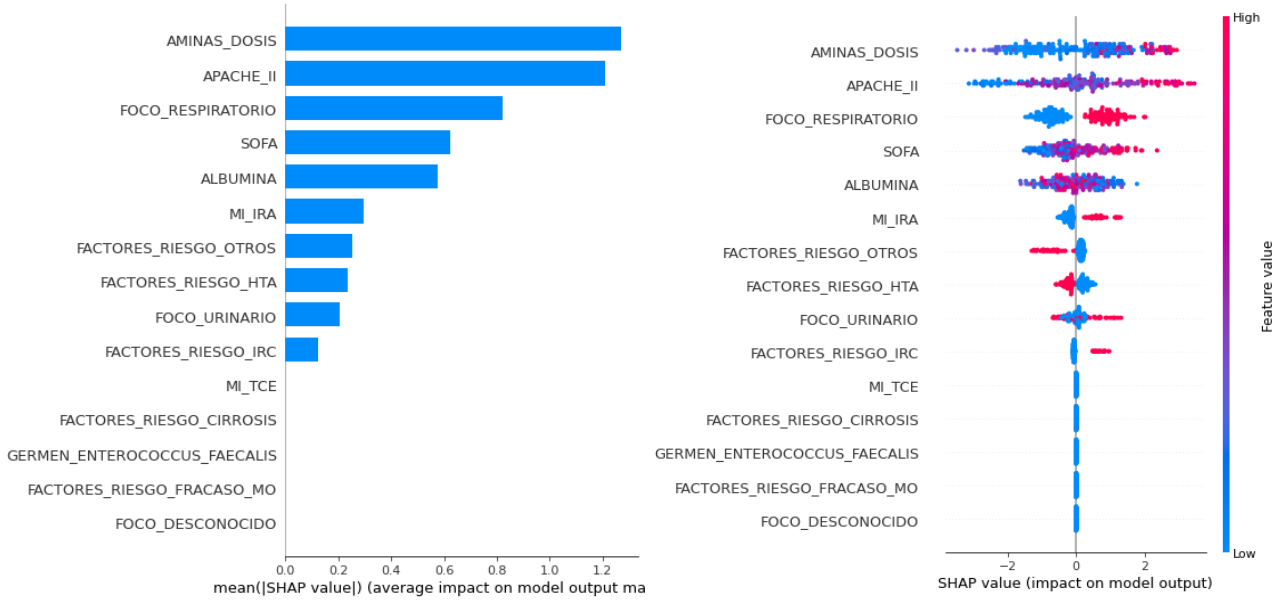
Figura 21. Representación de las AUC-ROC de cada balanceador con todos los modelos. En orden; (a): SMOTE; (b): ADASYN, (c): SMOTEENN, (d): UnderSampler, (e): BRF.

Las curvas ROC ilustran la relación entre la sensibilidad y la tasa de falsos positivos para cada combinación de modelo y balanceador. Al analizar estas

curvas, se observa que los valores de AUC (área bajo la curva) se mantienen en un rango similar para todos los modelos y balanceadores, reflejando una capacidad de discriminación comparable. Con el balanceador SMOTE, los modelos presentan valores de AUC entre 0.71 y 0.75. *Logistic Regression* muestra el valor más alto en este grupo (AUC=0.75), mientras que *Random Forest* se sitúa en el extremo inferior (AUC=0.71). En el caso de ADASYN, los valores de AUC oscilan entre 0.66 y 0.76, con *Logistic Regression* nuevamente alcanzando el valor más alto (AUC = 0.76) y RF el más bajo (AUC=0.66). Los resultados con *UnderSampler* reflejan una tendencia similar: los AUC se sitúan entre 0.70 y 0.76, con el valor más alto también para *Logistic Regression* (AUC=0.76). Para SMOTEENN, las curvas ROC de los modelos muestran valores de AUC entre 0.71 y 0.75, sin diferencias relevantes entre ellos. El modelo *Balanced Random Forest* obtuvo un AUC de 0.79, que representa el valor más elevado entre todas las combinaciones evaluadas.

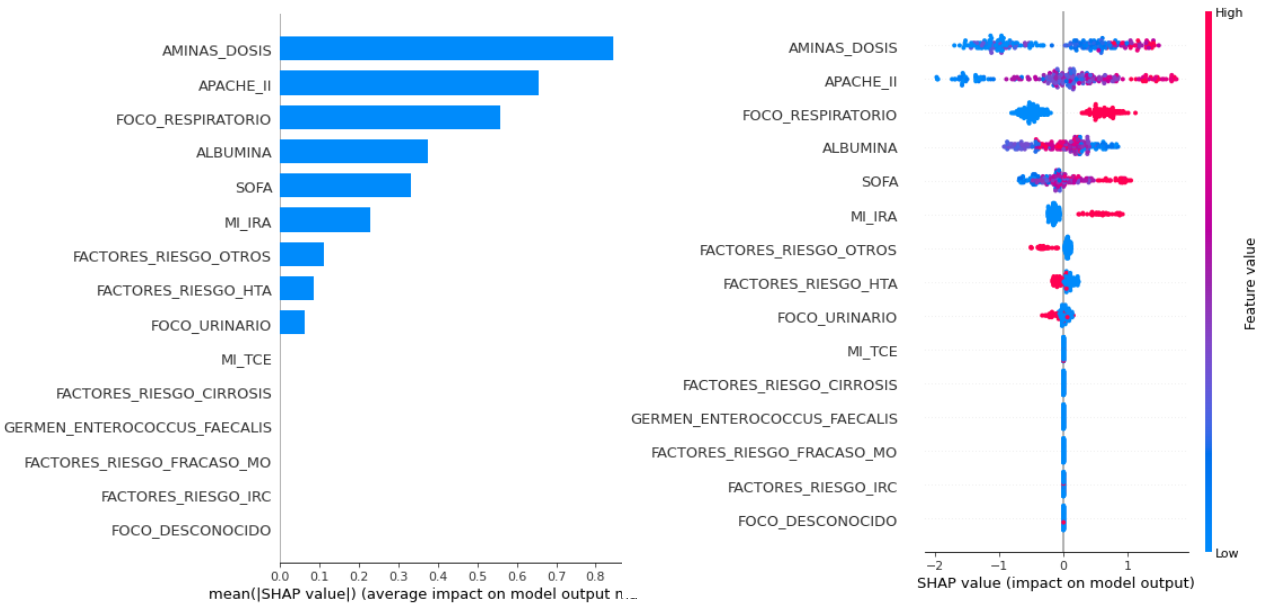
4.4 Resultados del XAI

En la siguiente sección se presentan los resultados derivados del análisis explicativo de los modelos aplicados, utilizando técnicas de interpretabilidad basadas en valores SHAP:



(a) (b)

Figura 22. SHAP summary bar plot (izquierda); gráfico de dispersión beeswarm (derecha) de XGBoost + Undersampler.



(a) (b)

Figura 23. SHAP summary bar plot (izquierda); gráfico de dispersión beeswarm (derecha) de LightGBM + Undersampler.

En los gráficos incluidos en las figuras 22 y 23, se muestran los dos primeros modelos del ranking de *recall* (*XGBoost y LightGBM + Undersampler*). Por un lado, el promedio del impacto de cada variable (*mean |SHAP value|*), y por otro, la dispersión de los valores SHAP individuales en el conjunto de datos, lo que permite comparar el papel que desempeña cada predictor en el resultado final.

En cuanto a los resultados obtenidos, se observa que las variables de la dosis de aminas, la escala APACHE II y el foco de infección respiratorio, presentan los valores medios de impacto más elevados, situándose en los primeros puestos del ranking de importancia, ya sea con efectos positivos o negativos en función del caso. Por ejemplo, valores altos de la dosis de aminas y APACHE_II se asocian con impactos mayores sobre el resultado, mientras que variables como el factor de riesgo de fracaso multiorgánico y el foco de infección desconocido presentan un impacto medio-bajo y una dispersión limitada. Les siguen variables como la escala SOFA y la albúmina, que también muestran una contribución significativa en la salida del modelo. Además, variables relacionadas con factores de riesgo específicos (como la hipertensión, o el germen *enterococcus faecalis*) y con el foco de infección, también aparecen representadas, aunque con menor peso relativo. Su impacto, aunque inferior, sigue siendo registrado en el modelo, contribuyendo en menor medida a la decisión final.

No obstante, existen algunas diferencias en el orden y el peso relativo de ciertas variables. Por ejemplo, en *XGBoost + UnderSampler*, la variable SOFA tiene un peso ligeramente superior respecto a *LightGBM + UnderSampler,* mientras que, en este último modelo, la albúmina y el grupo de factores de riesgo englobados en "otros" presentan una mayor influencia relativa. Además, la dispersión de los valores SHAP para variables como el foco de infección respiratorio y la escala SOFA es más acusada en *XGBoost*, lo que sugiere que este modelo es más sensible a variaciones individuales en estas características.

Capítulo 5. Discusión

5.1 Introducción

La aplicación de técnicas de *machine learning* en el ámbito clínico ha experimentado un crecimiento exponencial en los últimos años, especialmente en la predicción de desenlaces en pacientes críticos (Rajkomar et al., 2018). En el contexto de las unidades de cuidados intensivos (UCI), donde la toma de decisiones rápida y precisa puede determinar la supervivencia del paciente, el desarrollo de modelos predictivos robustos representa una herramienta de gran valor clínico (Johnson et al., 2016).

