Aplicación de machine learning para la predicción de resistencia antibiótica en infecciones intraabdominales en pacientes hospitalizados

Olga Rodríguez Núñez, MD, PhD

RESUMEN

Introducción: La antibioterapia empírica inadecuada frente a microorganismos resistentes a los antibióticos (MRA) en infecciones intraabdominales (IIA) se asocia a elevada morbilidad y mortalidad. El objetivo de este estudio fue desarrollar y evaluar modelos de machine learning (ML) para predecir el riesgo de IIA por MRA en pacientes hospitalizados.

Métodos: Se llevó a cabo un estudio retrospectivo en 1.532 pacientes adultos con IIA bacteriémica en un hospital terciario. Para abordar el desbalance de clases (17,2 % de MRA), se generaron 2.500 registros sintéticos mediante Synthetic Data Vault (SDV). Se entrenaron y compararon cuatro modelos individuales (regresión logística, random forest, XGBoost y CatBoost) y dos modelos stacking (Stacking 3c y Stacking 4c). El rendimiento se evaluó mediante AUC, recall, F1-score y precisión. Se priorizó la sensibilidad ajustando el umbral de clasificación.

Resultados: El modelo Stacking 3c, integrado por regresión logística, random forest y XGBoost, y entrenado con el 100 % de datos sintéticos, alcanzó un recall de 0,774, identificando la mayoría de los casos positivos, con una precisión de 0,228 y un AUC test de 0,705. El modelo Stacking 4c, que incluyó CatBoost y utilizó un 40 % de datos sintéticos, obtuvo un AUC test de 0,714, una precisión de 0,355 y un recall de 0,509. El análisis de importancia de variables por permutación identificó como predictores claves los antibióticos previos, la edad y los días de ingreso.

Conclusiones: El modelo Stacking 3c mostró una alta capacidad para detectar infecciones por MRA, contribuyendo a mejorar el inicio del tratamiento empírico y los desenlaces clínicos.

Palabras clave: infecciones intraabdominales, microorganismos resistentes a antibióticos, machine learning, modelos predictivos, antibioterapia empírica.

ABSTRACT

Introduction: Inadequate empirical antibiotic therapy against multidrug-resistant organisms (MDROs) in intra-abdominal infections (IAIs) is associated with high morbidity and mortality. The aim of this study was to develop and evaluate machine learning (ML) models to predict the risk of MDRO-related IAIs in hospitalized patients.

Methods: A retrospective study was conducted including 1,532 adult patients with bacteremic IAIs admitted to a tertiary hospital. To address class imbalance (17.2% MDRO), 2,500 synthetic records were generated using the Synthetic Data Vault (SDV). Four individual models (logistic regression, random forest, XGBoost, and CatBoost) and two stacking models (Stacking 3c and Stacking 4c) were trained and compared. Performance was assessed using AUC, recall, F1-score, and precision. Sensitivity was prioritized by adjusting the classification threshold.

Results: The Stacking 3c model, composed of logistic regression, random forest, and XGBoost, and trained with 100% synthetic data, achieved a recall of 0.774, identifying the majority of positive cases, with a precision of 0.228 and a test AUC of 0.705. The Stacking 4c model, which included CatBoost and used 40% synthetic data, reached a test AUC of 0.714, a precision of 0.355, and a recall of 0.509. Permutation-based feature importance analysis identified prior antibiotic exposure, age, and length of hospital stay as the most relevant predictors of MDRO infection risk.

Conclusions: The Stacking 3c model demonstrated strong performance in identifying patients with MDRO-related IAIs, supporting timely initiation of appropriate empirical therapy and potentially improving clinical outcomes.

Keywords: intra-abdominal infections, multidrug-resistant organisms, machine learning, predictive models, empirical antibiotic therapy.