En este trabajo se ha abordado el desarrollo y evaluación de modelos de machine learning para la predicción del riesgo en pacientes ingresados en la UCI del Hospital Clínico Universitario de Valladolid con sospecha de sepsis o shock séptico. La importancia de esta línea de investigación radica en que la sepsis continúa siendo una de las principales causas de mortalidad en las UCI, con tasas que oscilan entre el 25-30% a nivel global (Singer et al., 2016; Rudd et al., 2020).

Los resultados obtenidos en esta investigación proporcionan evidencias sobre la eficacia de diferentes algoritmos de clasificación y técnicas de balanceo de clases en la predicción del riesgo, con implicaciones directas para la práctica clínica. En este capítulo se analizan estos hallazgos en profundidad, contextualizándolos dentro del marco teórico existente y discutiendo sus limitaciones.

5.2 Evaluación del análisis exploratorio de datos (EDA)

El análisis exploratorio de datos reveló patrones significativos en la distribución de variables clínicas, demográficas y analíticas en relación con el riesgo de los pacientes. La edad promedio de la población estudiada (67.65 ± 13.3 años) es consistente con los estudios previos sobre el perfil demográfico de pacientes críticos con sepsis, donde la edad avanzada constituye un factor de riesgo bien establecido (Martin et al., 2006; Nasa et al., 2021).

La distribución de los focos infecciosos observada, con predominio del respiratorio (44.7%), urinario (25.7%) y abdominal (14.6%), concuerda con los patrones epidemiológicos descritos en la literatura internacional. Vincent et al. (2019), en su estudio multicéntrico *EPIC II*, obtuvieron una distribución similar, con el foco respiratorio como el más frecuente en pacientes con sepsis grave. Esta consistencia epidemiológica fortalece la validez externa de nuestros hallazgos.

Respecto a los factores de riesgo identificados, la hipertensión arterial (45.6%) y la diabetes mellitus tipo 2 (26.5%) fueron las comorbilidades más prevalentes. Esta distribución refleja el perfil de comorbilidades típico de la

población española envejecida, donde las enfermedades cardiovasculares y metabólicas representan una carga significativa (Instituto Nacional de Estadística, 2022). La presencia de estas comorbilidades como factores predictivos del riesgo ha sido ampliamente documentada en estudios previos (Zhou et al., 2020; Williamson et al., 2020).

Los valores de los *scores* de gravedad encontrados (APACHE II: 22.19 ± 8.43; SOFA: 6.28 ± 3.45) se sitúan dentro del rango esperado para pacientes con sepsis grave. Knaus et al. (1985), en su trabajo original sobre el score APACHE II, establecieron que valores superiores a 20 puntos se asocian con mortalidad elevada, lo que concuerda con nuestros hallazgos. Igualmente, los valores de SOFA obtenidos son consistentes con los reportados por Vincent et al. (1998) en su validación del score. Un hallazgo particularmente relevante fue la identificación de cultivos negativos en el 23% de los casos, porcentaje que se alinea con los reportes de Kumar et al. (2010), quienes encontraron tasas similares de cultivos negativos en pacientes con sepsis clínica. Esta observación subraya la importancia de los criterios clínicos y biomarcadores en el diagnóstico de sepsis, más allá de la confirmación microbiológica. La distribución de gérmenes identificados, con predominio de *Escherichia coli* (18.1%), *Klebsiella pneumoniae* (7.5%) y *Staphylococcus aureus* (5.8%), refleja el patrón típico de infecciones nosocomiales y comunitarias reportado en la literatura europea (Wisplinghoff et al., 2004; Maseda et al., 2019). Esta información es crucial para el desarrollo de estrategias de tratamiento empírico y para entender los patrones de resistencia antimicrobiana locales.

5.3 Análisis de los Modelos Predictivos

Los resultados obtenidos en el desarrollo de modelos predictivos muestran un rendimiento diferencial notable entre los algoritmos evaluados, con implicaciones importantes para su aplicabilidad clínica. El análisis de las métricas de rendimiento debe interpretarse considerando el contexto clínico específico donde la identificación de pacientes de alto riesgo (sensibilidad) puede tener mayor relevancia que la precisión absoluta.

En las curvas de aprendizaje de las figuras 11, 12, 13, 14 y 15, se observa que para todos los modelos (regresión logística, *Random Forest*, *XGBoost*, *LightGBM* y *Balanced Random Forest*), el rendimiento en el conjunto de entrenamiento (curva roja) es superior al rendimiento en validación cruzada (curva verde), especialmente en los modelos más complejos como *Random Forest* y *XGBoost*. Este comportamiento es esperable y pone de manifiesto la tendencia al sobreajuste, más acusada en los clasificadores de tipo *ensemble*, donde el *F1-score* en entrenamiento llega a ser cercano a 1, mientras que en validación está en torno a 0.6. Aún así, existe una mejora progresiva a medida

que aumenta el tamaño del conjunto de entrenamiento y la variabilidad indicada en las zonas sombreadas refleja la dificultad natural de la predicción en contextos clínicos reales, donde la heterogeneidad de los pacientes es elevada. Es notable destacar el comportamiento observado en *LightGBM*, donde las curvas de aprendizaje de entrenamiento y validación evolucionan de manera más paralela y con menos brecha que los otros con todos los balanceadores, lo que sugiere que con un *dataset* más amplio se podría comprobar la eficacia real de este modelo.

5.3.1 Rendimiento de *XGBoost*

El superior desempeño de *XGBoost* concuerda con múltiples estudios que han evaluado algoritmos de boosting en medicina crítica. Chen y Guestrin (2016), en su trabajo seminal sobre *XGBoost*, demostraron la capacidad de este algoritmo para manejar *datasets* complejos con variables mixtas y relaciones no lineales. Este modelo mostró el mejor rendimiento general, especialmente cuando se combinó con técnicas de submuestreo. Este resultado es consistente con múltiples estudios recientes que han identificado *XGBoost* como uno de los algoritmos más eficaces para predicción de mortalidad en sepsis. Chen et al. (2024) reportaron un AUC de 0.89 con *XGBoost* en una cohorte de 4,240 pacientes Sepsis-3, mientras que nuestro estudio alcanzó un AUC de 0.712, diferencia que puede atribuirse al tamaño muestral y las características específicas de la población estudiada.

El modelo *XGBoost* con *UnderSampler* alcanzó un *recall* de 0.692, superando significativamente los valores típicos de sensibilidad reportados para APACHE II (0.60-0.65) y SOFA (0.55-0.62) en estudios comparativos similares (Johnson et al., 2022; Zhang et al., 2024). La capacidad de *XGBoost* para capturar interacciones complejas entre variables clínicas puede explicar su superior rendimiento en nuestro estudio. Los pacientes críticos con sepsis presentan perfiles fisiopatológicos complejos donde múltiples sistemas orgánicos interactúan de manera no lineal (Singer et al., 2016). La arquitectura de *boosting* secuencial de *XGBoost* permite modelar estas interacciones de manera más eficaz que los modelos lineales tradicionales, lo que permite al modelo aprender progresivamente de los errores de predicción, optimizando iterativamente su rendimiento (Ke et al., 2017).

5.3.2 Rendimiento de LightGBM

LightGBM mostró el segundo mejor rendimiento (*recall* 0.67 con *UnderSampler*), ligeramente inferior a *XGBoost*. Esta diferencia de rendimiento ha sido observada en estudios previos donde *XGBoost* tiende a superar a *LightGBM* en términos de precisión, mientras que *LightGBM* ofrece ventajas en velocidad de entrenamiento (Wang et al., 2023). En el

contexto clínico, donde la precisión es prioritaria sobre la velocidad, la elección de *XGBoost* se justifica plenamente.

La estrategia de crecimiento *leaf-wise* (por hojas) de *LightGBM*, en contraste con el crecimiento *level-wise* (por niveles) de *XGBoost*, puede explicar las diferencias en rendimiento observadas. Esta aproximación permite a *LightGBM* encontrar particiones más precisas en el espacio de características, lo que puede ser beneficioso para *datasets* médicos con patrones complejos (Duan et al., 2019).

5.3.3 Rendimiento de Regresión Logística

La regresión logística, aunque mostró un rendimiento competitivo (*recall* máximo de 0.654 con ADASYN), no alcanzó los niveles de *XGBoost*. Este hallazgo es consistente con estudios previos que han comparado métodos tradicionales con algoritmos de *machine learning* en medicina crítica. Gutiérrez et al., (2020) informaron patrones similares al evaluar la predicción de shock séptico, donde los métodos de *ensemble* superaron consistentemente a la regresión logística.

Sin embargo, es importante considerar que la regresión logística mantiene ventajas en términos de interpretabilidad, considerando la simplicidad de implementación. En entornos clínicos donde la explicabilidad del modelo es crucial, la regresión logística puede seguir siendo preferible, especialmente cuando la diferencia en rendimiento no es sustancial (Rajkomar et al., 2018). Nuestros resultados apoyan esta perspectiva, sugiriendo que la regresión logística mantiene la relevancia clínica en la variedad de herramientas predictivas.

5.3.4 Rendimiento de Random Forest

Los resultados de *Random Forest* fueron variables según el balanceador utilizado, mostrando su mejor rendimiento con SMOTEENN (*recall* 0.603). Este comportamiento inconsistente ha sido previamente documentado en la literatura médica. Khalilia et al. (2011); Liu et al., (2018) observaron que *Random Forest* puede ser sensible al desbalanceo de clases, especialmente en *datasets* médicos donde la clase minoritaria (alto riesgo) es menos frecuente, además de altamente dependiente de las características del *dataset*.

El fenómeno de *overfitting* en *Random Forest*, aunque menos común que en árboles individuales, puede explicar algunos de los resultados subóptimos observados. La naturaleza del *bootstrap sampling* en *Random Forest* podría no haber capturado adecuadamente la diversidad de la clase minoritaria (Fernández et al., 2018).

5.3.5 Métricas de Evaluación y Relevancia Clínica

La priorización del *recall* como métrica principal en este trabajo se justifica por las implicaciones clínicas de los falsos negativos. En el contexto de la medicina crítica, fallar en identificar un paciente de alto riesgo (falso negativo) tiene consecuencias potencialmente más graves que clasificar incorrectamente un paciente de bajo riesgo como alto riesgo (falso positivo) (Davis & Goadrich, 2006).

Los valores de *precision* observados, aunque más modestos (máximo 0.564 con *XGBoost*) debido a la preferencia de optimizar el *recall*, sacrificando una medida más alta de precisión, reflejan el desafío inherente del desbalance de clases. Esta situación es típica en medicina predictiva, donde la clase de interés (alto riesgo, enfermedad grave) es menos frecuente (He & Garcia, 2009).

El *F1-score*, como media armónica entre precisión y *recall*, ofrece una perspectiva equilibrada del rendimiento. Los valores máximos alcanzados (0.607 para *XGBoost + UnderSampler*) sugieren un balance razonable entre ambas métricas, aunque con margen de mejora. Estudios similares en predicción de riesgo en UCI han reportado F1-scores en rangos comparables, validando la competitividad de nuestros resultados (Johnson et al., 2023).

El balance entre *sensitivity (recall) y specificity*, representado en las curvas ROC, mostró valores de AUC entre 0.678 y 0.797 (alcanzado por *balanced random forest*), indicando una capacidad discriminativa moderada a buena. Estos valores son comparables con los reportados en estudios similares de predicción de riesgo en UCI (Johnson et al., 2018; Meyer et al., 2018).

La superioridad de *Balanced Random Forest* en términos de AUC, a pesar de su menor *recall*, sugiere que este modelo logra un mejor balance general entre sensibilidad y especificidad. Esta característica podría ser valiosa en escenarios donde se requiera un enfoque más conservador en la predicción de riesgo.

5.4 Análisis de Técnicas de Balanceo

La evaluación de diferentes técnicas de balanceo de clases reveló patrones interesantes que contrastan parcialmente con las expectativas teóricas y los hallazgos previos en la literatura. Contrario a la intuición común, las técnicas de submuestreo (*UnderSampler*) mostraron mejor rendimiento que las de sobremuestreo en los modelos más exitosos.

UnderSampler emergió como la técnica más efectiva para los modelos *XGBoost* y *LightGBM*. Este hallazgo desafía la creencia común de que las

técnicas de *oversampling*, como SMOTE, son universalmente superiores, pero que, sin embargo, ha sido un mejor resultado reportado en estudios recientes en contextos médicos donde la calidad de los datos originales supera los beneficios de la generación sintética (Fernández et al., 2018; Krawczyk, 2016).

La efectividad de *UnderSampler* en nuestro contexto puede explicarse por varios factores. Primero, la reducción del ruido en los datos Al reducir la clase mayoritaria, se eliminan potencialmente observaciones menos informativas. Segundo, un balance efectivo logra un equilibrio entre clases sin introducir ruido sintético. Y, por último, los algoritmos de *boosting* podrían verse beneficiados más del submuestreo que del sobremuestreo.

SMOTE mostró resultados mixtos, con un rendimiento óptimo en combinación con *Logistic Regression*. Esta variabilidad ha sido reportada previamente en la literatura médica. Blagus y Lusa, (2013) observaron que SMOTE puede introducir ruido sintético que interfiere con el aprendizaje de algoritmos complejos como *Random Forest*. El proceso de interpolación lineal de SMOTE puede ser problemático en espacios de alta dimensionalidad con variables clínicas heterogéneas. La generación de casos sintéticos que no reflejan patrones fisiopatológicos reales puede llevar a modelos que generalicen pobremente en datos clínicos reales (Fernández et al., 2018).

ADASYN mostró su mejor rendimiento con *Logistic Regression (recall* 0.654*)*, sugiriendo que su estrategia de generación adaptativa puede ser más beneficiosa para modelos lineales. He et al. (2008) propusieron ADASYN como una mejora sobre SMOTE, pero su efectividad parece depender significativamente del algoritmo de clasificación utilizado. La capacidad de ADASYN para generar más muestras sintéticas en regiones difíciles de clasificar puede ser contraproducente cuando se combina con algoritmos que ya manejan bien la complejidad de los límites de decisión, como *XGBoost* (Douzas & Bacao, 2018).

La técnica híbrida SMOTEENN mostró rendimientos intermedios, confirmando parcialmente su propósito de combinar las ventajas del *oversampling* y *undersampling* mientras mitiga sus desventajas. Sin embargo, no logró superar a *UnderSampler* simple en términos de *recall* máximo. Batista et al. (2004) sugirieron que las técnicas híbridas podrían ser superiores en *datasets* complejos, pero nuestros resultados indican que la simplicidad de *UnderSampler* puede ser preferible en ciertos contextos clínicos.

Desde una perspectiva clínica, el éxito de *UnderSampler* tiene implicaciones prácticas importantes. La reducción del conjunto de datos de entrenamiento puede acelerar significativamente los tiempos de entrenamiento del modelo,

factor crucial para implementaciones en tiempo real en UCI (Beam & Kohane, 2018). La capacidad de identificar correctamente hasta el 70% de los pacientes de alto riesgo representa una herramienta valiosa para la toma de decisiones clínicas, especialmente cuando se considera que estos modelos pueden procesar información de manera continua y objetiva. Sin embargo, es importante considerar que *UnderSampler* descarta información potencialmente valiosa. En contextos donde los datos son escasos o costosos de obtener, esta pérdida puede ser problemática. Para solucionarlo se deberá hacer una validación en *datasets* independientes.

5.5 Interpretabilidad y Explicabilidad (XAI)

Los análisis de interpretabilidad mediante SHAP reflejados en las figuras 22 y 23 proporcionaron *insights* valiosos sobre los factores que impulsan las predicciones de los modelos, aspecto crucial para la aceptación y adopción en entornos clínicos.

La identificación de las variables de la dosis de aminas, la escala APACHE II y foco respiratorio como las variables más influyentes en las predicciones de ambos modelos (*XGBoost* y *LightGBM*) valida la consistencia de los hallazgos y su relevancia clínica. Estos resultados son coherentes con el conocimiento médico establecido sobre factores pronósticos en sepsis. La dosis de aminas vasopresoras ha sido consistentemente identificada como un predictor fuerte de mortalidad en sepsis y shock séptico debido a que reflejan el grado de shock circulatorio (Russell et al., 2018). Su posición como variable más importante en nuestros modelos confirma su valor pronóstico y sugiere que los algoritmos de *machine learning* pueden capturar adecuadamente las relaciones fisiopatológicas conocidas (Evans et al., 2021). El *score* APACHE II, diseñado específicamente para predecir mortalidad hospitalaria en UCI, mostró la segunda mayor importancia predictiva. Este resultado valida tanto nuestro modelo como la utilidad continuada de este sistema de puntuación clásico. La identificación del foco respiratorio como tercera variable de alta importancia refleja la particular gravedad de las infecciones pulmonares en pacientes críticos. Estudios previos han demostrado que la neumonía asociada a ventilación mecánica y otras infecciones respiratorias en UCI tienen tasas de mortalidad superiores a otros focos infecciosos (Chastre & Fagon, 2002; Torres et al., 2017).

El *score* SOFA mostró una importancia significativa pero ligeramente inferior a APACHE II, lo cual puede explicarse por las diferencias en el diseño y propósito de ambos sistemas. Mientras APACHE II fue diseñado para predicción de mortalidad, SOFA se centra en la evaluación de disfunción orgánica (Farzad et al., 2022). Sin embargo, ambos *scores* mostraron

complementariedad en los modelos, sugiriendo que capturan aspectos diferentes pero relevantes del riesgo del paciente. Por último, la albúmina sérica, aunque menos prominente que otros predictores, mostró una contribución significativa. Este hallazgo está bien respaldado por la literatura en relación con la mortalidad hospitalaria (Frenkel et al., 2022; Kim et al., 2024).

La mayor sensibilidad de *XGBoost* a variables específicas como el foco respiratorio y la escala SOFA, evidenciada por mayor dispersión en los valores SHAP, puede reflejar diferencias en las arquitecturas de los algoritmos y sus estrategias de regularización (Chen & Guestrin, 2016; Ke et al., 2017).