E-mail de contacto: olrodrig@clinic.cat
Trabajo tutorizado por: Edwar Macias Toro
Curso: 2025

INTRODUCCIÓN

Las infecciones causadas por microorganismos resistentes a los antibióticos (MRA) constituyen una de las principales amenazas para la seguridad de los pacientes hospitalizados y un desafío creciente en la práctica clínica [1]. El tratamiento empírico inicial plantea un reto cuando existen factores de riesgo de resistencia, algo que ocurre con frecuencia en las IIA, donde la tasa de MRA es especialmente alta [2]. Una IIA causada por un MRA se asocia a elevadas tasas de morbilidad, mortalidad (pudiendo variar hasta el 60%) y uso de recursos hospitalarios [3,4]. El tratamiento de los pacientes con IIA se debe abordar con cierta urgencia, frecuentemente mediante una intervención quirúrgica para control del foco combinada con terapia antibiótica. Sin embargo, la elección de la antibioterapia empírica adecuada no siempre se logra; en una revisión sistemática de infecciones hospitalarias graves, la inadecuación empírica por espectro insuficiente se halló entre 14% y 79% de los casos, y en más del 50 % de los trabajos, la incidencia fue ≥ 50 % [5].

Tradicionalmente, la identificación de los factores de riesgo de MRA se ha basado en análisis estadísticos clásicos, como modelos de regresión logística multivariante [6-10]. Estudios previos han identificado así múltiples factores de riesgo asociados a la aparición de infecciones por MRA en el foco IIA, como la inmunosupresión, la cirrosis hepática, la hospitalización prolongada, la manipulación biliar o el uso previo de antibióticos de amplio espectro [10]. Sin embargo, estas técnicas pueden ser insuficientes para capturar relaciones no lineales o complejas entre múltiples variables clínicas. En este contexto, los algoritmos de machine learning (ML) ofrecen nuevas posibilidades para la predicción individualizada del riesgo de infección por MRA. Herramientas como el random forest han demostrado un rendimiento superior al de los modelos clásicos estadísticos en tareas similares, como la predicción de candidemia en pacientes hospitalizados, proporcionando mayor sensibilidad y especificidad, así como una mejor calibración [11]. Más recientemente, Hu et al. propusieron un modelo predictivo temprano del riesgo de bacteriemia, mostrando resultados prometedores y reforzando la utilidad del ML para anticiparse a este tipo de infecciones en contextos clínicos diversos [12].

El presente proyecto propone el desarrollo de diferentes modelos predictivos basado en ML con el objetivo de identificar precozmente a los pacientes hospitalizados con alto riesgo de desarrollar infecciones por los MRA más prevalentes en el medio hospitalario.

MATERIAL Y MÉTODOS

Diseño del estudio

Se trata de un estudio retrospectivo que incluyó a todos los pacientes adultos consecutivos con IIA con hemocultivos positivos que ingresaron en un hospital terciario de 750 camas en Barcelona entre enero de 2007 y julio de 2019. El estudio fue aprobado por el comité de ética de la institución (código HCB/2025/0739).

Conjunto de datos y análisis

Para este estudio, se creó una base de datos en formato xlsx desde el sistema operativo SAP que integró 269 variables (numéricas, categóricas y fechas), organizadas en bloques temáticos, abarcando características clínicas, microbiológicas, analíticas, de ingreso y procedimientos previos. Del paciente se registraron edad, sexo, comorbilidades específicas (ej. cirrosis, enfermedad renal crónica,

neoplasias) y el pronóstico de la enfermedad de base según McCabe [13] (no fatal, fatal a medio o rápidamente fatal). Respecto al ingreso, se recogieron días de hospitalización, procedencia y si la infección ocurrió en UCI. En relación con la infección, se evaluaron tipo de foco IIA, origen (comunitaria, nosocomial o relacionada con la asistencia sanitaria), presencia de shock, CID, SDRA y si el microorganismo aislado fue un MRA, variable dependiente. Se incluyeron factores de riesgo como hospitalización previa, cirugía o manipulaciones invasivas recientes (ej. biliares), y exposición a antibióticos en los tres meses previos. También se registraron condiciones de inmunosupresión (granulocitopenia, corticoides). Se incorporaron variables analíticas (creatinina, leucocitos, PCR) y constantes vitales (temperatura, frecuencia cardíaca, presión arterial, fiebre, insuficiencia respiratoria). Finalmente, se consideró el uso de dispositivos invasivos (intubación orotraqueal, sondaje vesical, catéteres vasculares) y los días con catéter.