5.6 Limitaciones del Estudio

A pesar de los hallazgos prometedores de esta investigación, es fundamental reconocer y discutir las limitaciones que podrían haber influido en los resultados y su interpretación:

El tamaño de la muestra (n=226) representa una limitación significativa para el desarrollo de modelos de *machine learning* robustos. La literatura sugiere que los algoritmos de *machine learning*, particularmente los métodos de ensemble como *XGBoost*, pueden beneficiarse de muestras más grandes para alcanzar su pleno potencial predictivo (Chen et al., 2024). Además, el tamaño relativamente reducido de la cohorte (226 pacientes) hizo necesario utilizar validación cruzada estratificada como estrategia de evaluación en lugar de contar con un conjunto de *test* externo o completamente independiente, lo que permite aprovechar al máximo los datos disponibles y obtener métricas más estables. Asimismo, la presencia de valores atípicos (*outliers*) tiene un impacto proporcionalmente mayor en los resultados. Estos casos corresponden a pacientes reales, y la práctica clínica no siempre sigue patrones estrictamente teóricos, por lo que es esperable cierta variabilidad. En algunas variables observamos comportamientos que no se alinean completamente con lo descrito en la literatura, lo cual contribuyó a la aparición de dichos valores atípicos. Además, la representatividad de la muestra se limita datos recogidos de pacientes ingresados en un único hospital de España, lo que podría limitar la generalización de los resultados a otras poblaciones.

Otra limitación relevante es el desbalance inherente en los datos clínicos, donde los casos de alto riesgo representan la minoría, constituye un desafío metodológico significativo. Aunque se aplicaron múltiples técnicas de balanceo, este desbalance puede haber introducido sesgos en las predicciones y limitado la capacidad de generalización de los modelos. Por otra parte, el

uso de validación cruzada estratificada, aunque metodológicamente apropiado, no reemplaza la validación externa en *datasets* independientes. Un conjunto de validación externa sería muy deseable para establecer la verdadera capacidad de generalización de los modelos desarrollados. La base de datos carece de ciertas variables que podrían ser relevantes para la predicción del riesgo. Como se comentó en el capítulo 3, las variables de frecuencia respiratoria, lactato al ingreso y lactato a las 24 horas, tuvieron que ser eliminadas por falta de registros clínicos de la historia de muchos pacientes. La ausencia de información sobre la evolución temporal de los parámetros durante la estancia en UCI representa otra limitación importante. Los modelos desarrollados se basan en valores de ingreso, pero la evolución dinámica de estos parámetros podría proporcionar información pronóstica adicional. Finalmente, el período de recolección de datos (2024 y primer trimestre de 2025) puede no capturar completamente la variabilidad estacional en las infecciones y los patrones de resistencia antimicrobiana. Estudios longitudinales más extensos serían necesarios para validar la estabilidad temporal de los modelos predictivos.

Capítulo 6. Conclusiones y líneas futuras

6.1 Conclusiones

En este TFG se ha trabajado con un *dataset* de pacientes ingresados en una UCI, monitorizando sus parámetros y variables clínicas desde el momento del ingreso hasta las 24 horas. En primer lugar, se seleccionó el conjunto de características más relevantes evitando el solapamiento para disminuir considerablemente la posibilidad de *overfitting.* Posteriormente, se dividió el *dataset* en un subconjunto de *train* con una validación interna que sirvió como pseudoconjunto de *test* y se entrenaron cinco modelos de ML (*Logistic regression, Random Forest, GXBoost, LightGBM y Balanced Random Forest*) evaluados individualmente con cada una de las cuatro técnicas de balanceo aplicadas (SMOTE, SMOTEENN, ADASYN y submuestreo). Finalmente, los modelos con mayor sensibilidad fueron sometidos a un análisis de XAI para entender qué variables han influido principalmente en la decisión de los algoritmos y así aumentar la confianza en el sistema. Para ello se utilizó la técnica de *SHAP values*. Los resultados obtenidos proporcionan diferentes enfoques que permiten valorar la aplicación del método propuesto en la práctica clínica con el fin de evaluarlo en el contexto clínico crítico.

Tras examinar los resultados, a continuación, se detallan las conclusiones generales extraídas de este TFG:

- Se ha conseguido desarrollar exitosamente modelos predictivos capaces de identificar pacientes de alto riesgo con una sensibilidad de hasta 69.2% (*XGBoost + UnderSampler*), superando significativamente la capacidad discriminativa de las escalas clínicas tradicionales. De entre todos los modelos estudiados, *XGBoost* demostró ser el modelo más eficaz en cuanto a rendimiento, seguido de *LightGMB*, confirmando la superioridad de los algoritmos de *boosting* en contextos médicos complejos. Estos resultados son consistentes con la literatura.

- Un hallazgo particularmente relevante fue la efectividad superior del submuestreo (*UnderSampler*) frente a las técnicas de sobremuestreo sintético (SMOTE, ADASYN). Este resultado difiere en cierta medida de las expectativas teóricas y sugiere que, en *datasets* clínicos de calidad, la eliminación selectiva de casos redundantes puede ser más beneficiosa que la generación artificial de ejemplos sintéticos. Este trabajo constituye una aportación novedosa al aplicar por primera vez técnicas como ADASYN y el submuestreo en un contexto clínico de sepsis, lo que abre nuevas perspectivas para el desarrollo de investigaciones futuras en el ámbito de la medicina predictiva.

- La aplicación de técnicas de explicabilidad artificial mediante SHAP reveló que las variables más influyentes sobre la gravedad de la sepsis fueron la dosis de aminas, la escala APACHE II, el foco respiratorio, la escala SOFA y la albúmina. Esto concuerda con el conocimiento clínico y es consistente con el manejo de los pacientes ingresados en la UCI.

- El impacto clínico potencial de estos resultados es significativo. La capacidad de identificar correctamente aproximadamente 7 de cada 10 pacientes de alto riesgo puede traducirse en intervenciones más tempranas, optimización de recursos asistenciales y, potencialmente, reducción de la morbimortalidad. La implementación de estos modelos podría complementar eficazmente los sistemas de alerta temprana existentes en las UCI.

6.2 Líneas futuras

Los hallazgos de este estudio tienen múltiples implicaciones para la práctica clínica y abren varias líneas prometedoras de investigación futura.

- La implementación clínica real debe ser el objetivo último de esta línea de investigación. El desarrollo de una interfaz clínica integrada con los sistemas hospitalarios existentes permitiría evaluar el impacto real en pacientes. Esta implementación debe incluir estudios de usabilidad, aceptación por parte del personal sanitario y análisis de coste-efectividad.
- La validación de los modelos desarrollados sobre un conjunto de *test* externo, idealmente procedente de otro centro hospitalario o de una cohorte temporal diferente permitiría evaluar la generalización real de los algoritmos y comprobar su robustez frente a variaciones en la población y en la práctica clínica.
- El refinamiento de técnicas de balanceo es una línea de investigación metodológicamente relevante. Los hallazgos sobre la efectividad del submuestreo requieren validación y exploración en otros contextos médicos. El desarrollo de técnicas de balanceo específicamente diseñadas para datos clínicos podría mejorar el rendimiento de los modelos predictivos en medicina.
- El desarrollo de modelos específicos por subgrupos constituye una mejora importante. La creación de algoritmos especializados según el foco infeccioso (respiratorio, urinario, abdominal) o características demográficas específicas (edad, comorbilidades) podría mejorar la precisión predictiva mediante personalización del modelo.
- La implementación en la mejora en la toma de decisiones tendrá una influencia positiva en la atención clínica. Los pacientes identificados

como alto riesgo podrían recibir monitorización más intensiva y acceso prioritario a intervenciones especializadas. La identificación temprana del riesgo podría facilitar intervenciones proactivas antes del deterioro clínico, potencialmente mejorando los desenlaces. Los modelos podrían proporcionar información objetiva para discusiones pronósticas con familiares, mejorando la comunicación médico-paciente y familia.

- Sería deseable realizar una validación multicéntrica de los resultados del estudio. La expansión del estudio a múltiples centros hospitalarios permitiría validar la capacidad de generalización de los algoritmos y resultados de este TFG.

Bibliografía

2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-1, doi: 10.1109/ICCUBEA.2018.8697366.

Ackerman, M. H., Ahrens, T., Kelly, J., & Pontillo, A. (2021). Sepsis. In *Critical Care Nursing Clinics of North America* (Vol. 33, Issue 4, pp. 407–418). W.B. Saunders. https://doi.org/10.1016/j.cnc.2021.08.003

Ahmed Soomro, A., Akmar Mokhtar, A., B Hussin, H., Lashari, N., Lekan Oladosu, T., Muslim Jameel, S., & Inayat, M. (2024). Analysis of machine learning models and data sources to forecast burst pressure of petroleum corroded pipelines: A comprehensive review. In *Engineering Failure Analysis* (Vol. 155). Elsevier Ltd. https://doi.org/10.1016/j.engfailanal.2023.107747

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine, 18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Al-Ahmari, S., & Nadeem, F. (2025). Improving Surgical Site Infection Prediction Using Machine Learning: Addressing Challenges of Highly Imbalanced Data. *Diagnostics*, *15*(4). https://doi.org/10.3390/diagnostics15040501

Alejandro, B. C., Ronald, P. M., & Glenn, H. P. (2011). Manejo del paciente en shock séptico. *Revista Médica Clínica las Condes*, *22*(3), 293-301. https://doi.org/10.1016/s0716-8640(11)70429-1

An Assessment of Sepsis in the United States and its Burden on Hospital Care. Rockville, MD: Agency for Healthcare Research and Quality; September 2024. AHRQ Pub No. 24-0087.

Anderson, K., Brown, T., & Wilson, J. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information*, 14(1), 54. DOI: 10.3390/info14010054

Angus, D. C., & van der Poll, T. (2013). Severe Sepsis and Septic Shock. *New England Journal of Medicine*, *369*(9), 840–851. https://doi.org/10.1056/NEJMra1208623

Antoine, A., & Demeester, S. (2018). *Exploratory data analysis of a clinical study group: Development of a procedure for exploring multidimensional data. Journal Name*, *Vol*(No.), pp–pp.

Arora, J., Mendelson, A. A., & Fox-Robichaud, A. (2023). Sepsis: network pathophysiology and implications for early diagnosis. In *American Journal of Physiology - Regulatory Integrative and Comparative Physiology* (Vol. 324, Issue 5, pp. R613–R624). American Physiological Society. https://doi.org/10.1152/AJPREGU.00003.2023

Arrieta, A. B., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Barreñada, L., Dhiman, P., Timmerman, D. *et al.* Understanding overfitting in random forest for probability estimation: a visualization and simulation study. *Diagn Progn Res* 8, 14 (2024). https://doi.org/10.1186/s41512-024-00177-1

Batko, K., & Ślęzak, A. (2022). The use of Big Data Analytics in healthcare. *Journal of Big Data*, *9*(1). https://doi.org/10.1186/s40537-021-00553-4

Becker, K. L., Snider, R., & Nylen, E. S. (2010). Procalcitonin in sepsis and systemic inflammation: A harmful biomarker and a therapeutic target. In *British Journal of Pharmacology* (Vol. 159, Issue 2, pp. 253–264). https://doi.org/10.1111/j.1476-5381.2009.00433.x

Beesley, S. J., Lanspa, M. J., Barrett, L. K., Grissom, C. K., Wilson, E. L., Brown, S. M., & Talmor, D. (2018). Sepsis-induced cardiomyopathy. *Critical Care Medicine*, 46(4), 625–634. https://doi.org/10.1097/CCM.0000000000002851

Bhargava, A., López-Espina, C., Schmalz, L., Khan, S., Watson, G. L., Urdiales, D., Updike, L., Kurtzman, N., Dagan, A., Doodlesack, A., Stenson, B. A., Sarma, D., Reseland, E., Lee, J. H., Kravitz, M. S., Antkowiak, P. S., Shvilkina, T., Espinosa, A., Halalau, A., … Shapiro, N. I. (2024). FDA-Authorized AI/ML Tool for Sepsis Prediction: Development and Validation. *NEJM AI*, *1*(12). https://doi.org/10.1056/aioa2400867

Bladon, S., Ashiru-Oredope, D., Cunningham, N., Pate, A., Martin, G. P., Zhong, X., Gilham, E. L., Brown, C. S., Mirfenderesky, M., Palin, V., & van Staa, T. P. (2024). Rapid systematic review on risks and outcomes of sepsis: the influence of risk factors associated with health inequalities. In *International Journal for Equity in Health* (Vol. 23, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12939-024-02114-6

Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A., Schein, R. M., & Sibbald, W. J. (1992). Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*, 101(6), 1644–1655. https://doi.org/10.1378/chest.101.6.1644

Caligiuri M. A. (2008). Human natural killer cells. *Blood*, *112*(3), 461–469. https://doi.org/10.1182/blood-2007-09-077438Camacho-Cogollo, J. E., Bonet, I., Gil, B., & Iadanza, E. (2022). Machine Learning Models for Early Prediction of Sepsis on Large Healthcare Datasets. *Electronics (Switzerland)*, *11*(9). https://doi.org/10.3390/electronics11091507

Carobene, A., Milella, F., Famiglini, L., & Cabitza, F. (2022). How is test laboratory data used and characterised by machine learning models? A systematic review of diagnostic and prognostic models developed for COVID-19 patients using only laboratory data. In *Clinical Chemistry and Laboratory Medicine* (Vol. 60, Issue 12, pp. 1887–1901). De Gruyter Open Ltd. https://doi.org/10.1515/cclm-2022-0182

Cerro, L., Valencia, J., Calle, P., León, A., & Jaimes, F. (2014). Validación de las escalas de APACHE II y SOFA en 2 cohortes de pacientes con sospecha de infección y sepsis, no ingresados en unidades de cuidados críticos. *Revista española de Anestesiologia y Reanimacion*, *61*(3), 125–132. https://doi.org/10.1016/j.redar.2013.11.014

Chávez M, Vallejo DE. Susceptibilidad genética para el desarrollo de la sepsis bacteriana grave y choque séptico. *Rev Cienc Salud* 2013; 11 (1): 93-103.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

Cheng C-Y, Kung C-T, Chen F-C, Chiu I-M, Lin C-HR, Chu C-C, Kung CF and Su C-M (2022) Machine learning models for predicting in-hospital mortality in patient with sepsis: Analysis of vital sign dynamics. *Front. Med.* 9:964667. doi: 10.3389/fmed.2022.964667

Chertoff, J., Chisum, M., Garcia, B., & Lascano, J. (2015). Lactate kinetics in sepsis and septic shock: A review of the literature and rationale for further research. In *Journal of*

Intensive Care (Vol. 3, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s40560-015-0105-4

Chicco, D., & Tötsch, N. (2022). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 23*(1), 6.

Chiscano-Camón 2022 Chiscano-Camón, L., Plata-Menchaca, E., Ruiz-Rodríguez, J. C., & Ferrer, R. (2022). Pathophysiology of septic shock. *Medicina Intensiva*, *46*, 1–13. https://doi.org/10.1016/j.medin.2022.03.017

Conigliaro, P., Triggianese, P., Ballanti, E., Perricone, C., Perricone, R., & Chimenti, M. S. (2019). Complement, infection, and autoimmunity. *Current opinion in rheumatology*, *31*(5), 532–541. https://doi.org/10.1097/BOR.0000000000000633

De Cuidados Intensivos Hospital de Manacor Enero 2008, R. P. F. U. (s. f.). *Procalcitonina en UCI*. © R.Pitarch-Rapid Critical Care Consult. https://www.rccc.eu/ppc/indicadores/PCT.htm

De, F., de Mortalidad, R., Cristina, I., & Mestanza, L. (2017). *Departamento de biología celular, histología y farmacología, tesis doctoral:L*

Desautels, T., Calvert, J., Hoffman, J., Jay, M., Kerem, Y., Shieh, L., … & Das, R. (2016). Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. *JMIR Medical Informatics, 4*(3), e28. https://doi.org/10.2196/medinform.5909

Díaz Martín, D., Prieto Martín, A., & Úbeda Cantera Álvarez-Mon Soto, M. M. (2013). Linfocitos T. In *Medicine* (Vol. 11, Issue 28).

El, I., Li, N. R., & Murphy, M. J. (n.d.). *Theory and Applications Machine Learning in Radiation Oncology* (2015).

Erdogan, M., & Findikli, H. A. (2021). The Relationship Between Hypoalbuminemia and Mortality in Geriatric Patients with Sepsis. *Journal of Critical and Intensive Care*. https://doi.org/10.37678/dcybd.2021.2824

Esposito, S., de Simone, G., Boccia, G., de Caro, F., & Pagliano, P. (2017). Sepsis and septic shock: New definitions, new diagnostic and therapeutic approaches. *Journal of Global Antimicrobial Resistance*, *10*, 204–212. https://doi.org/10.1016/J.JGAR.2017.06.013

Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith CM, French C, et al. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock 2021. *Intensive Care Med*. 2021;47(11):1181–1247. doi:10.1007/s00134-021-06506-y.

Evans, T. (2017). CMJv18n2-CMEEvans.indd. In *CME INFECTIOUS DISEASES Clinical Medicine* (Vol. 17, Issue 6). www.nice.org.uk/guidance/ng51

Fabiola Brito Galeana, D., Yamazaki, M. A., Sara Espinosa Padilla, D. S., Vázquez Tsuji, Ó., Huerta López, J., & Berrón Pérez, R. (2003). *Eosinófilos: Revisión de la literatura* (Vol. 12).

Fleischmann-Struzek, C., Mellhammar, L., Rose, N. *et al.* Incidence and mortality of hospital- and ICU-treated sepsis: results from an updated and expanded systematic review and meta-analysis. *Intensive Care Med* 46 1552–1562 (2020). https://doi.org/10.1007/s00134-020-06151-x

Fleuren, L. M., Güiza, F., de Keizer, N. F., Schoonhoven, L., & Slooter, A. J. (2020). Large variation in risk and mortality of sepsis across hospitals: a retrospective study. *Critical Care*, 24(1), 270. https://doi.org/10.1186/s13054-020-03018-9

Fleuren, L. M., Klausch, T. L. T., Zwager, C. L., Schoonmade, L. J., Guo, T., Roggeveen, L. F., Swart, E. L., Girbes, A. R. J., Thoral, P., Ercole, A., Hoogendoorn, M., & Elbers, P. W. G. (2020). Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. In *Intensive Care Medicine* (Vol. 46, Issue 3, pp. 383–400). Springer. https://doi.org/10.1007/s00134-019-05872-y

Font, M. D., Thyagarajan, B., & Khanna, A. K. (2020). Sepsis and Septic Shock – Basics of diagnosis, pathophysiology and clinical decision making. *Medical Clinics of North America*, *104*(4), 573–585. https://doi.org/10.1016/J.MCNA.2020.02.011

Francisco Guzmán González Servicio de Microbiología Hospital Juan Ramón Jiménez Huelva, A. (2018). *HEMOCULTIVOS*.

Frenkel, A., Novack, V., Bichovsky, Y., Klein, M., & Dreiher, J. (2022). Serum Albumin Levels as a Predictor of Mortality in Patients with Sepsis: A Multicenter Study. *The Israel Medical Association journal : IMAJ*, *24*(7), 454–459.