Definiciones

Se consideraron MRA a aquellos que no son habituales en pacientes sin factores de riesgo de resistencia como: enterobacterias resistentes a cefalosporinas de tercera generación o a carbapenémicos, *Pseudomonas aeruginosa* y otros bacilos Gram negativos no fermentadores, *Enterococcus faecium* y enterococos resistentes a glicopéptidos, *Staphylococcus aureus* resistente a meticilina y *Candida spp*. Las infecciones se clasificaron como de adquisición comunitaria, asociadas a la atención sanitaria y nosocomiales, según los criterios de Friedman et al. [14]. Dentro de foco de IIA se incluyeron: diverticulitis, apendicitis, colangitis y colecistitis, peritonitis secundaria, abscesos intraabdominales, hepáticos, esplénicos y pélvicos. Se consideraron antecedentes de riesgo la realización de cirugía u otros procedimientos invasivos, hospitalización previa o tratamiento antibiótico previo durante el mes anterior al inicio de la infección.

Limpieza y preprocesado de datos

Se eliminaron los números de identificación de paciente y de los hemocultivos y analíticas para preservar el anonimato. Durante la preparación del conjunto de datos, se eliminaron registros duplicados basándose en el identificador del hemocultivo, incluyendo solo el primer episodio por paciente. Tras esta depuración inicial, se obtuvieron 1.532 registros válidos. Todas las variables fueron revisadas para detectar valores atípicos. Los registros clínicos originales fueron consultados para evaluar su validez, eliminando los erróneos y conservando los coherentes clínicamente. Las variables con más del 40% de valores ausentes (ej., lactato, bilirrubina total) fueron excluidas para evitar sesgos por imputación. Para el resto, se aplicaron estrategias específicas: imputación por regresión lineal para numéricas con proporciones moderadas (ej., PCR), y por la moda para categóricas con menos del 20% de ausentes. Las variables numéricas fueron normalizadas y transformadas. Se calcularon derivados como edad o días de ingreso a partir de campos de fecha, los cuales fueron eliminados. También se excluyeron variables no disponibles al diagnóstico, para asegurar que la predicción se basara solo en información accesible en ese momento. Las variables binarias se codificaron numéricamente y las categóricas no binarias se transformaron mediante codificación one-hot. Se crearon variables derivadas binarias para representar la presencia o ausencia de condiciones relevantes, como el número de antibióticos administrados antes del hemocultivo, exposición previa a antibióticos y factores de riesgo clínico. La variable objetivo fue la presencia de un MRA (17% de la muestra: 263 casos positivos, 1.269 negativos), combinando diferentes tipos de resistencia. Tras el procesamiento, quedaron 62 variables (51 categóricas binarias y 11 numéricas).

Finalmente, el conjunto de datos preprocesado fue dividido en dos subconjuntos: entrenamiento (80%, 1.225 pacientes) y test (20%, 307 pacientes). La partición se realizó mediante estratificación por la variable objetivo MRA para conservar la proporción de casos resistentes. Dada la

desproporción de clases, se implementaron estrategias para manejar el desequilibrio, como la generación de 2.500 registros sintéticos a partir del conjunto de entrenamiento usando Synthetic Data Vault (SDV). Este conjunto sintético fue balanceado manualmente para alcanzar una proporción del 60% de casos MRA, reduciendo el desequilibrio y facilitando el entrenamiento.