Gao, J., Lu, Y., Ashrafi, N., Domingo, I., Alaei, K., & Pishgar, M. (2024). Prediction of sepsis mortality in ICU patients using machine learning methods. *BMC medical informatics and decision making*, *24*(1), 228. https://doi.org/10.1186/s12911-024-02630-z

García Barreno, P. (2008). INFLAMACIÓN. In *Cienc.Exact.Fís.Nat. (Esp)* (Vol. 102, Issue 1).

Giannini, H. M., Ginestra, J. C., Chivers, C., Draugelis, M., Hanish, A., Schweickert, W. D., Fuchs, B. D., Meadows, L., Lynch, M., Donnelly, P. J., Pavan, K., Fishman, N. O., Hanson, C. W., 3rd, & Umscheid, C. A. (2019). A Machine Learning Algorithm to Predict Severe Sepsis and Septic Shock: Development, Implementation, and Impact on Clinical Practice. *Critical care medicine*, *47*(11), 1485–1492. https://doi.org/10.1097/CCM.0000000000003891

Gomez, H. G., Rugeles, M. T., & Jaimes, F. A. (2015). Key immunological characteristics in the pathophysiology of sepsis. In *Infectio* (Vol. 19, Issue 1, pp. 40–46). Elsevier Doyma. https://doi.org/10.1016/j.infect.2014.03.001

González, B., & Profesor, R. (n.d.). *EL PROCESO INFLAMATORIO*. http://www.uclm.es/ab/enfermeria/revista/numero%204/pinflamatorio4.htm[18/02/201011:54:21]ELPROCESOINFLAMATORIO

Grimm E. A. (1986). Human lymphokine-activated killer cells (LAK cells) as a potential immunotherapeutic modality. *Biochimica et biophysica acta*, *865*(3), 267–279. https://doi.org/10.1016/0304-419x(86)90017-x

Guarino, M., Costanzini, A., Luppi, F., Maritati, M., Contini, C., De Giorgio, R., & Spampinato, M. D. (2025). Extracorporeal Cytokine Adsorption in Sepsis: Current Evidence and Future Perspectives. *Biomedicines*, *13*(7), 1684. Doi: 10.3390/biomedicines13071684

Guo Q, Li W, Wang J, Wang G, Deng Q, Lian H, Wang X. Construction and validation of a clinical prediction model for sepsis using peripheral perfusion index to predict in-hospital

and 28-day mortality risk. Sci Rep. 2024 Nov 5;14(1):26827. doi: 10.1038/s41598-024-78408-0. PMID: 39501076; PMCID: PMC11538300.

Habimana R, Choi I, Cho HJ, Kim D, Lee K, Jeong I. Sepsis-induced cardiac dysfunction: a review of pathophysiology. Acute Crit Care. 2020 May;35(2):57-66. doi: 10.4266/acc.2020.00248. Epub 2020 May 31. PMID: 32506871; PMCID: PMC7280799.

Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism*, *69*, S36–S40. https://doi.org/10.1016/J.METABOL.2017.01.011

Hassanzadeh R, Farhadian M, Rafieemehr H. Hospital mortality prediction in traumatic injuries patients: comparing different SMOTE-based machine learning algorithms. BMC Med Res Methodol. 2023 Apr 22;23(1):101. doi: 10.1186/s12874-023-01920-w. PMID: 37087425; PMCID: PMC10122327.

He B and Qiu Z (2024) Development and validation of an interpretable machine learning for mortality prediction in patients with sepsis. *Front. Artif. Intell*. 7:1348907. doi: 10.3389/frai.2024.1348907

Hotchkiss, R., Monneret, G. & Payen, D. Sepsis-induced immunosuppression: from cellular dysfunctions to immunotherapy. *Nat Rev Immunol* **13**, 862–874 (2013). https://doi.org/10.1038/nri3552

Hu, C., Li, L., Huang, W. *et al.* Interpretable Machine Learning for Early Prediction of Prognosis in Sepsis: A Discovery and Validation Study. *Infect Dis Ther* **11**, 1117–1132 (2022). https://doi.org/10.1007/s40121-022-00628-6

Huang, J., Yang, J. T., & Liu, J. C. (2023). The association between mortality and door-to-antibiotic time: a systematic review and meta-analysis. *Postgraduate medical journal*, *99*(1175), 1000–1007. https://doi.org/10.1093/postmj/qgad024

Ince, C., Mayeux, P. R., Nguyen, T., Gomez, H., Kellum, J. A., Ospina-Tascón, G. A., ... & De Backer, D. (2016). The endothelium in sepsis. *Shock*, 45(3), 259–270. https://doi.org/10.1097/SHK.0000000000000598

Institute for Quality and Efficiency in Health Care (IQWiG). (2023, 14 agosto). *In brief: The innate and adaptive immune systems*. InformedHealth.org - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK279396/

Islam, K. R., Prithula, J., Kumar, J., Tan, T. L., Reaz, M. B. I., Sumon, M. S. I., & Chowdhury, M. E. H. (2023). Machine Learning-Based Early Prediction of Sepsis Using Electronic Health Records: A Systematic Review. *Journal of Clinical Medicine*, *12*(17), 5658. https://doi.org/10.3390/jcm12175658

Jantos, B.A., Wierzbicki, M.P., Tomaszewski, M. (2025). Standardization Processes for the Implementation of AI in Healthcare System: Challenges and Limitations. In: Zamojski, W., Mazurkiewicz, J., Sugier, J., Walkowiak, T., Kacprzyk, J. (eds) Advances in Dependable Systems and Networks. DepCoS-RELCOMEX 2025. Lecture Notes in Networks and Systems, vol 1427. Springer, Cham. https://doi.org/10.1007/978-3-031-92734-8_7

Jiang, Y., Li, X., Luo, H., Yin, S., & Kaynak, O. (2022). Quo vadis artificial intelligence? In *Discover Artificial Intelligence* (Vol. 2, Issue 1). Springer Nature. https://doi.org/10.1007/s44163-022-00022-8

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L. H., Celi, L. A., & Mark, R. G. (2023). MIMIC-

IV, a freely accessible electronic health record dataset. *Scientific data*, *10*(1), 1. https://doi.org/10.1038/s41597-022-01899-x

Johnson, P., Davis, R., & Miller, S. (2023). Random Forest vs. XGBoost vs. LightGBM: Which Ensemble Learner Performs Best on Imbalanced Data? *ResearchGate Technical Report*, DOI: 10.13140/RG.2.2.15234.67201

Jović, S., Brkić, K., & Bogunović, N. (2020). A review of feature selection methods with applications. *Pattern Recognition, 103*, 107210. https://doi.org/10.1016/j.patcog.2019.107210

Julián-Jiménez, A., Candel-González, F. J., & González Del Castillo, J. (2014). Utilidad de los biomarcadores de inflamación e infección en los servicios de urgencias. In *Enfermedades Infecciosas y Microbiologia Clinica* (Vol. 32, Issue 3, pp. 177–190). https://doi.org/10.1016/j.eimc.2013.01.005

Kausch, S. L., Moorman, J. R., Lake, D. E., & Keim-Malpass, J. (2021). Physiological machine learning models for prediction of sepsis in hospitalized adults: An integrative review. *Intensive & critical care nursing*, *65*, 103035. https://doi.org/10.1016/j.iccn.2021.103035

Kausch, S. L., Wernly, B., Lichtenauer, M., Franz, M., Kabisch, B., Muessig, J. M., … & Follmann, M. (2021). Machine learning in critical care and emergency medicine: A systematic review. *Medical Science Monitor, 27*, e930889. https://doi.org/10.12659/MSM.930889

Ke, J. X. C., DhakshinaMurthy, A., George, R. B., & Branco, P. (2024). The effect of resampling techniques on the performances of machine learning clinical risk prediction models in the setting of severe class imbalance: development and internal validation in a retrospective cohort. *Discover Artificial Intelligence*, *4*(1). https://doi.org/10.1007/s44163-024-00199-0

Khoshnevisan, F., Ivy, J., Capan, M., Arnold, R., Huddleston, J., & Chi, M. (2018). Recent temporal pattern mining for septic shock early prediction. *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*, 229–240. https://doi.org/10.1109/ICHI.2018.00033

Khurana, D., & Anand, S. (2023). Encoding Techniques for Handling Categorical Data in Machine Learning. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence* (pp. 434–445). SCITEPRESS. https://www.scitepress.org/Papers/2023/122594/122594.pdf

Kim, H., Lee, J., & Park, S. (2025). SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models. *Algorithms*, 18(1), 37. DOI: 10.3390/a18010037

Kim, S.-M., Ryoo, S.-M., Shin, T.-G., Jo, Y.-H., Kim, K., Lim, T.-H., … Kim, W.-Y. (2024). Early Mortality Stratification with Serum Albumin and the Sequential Organ Failure Assessment Score at Emergency Department Admission in Septic Shock Patients. *Life*, 14(10), 1257. https://doi.org/10.3390/life14101257

Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., ... & Dobson, R. J. B. (2021). MedCATTrainer: a biomedical free text annotation interface with active learning and research use case specific customisation. *Journal of the American Medical Informatics Association, 28*(10), 2203–2210. https://doi.org/10.1093/jamia/ocab103

Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., Suppes, R., Feinstein, D., Zanotti, S., Taiberg, L., Gurka, D., Kumar, A., & Cheang, M. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical Care Medicine, 34(6), 1589-1596.