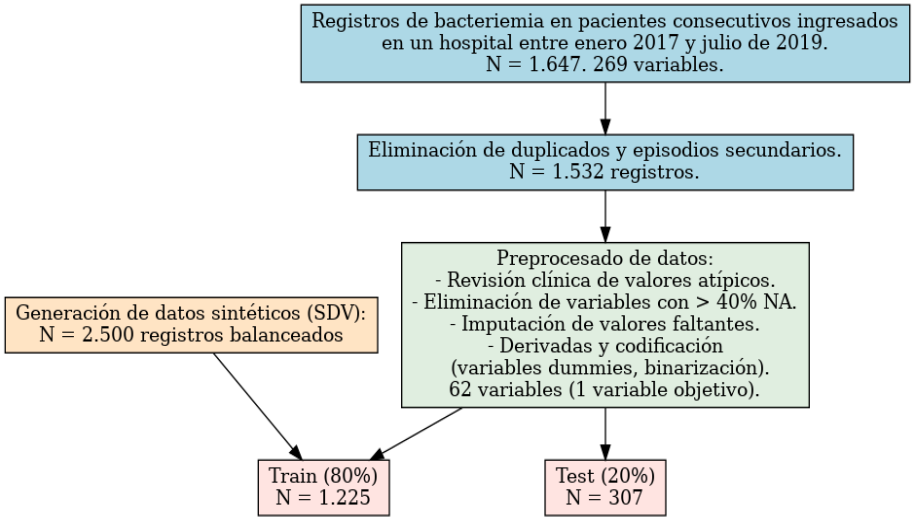


Figura 1. Diagrama de flujo del procesamiento de los datos clínicos utilizados para el desarrollo del modelo de predicción.

MODELO

En este estudio, se exploró el rendimiento de varios algoritmos de ML para la predicción de MRA. Se evaluaron modelos individuales como la regresión logística (RL), random forest (RF), XGBoost y CatBoost con el fin de establecer una línea base de rendimiento y comprender el comportamiento de cada uno de los enfoques. La RL es un modelo clásico muy interpretable en medicina, conocido por su utilidad para predecir enfermedades e identificar factores de riesgo [15,16]. Random forest es útil como herramienta de diagnóstico y pronóstico en situaciones en las que hay muchas variables, su importancia y las relaciones complejas entre ellas [11,12,17]. XGBoost y CatBoost son algoritmos de gradient boosting basados en árboles de decisión muy reconocidos a nivel médico por su precisión [16,18,19]. Adicionalmente, para potenciar la capacidad predictiva y la robustez del sistema, se diseñaron varios modelos de Stacking en los que se evaluaron varios clasificadores base en diferentes combinaciones. Para la evaluación de los modelos se emplearon las siguientes métricas: sensibilidad (recall o tasa de verdaderos positivos), especificidad, exactitud, precisión (valor predictivo positivo), F1-score, área bajo la curva ROC (AUC-ROC) y área bajo la curva de precisión-recall (AUC-PR). En la evaluación de los distintos modelos, se priorizó el recall para la clase minoritaria (MRA = 1) y el F1-score. Esta decisión se fundamentó en la relevancia clínica de detectar de forma temprana los casos de MRA, dado que los falsos negativos pueden tener consecuencias graves para el paciente y comprometer el tratamiento adecuado de la infección. La robustez de los modelos y la prevención del sobreajuste se abordaron mediante diversas estrategias implementadas durante el desarrollo y validación de cada algoritmo.

RESULTADOS

El estudio incluyó 1.532 episodios de IIA con hemocultivos positivos, de los cuales 263 (17,2%) correspondieron a infecciones por un MRA. La cohorte presentó una mediana de edad de 70 años (RIQ: 59 - 80) y la mayoría de los pacientes eran de sexo masculino (76,2%).

Modelos de clasificación individuales

Se exploraron diferentes arquitecturas de modelos individuales para evaluar su capacidad predictiva. Para RL se configuró class_weight = balanced para manejar el desbalance de clases y se estandarizaron las características. Se aplicó un análisis de componentes principales (PCA) que conservó los componentes que explicaban el 95% de la varianza total, y se utilizó una penalización L2 para prevenir el sobreajuste. Con el 100% de datos sintéticos, el modelo alcanzó un recall de 0,660, un F1-score de 0,438 y un AUC en test de 0,702. La diferencia con el AUC de entrenamiento (0,793) fue pequeña, lo que sugiere un overfitting moderado y buena capacidad de generalización. El AUC en validación cruzada fue de 0,767, lo que refuerza la consistencia del rendimiento. Sin embargo, su punto débil fue la baja precisión (PPV = 0,327), indicando un número elevado de falsos positivos, a pesar de que el valor predictivo negativo (NPV) alcanzó 0,910.