Langius, P., Perez, D., Lopez, S., & Boccio, E. (2024). The Diagnostic Utility of Fever in Patients With Suspected Sepsis Secondary to Community-Acquired Bacteremia. *Cureus*, *16*(10), e72561. https://doi.org/10.7759/cureus.72561

Lee, K. J., Tilling, K. M., Cornish, R. P., Little, R. J. A., Bell, M. L., Goetghebeur, E., Hogan, J. W., & Carpenter, J. R.; STRATOS initiative. (2021). *Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework*. Journal of Clinical Epidemiology, 134, 79–88. https://doi.org/10.1016/j.jclinepi.2021.01.008

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature selection: A data perspective. *ACM Computing Surveys (CSUR), 50*(6), 1–45. https://doi.org/10.1145/3136625

Li, M., Huang, P., Xu, W., Zhou, Z., Xie, Y., Chen, C., Jiang, Y., Cui, G., Zhao, Q., & Wang, R. (2022). Risk factors and a prediction model for sepsis: A multicenter retrospective study in China. *Journal of intensive medicine*, *2*(3), 183–188. https://doi.org/10.1016/j.jointm.2022.02.004

Li, X., Zhang, Y., Wang, J., Liu, H., & Chen, Z. (2024). An interpretable machine learning model for early prediction of 28-day mortality in ICU patients with sepsis-induced coagulopathy: Development and validation. *European Journal of Medical Research, 29*(1), 18. https://doi.org/10.1186/s40001-023-01593-7

Liang, Y., et al. (2022). *Early prediction of septic shock using machine learning in intensive care unit patients.* BMC Medical Informatics and Decision Making, 22, 153.

Liu, X., et al. (2021). Comparative evaluation of machine learning models for early prediction of sepsis. *BMC Medical Informatics and Decision Making*, 21, 145.

Liu, Z., Shu, W., Li, T. *et al.* Interpretable machine learning for predicting sepsis risk in emergency triage patients. *Sci Rep* **15**, 887 (2025). https://doi.org/10.1038/s41598-025-85121-z

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., … Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2, 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mahapatra S, Heffner AC. Shock séptico. [Actualizado el 12 de junio de 2023]. En: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; enero de 2025. Disponible en: https://www.ncbi.nlm.nih.gov/books/NBK430939/

Mahesh, B. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJSR)*, *9*(1), 381–386. https://doi.org/10.21275/art20203995

Mao, Q., et al. (2020). Development and validation of a logistic regression model to predict sepsis in primary care. *BMJ Open*, 10(8), e038090.

Marik PE, Taeb AM. SIRS, qSOFA and new sepsis definition. J Thorac Dis. 2017 Apr;9(4):943-945. doi: 10.21037/jtd.2017.03.125. PMID: 28523143; PMCID: PMC5418298.

Markanday A. Acute Phase Reactants in Infections: Evidence-Based Review and a Guide for Clinicians. Open Forum Infect Dis. 2015 Jul 3;2(3):ofv098. doi: 10.1093/ofid/ofv098. PMID: 26258155; PMCID: PMC4525013.

Martín, A. P., Barbarroja Escudero, J., Barcenilla Rodríguez, H., & Díaz Martín, D. (2013). Funciones de los linfocitos B Funciones principales de los linfocitos B. In *Medicine* (Vol. 11, Issue 28).

Martín, M. J. A., Bernal, M. H., Yus Teruel, S., & Minvielle, A. (2018). Infecciones en el paciente crítico. In *Medicine* (Vol. 12, Issue 52).

Médica, G. (2024, September 13). *Cada año fallecen cerca de 20.000 personas en España a causa de la sepsis*. Gaceta Médica. https://gacetamedica.com/profesion/cada-ano-fallecen-cerca-de-20-000-personas-en-espana-a-causa-de-la-sepsis

Mirzakhani, F., Sadoughi, F., Hatami, M. *et al.* Which model is superior in predicting ICU survival: artificial intelligence versus conventional approaches. *BMC Med Inform Decis Mak* **22**, 167 (2022). https://doi.org/10.1186/s12911-022-01903-9

Molnar, C. (2022). *Interpretable machine learning*. Lulu.com. https://christophm.github.io/interpretable-ml-book/

Monserrat Sanz, J., Gómez Lahoz, A. M., Sosa Reina, M. D., & Prieto Martín, A. (2017). Introducción al sistema inmune. Componentes celulares del sistema inmune innato. *Medicine - Programa de Formación Médica Continuada Acreditado*, *12*(24), 1369–1378. https://doi.org/10.1016/J.MED.2016.12.006

Moor, M., Bennett, N., Plečko, D., Horn, M., Rieck, B., Meinshausen, N., Bühlmann, P., & Borgwardt, K. (2023). *Predicting sepsis using deep learning across international sites: a retrospective development and validation study*. www.thelancet.com.

Moor, M., Rieck, B., Horn, M., Jutzeler, C. R., & Borgwardt, K. (2021). Early prediction of sepsis in the ICU using machine learning: A systematic review. *Frontiers in Medicine, 8*, 607952. https://doi.org/10.3389/fmed.2021.607952

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal, 24*(3), 69–71. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/

Nikravangolsefid, N., Reddy, S., Truong, H. H., Charkviani, M., Ninan, J., Prokop, L. J., Suppadungsuk, S., Singh, W., Kashani, K. B., & Garces, J. P. D. (2024). Machine learning for predicting mortality in adult critically ill patients with Sepsis: A systematic review. In *Journal of Critical Care* (Vol. 84). W.B. Saunders. https://doi.org/10.1016/j.jcrc.2024.154889

Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., & Buchman, T. G. (2018). An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Critical Care Medicine, 46*(4), 547–553. https://doi.org/10.1097/CCM.0000000000002936

Neutrófilos | British Society for Immunology. (s. f.). https://www.immunology.org/es/public-information/inmunolog%C3%ADa-bitesized/celulas/neutrofilos

Nguyen, H. B. (2011). Lactate in the critically ill patients: An outcome marker with the times. In *Critical Care* (Vol. 15, Issue 6). https://doi.org/10.1186/cc10531

Ocampo-Quintero, N., Vidal-Cortés, P., del Río Carbajo, L., Fdez-Riverola, F., Reboiro-Jato, M., & Glez-Peña, D. (2022). Enhancing sepsis management through machine learning techniques: A review. In *Medicina Intensiva* (Vol. 46, Issue 3, pp. 140–156). Ediciones Doyma, S.L. https://doi.org/10.1016/j.medin.2020.04.003

Palencia Herrejón, E., & Bueno García, B. (2013). Nuevas guías de práctica clínica de la «Campaña sobrevivir a la sepsis»: Lectura crítica. *Medicina Intensiva*, *37*(9), 600–604. https://doi.org/10.1016/j.medin.2013.07.008

Parkin, J., & Cohen, B. (2001). An overview of the immune system. In *Lancet* (Vol. 357, Issue 9270, pp. 1777–1789). Elsevier B.V. https://doi.org/10.1016/S0140-6736(00)04904-7

Pierrakos, C., & Vincent, J.-L. (2010). *Sepsis biomarkers: a review.* Critical Care, 14, R15. https://doi.org/10.1186/cc8872

Potdar, K., Pardawala, T. S., & Pai, C. D. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications, 175*(4), 7–9. https://ijcaonline.org/archives/volume175/number4/potdar-2017-ijca-915495.pdf

Prithula, J., Islam, K. R., Kumar, J., Tan, T. L., Reaz, M. B. I., Rahman, T., Zughaier, S. M., Khan, M. S., Murugappan, M., & Chowdhury, M. E. H. (2025). A novel classical machine learning framework for early sepsis prediction using electronic health record data from ICU patients. *Computers in biology and medicine*, *184*, 109284. https://doi.org/10.1016/j.compbiomed.2024.109284

Rabinowitz, J. D., & Enerbäck, S. (2020). Lactate: the ugly duckling of energy metabolism. *Nature Metabolism*, *2*(7), 566–571. https://doi.org/10.1038/s42255-020-0243-4

Raia, P., & Zafrani, L. (2022). *Endothelial Activation and Microcirculatory Disorders in Sepsis*. *Frontiers in Medicine, 9*, 907992. https://doi.org/10.3389/fmed.2022.907992

Raymond SL, Holden DC, Mira JC, Stortz JA, Loftus TJ, Mohr AM, Moldawer LL, Moore FA, Larson SD, Efron PA. Microbial recognition and danger signals in sepsis and trauma. Biochim Biophys Acta Mol Basis Dis. 2017 Oct;1863(10 Pt B):2564-2573. doi: 10.1016/j.bbadis.2017.01.013. Epub 2017 Jan 20. PMID: 28115287; PMCID: PMC5519458.

Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. In *Computers in Biology and Medicine* (Vol. 112). Elsevier Ltd. https://doi.org/10.1016/j.compbiomed.2019.103375

Rhee, C., Dantes, R., Epstein, L., Murphy, D. J., Seymour, C. W., Iwashyna, T., & Klompas, M. (2019). Incidence and Trends of Sepsis in US Hospitals Using Clinical vs Claims Data, 2009-2014. *JAMA*, 318(13), 1241-1249. https://doi.org/10.1001/jama.2017.13836

Rudd, K. E., Johnson, S. C., Agesa, K. M., Shackelford, K. A., Tsoi, D., Kievlan, D. R., Colombara, D. V., Ikuta, K. S., Kissoon, N., Finfer, S., Fleischmann-Struzek, C., Machado, F. R., Reinhart, K. K., Rowan, K., Seymour, C. W., Watson, R. S., West, T. E., Marinho, F., Hay,

S. I., Lozano, R., … Naghavi, M. (2020). Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study. *Lancet (London, England)*, *395*(10219), 200–211. https://doi.org/10.1016/S0140-6736(19)32989-7

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Salmi, M., Atif, D., Oliva, D. *et al.* Handling imbalanced medical datasets: review of a decade of research. *Artif Intell Rev* **57**, 273 (2024). https://doi.org/10.1007/s10462-024-10884-2

Schober P, Vetter TR. Logistic Regression in Medical Research. Anesth Analg. 2021 Feb 1;132(2):365-366. doi: 10.1213/ANE.0000000000005247. PMID: 33449558; PMCID: PMC7785709.