Para RF, se valoraron diversas configuraciones optimizadas con randomizedsearchCV, explorando el uso de diferentes porcentajes de datos sintéticos. A pesar de esta optimización, los modelos de RF mostraron una marcada tendencia al sobreajuste, con un AUC en entrenamiento muy alto (1,0) frente a un AUC en el conjunto de prueba significativamente más bajo (alrededor de 0,66-0,67) como se muestra en la Figura 2, lo que denota un sobreajuste severo. Crucialmente, el rendimiento en la clase minoritaria fue extremadamente bajo, con un recall para la clase MRA = 1 consistentemente bajo (entre 0 y 0,17) y un F1-score también muy bajo. La configuración sin datos sintéticos ni siquiera fue capaz de detectar la clase positiva. Estas limitaciones hicieron que el modelo RF no resultara adecuado para este estudio.

XGBoost se evaluó en diversas configuraciones. La estrategia clave fue el balance de clases mediante la incorporación de datos sintéticos, junto con la optimización de hiperparámetros para controlar la complejidad y reducir la varianza. Una versión más simple y robusta, con parámetros de regularización ajustados (learning_rate de 0,05 y max_depth de 2), demostró ser la más efectiva. La configuración con 40% de datos sintéticos presentó un AUC en test competitivo (0,701) con un AUC en entrenamiento aceptable (0,769), lo que sugiere un overfitting manejable. Su F1-score de 0,403 es sólido y su precisión en test alcanza 0,38, el valor más alto entre los modelos con buen rendimiento, lo que lo hace ventajoso si se prioriza minimizar los falsos positivos.

Para el modelo CatBoost se utilizó un pipeline que integraba el clasificador con una búsqueda aleatoria de hiperparámetros y class_weight = balanced. Se evaluaron tres proporciones de datos sintéticos (0%, 50% y 100%). La configuración con 0% de datos sintéticos fue la opción preferente por ofrecer el mejor equilibrio entre rendimiento y generalización, con un AUC en entrenamiento de 0,744 y un AUC en test de 0,711, lo que indica una mínima brecha de sobreajuste. Además, alcanzó un recall de 0,623 y un F1-score de 0,388 en la clase minoritaria. Aunque las configuraciones con 50% y 100% de datos sintéticos lograron un AUC en test ligeramente superior (hasta 0,723) y mejores valores de recall (hasta 0,755), el marcado aumento del AUC en entrenamiento (por encima de 0,97) sugiere un sobreajuste importante.

La Figura 3 compara la distribución de probabilidades predichas por los modelos RL, RF, XGBoost y CatBoost. XGBoost y CatBoost muestran una mayor separación entre las clases, lo que sugiere mejor capacidad discriminativa en la clasificación de MRA. La Figura 4 muestra las diez variables más relevantes según permutation importance para XGBoost y CatBoost. En ambos, el número de antibióticos previos fue el principal predictor. Cabe mencionar que no se aplicó PCA en los modelos basados en árboles (RF, XGBoost y CatBoost), ya que son robustos ante variables correlacionadas y no requieren escalado previo.

Figura 2. Comparación del desempeño de los modelos base (Regresión Logística, Random Forest, XGBoost y CatBoost) para la detección de MRA: curvas ROC, precision-recall y matrices de confusión

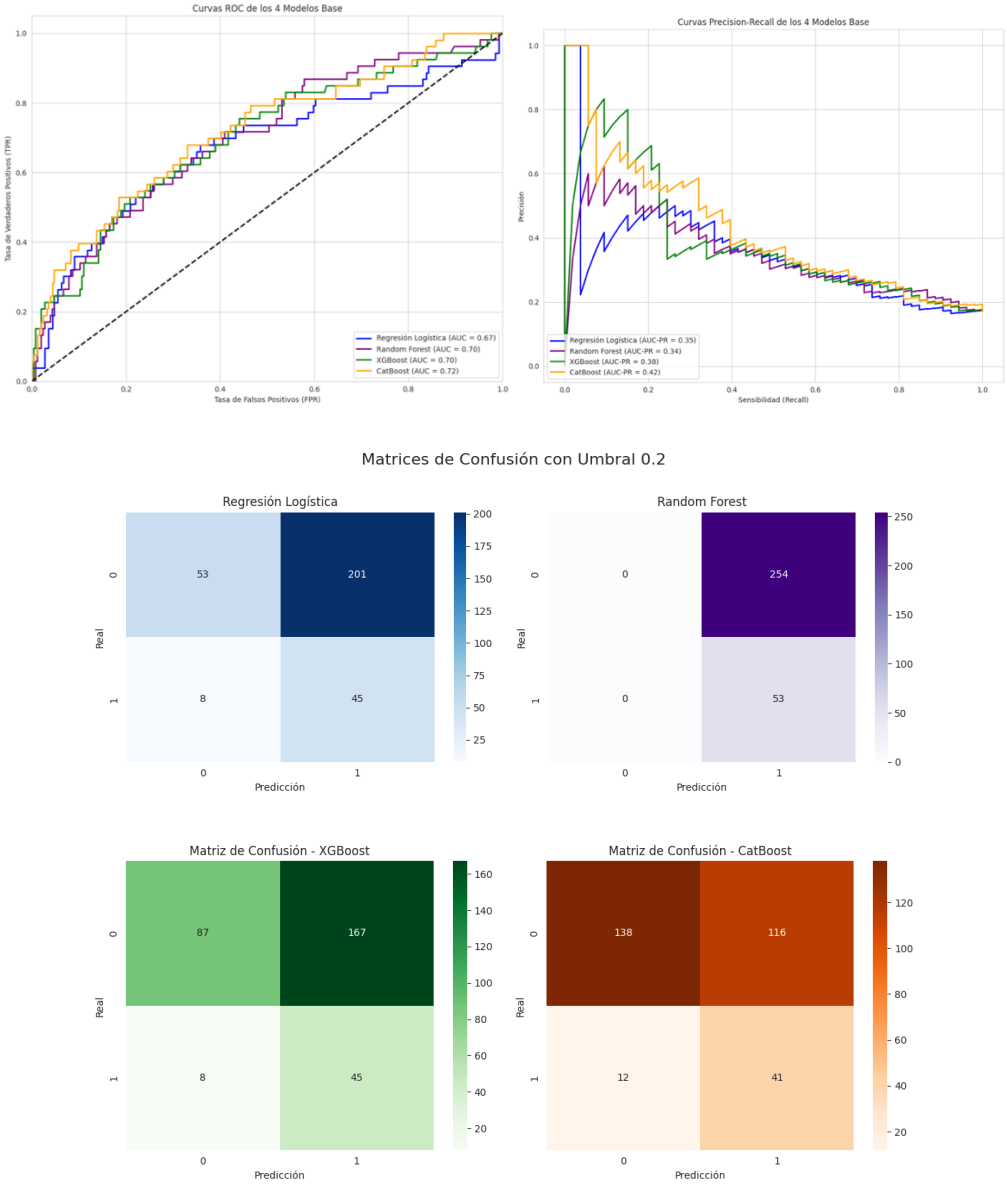
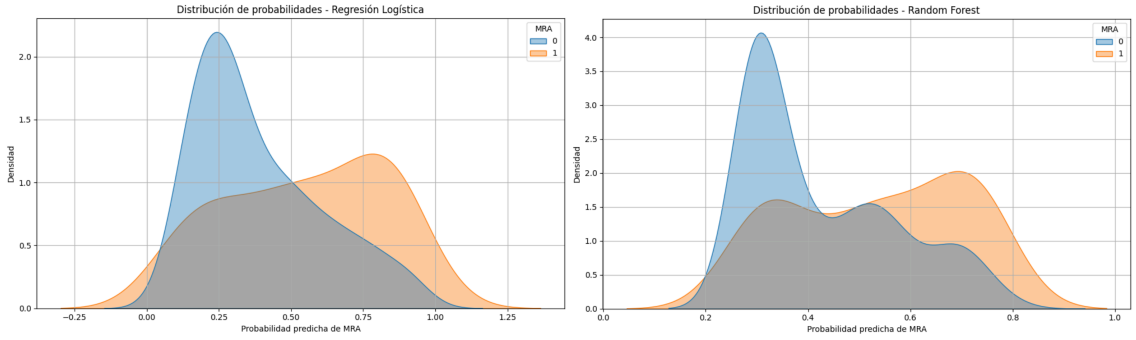