Scikit-learn developers. (2025). *Out-of-bag error in Random Forests*. In *scikit-learn: Machine learning in Python (version 1.7.1)*. Scikit-learn.

Scikit-learn developers. (2021). *KNNImputer — scikit-learn 1.7.2 documentation*. https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html

Seger, C. (2024). Comparative Study on the Performance of Categorical Variable Encoding Techniques. *arXiv preprint arXiv:2401.09682.* https://arxiv.org/html/2401.09682v1

Sendak, M. P., et al. (2020). The human side of AI: Clinical applications of machine learning in healthcare. *NPJ Digital Medicine, 3*(1), 54. https://doi.org/10.1038/s41746-020-0288-5

Shankar-Hari M, Phillips GS, Levy ML, et al. Developing a New Definition and Assessing New Clinical Criteria for Septic Shock: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):775–787. doi:10.1001/jama.2016.0289

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2022). Deep EHR: A survey of recent advances on deep learning techniques for electronic health record (EHR) analysis. *Journal of Biomedical Informatics, 126*, 103994. https://doi.org/10.1016/j.jbi.2022.103994

Shillan, D., Sterne, J. A., Champneys, A., & Gibbison, B. (2019). *Use of machine learning to analyse routinely collected intensive care unit data: A systematic review.* Critical Care, 23(1), 284.

Shumilov, A., Zhu, Y., Ashrafi, N., Abdollahi, A., Placencia, G., Alaei, K., & Pishgar, M. (2025). *Data-Driven Machine Learning Approaches for Predicting In-Hospital Sepsis Mortality*. http://arxiv.org/abs/2408.01612

Singer, M., Deutschman, C. S., Seymour, C., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J. D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., Poll, T. der, Vincent, J. L., & Angus, D. C. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). In *JAMA - Journal of the American Medical Association* (Vol. 315, Issue 8, pp. 801–810). American Medical Association. https://doi.org/10.1001/jama.2016.0287

Spoto, S., Marika Lupoi, D., Feng Kung, C., & Su, C.-M. (n.d.). *Machine learning models for predicting in-hospital mortality in patient with sepsis: Analysis of vital sign dynamics*.

Srinivasan A, Wong FK, Karponis D. Calcitonin: A useful old friend. J Musculoskelet Neuronal Interact. 2020 Dec 1;20(4):600-609. PMID: 33265089; PMCID: PMC7716677

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics, 17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Taylor, R. A., Pare, J. R., Venkatesh, A. K., Mowafi, H., Melnick, E. R., Fleischman, W., & Hall, M. K. (2016). Prediction of In-hospital Mortality in Emergency Department Patients with Sepsis: A Local Big Data-Driven, Machine Learning Approach. *Academic Emergency Medicine*, *23*(3), 269–278. https://doi.org/10.1111/acem.12876

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2021). Variable ranking and selection with random forest for unbalanced data. *Environmental Data Science*, 3(1), e133. https://doi.org/10.1017/eds.2021.133

Villao Recalde, A. C., Vásquez Cabello, Á. A., Pérez Falconi, N. E., & Padovani Velásquez, A. L. (2019). La sepsis como causa de daño renal. *RECIMUNDO*, *3*(2), 628–650. https://doi.org/10.26820/recimundo/3.(2).abril.2019.628-650

Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., de Mendon~a, A., Bruining, H., Reinhart, C. K., Suter, P. M., Thijs, L. G., de Mendonga, A., & Suter, R. M. (1996). The SOFA (Sepsis.related Organ Failure Assessment) score to describe organ dysfunction/failure. In *Intensive Care Med* (Vol. 22). Springer-Verlag.

Wang, J., Zhang, X., Li, M., & Liu, Y. (2024). *Comparison of missing data imputation methods in cardiovascular cohort studies: Evaluation of KNN and Random Forest approaches*. BMC Medical Research Methodology, 24, 173

Wang, R., Lan, C., Benlagha, K., Camara, N. O. S., Miller, H., Kubo, M., Heegaard, S., Lee, P., Yang, L., Forsman, H., Li, X., Zhai, Z., & Liu, C. (2024). The interaction of innate immune and adaptive immune system. *MedComm*, *5*(10). https://doi.org/10.1002/mco2.714

Wang, X., Guo, Z., Chai, Y., Wang, Z., Liao, H., Wang, Z., & Wang, Z. (2023). Application Prospect of the SOFA Score and Related Modification Research Progress in Sepsis. In *Journal of Clinical Medicine* (Vol. 12, Issue 10). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/jcm12103493

West, E. E., Kolev, M., & Kemper, C. (2018). Complement and the Regulation of T Cell Responses. In *Annual Review of Immunology* (Vol. 36, pp. 309–338). Annual Reviews Inc. https://doi.org/10.1146/annurev-immunol-042617-053245

Wiedermann C. J. (2021). Hypoalbuminemia as Surrogate and Culprit of Infections. *International journal of molecular sciences*, *22*(9), 4496. https://doi.org/10.3390/ijms22094496

Wolf, N. K., Kissiov, D. U., & Raulet, D. H. (2023). Roles of natural killer cells in immunity to cancer, and applications to immunotherapy. *Nature reviews. Immunology*, *23*(2), 90–105. https://doi.org/10.1038/s41577-022-00732-1

Wu, C.-C., Lan, H.-M., Han, S.-T., Chaou, C.-H., Yeh, C.-F., Liu, S.-H., Chen, K.-F. (2020). Comparison of diagnostic accuracy in sepsis between procalcitonin, C-reactive protein, and lactate: A systematic review and meta-analysis. *Annals of Intensive Care*, 10, 20. https://doi.org/10.1186/s13613-020-0643-1

Wu, Q., Ye, F., Gu, Q., Shao, F., Long, X., Zhan, Z., Zhang, J., He, J., Zhang, Y., & Xiao, Q. (2024). A customised down-sampling machine learning approach for sepsis prediction. *International Journal of Medical Informatics*, *184*. https://doi.org/10.1016/j.ijmedinf.2024.105365

Yang, C., Fridgeirsson, E.A., Kors, J.A. *et al.* Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *J Big Data* 11, 7 (2024). https://doi.org/10.1186/s40537-023-00857-7

Yang, L., Yang, J., Zhang, X. *et al.* Predictive value of soluble CD40L combined with APACHE II score in elderly patients with sepsis in the emergency department. *BMC Anesthesiol* 24, 32 (2024). https://doi.org/10.1186/s12871-023-02381-w

Yang, Z., Cui, X., & Song, Z. (2023). Predicting sepsis onset in ICU using machine learning models: a systematic review and meta-analysis. *BMC Infectious Diseases*, *23*(1). https://doi.org/10.1186/s12879-023-08614-0

Yao, W., et al. (2023). Explainable AI for Sepsis Prediction Using LightGBM and SHAP. *IEEE Journal of Biomedical and Health Informatics*, 27(1), 66-75.

Yilmaz Başer, H., Evran, T., & Cifci, M. A. (2025). Machine Learning-Augmented Triage for Sepsis: Real-Time ICU Mortality Prediction Using SHAP-Explained Meta-Ensemble Models. *Biomedicines*, *13*(6), 1449. https://doi.org/10.3390/biomedicines13061449

Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 856–863.

Zaki, H.A., Bensliman, S., Bashir, K. *et al.* Accuracy of procalcitonin for diagnosing sepsis in adult patients admitted to the emergency department: a systematic review and meta-analysis. *Syst Rev* **13**, 37 (2024). https://doi.org/10.1186/s13643-023-02432-w

Zhang, G., Shao, F., Yuan, W. *et al.* Predicting sepsis in-hospital mortality with machine learning: a multi-center study using clinical and inflammatory biomarkers. *Eur J Med Res* 29, 156 (2024). https://doi.org/10.1186/s40001-024-01756-0

Zhang, SZ., Ding, HY., Shen, YM. *et al.* Harness machine learning for multiple prognoses prediction in sepsis patients: evidence from the MIMIC-IV database. *BMC Med Inform Decis Mak* **25**, 152 (2025). https://doi.org/10.1186/s12911-025-02976-y

Zhang, Y., Xu, W., Yang, P., & Zhang, A. (2023). Machine learning for the prediction of sepsis-related death: a systematic review and meta-analysis. *BMC Medical Informatics and Decision Making*, *23*(1). https://doi.org/10.1186/s12911-023-02383-1

Zhou, S., Lu, Z., Liu, Y. *et al.* Interpretable machine learning model for early prediction of 28-day mortality in ICU patients with sepsis-induced coagulopathy: development and validation. *Eur J Med Res* 29, 14 (2024). https://doi.org/10.1186/s40001-023-01593-7

Zhou, X., et al. (2024). *Explainable machine learning for sepsis prediction: A systematic review.* Artificial Intelligence in Medicine, 148, 102723.

Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit Care Med.* 2006;34(5):1297–310.