Figura 3. Comparación de la distribución de probabilidades predichas por los modelos RL, RF, XGBoost y CatBoost en la clasificación de MRA.



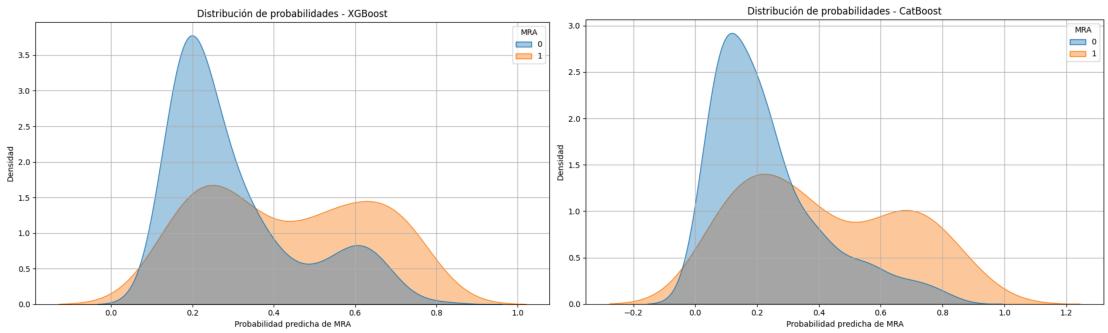
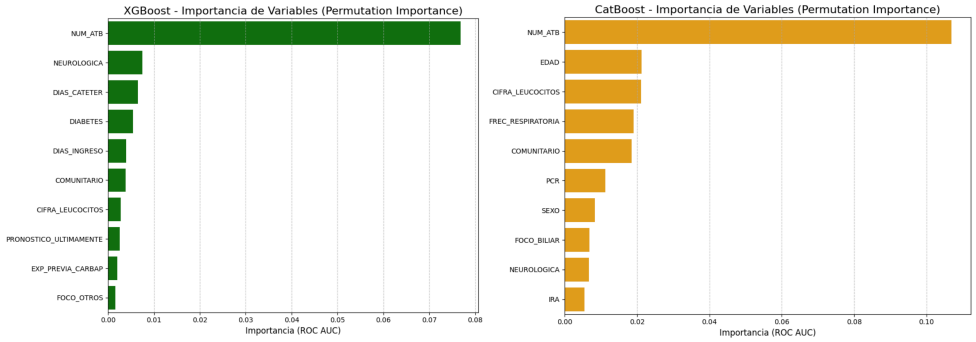


Figura 4. Importancia de las variables en los modelos predictivos de MRA según Permutation Importance (basado en AUC ROC).
*Se presentan las 10 variables más importantes en cada uno de los modelos (XGBoost, CatBoost), calculadas mediante el método de Permutation Importance en el conjunto de test.
**La métrica empleada para cuantificar la pérdida de rendimiento fue el AUC ROC; las barras reflejan la contribución individual de cada variable al rendimiento del modelo.
***Se excluyeron RL por realizar análisis con PCA y RF por su menor rendimiento.



Modelos de stacking classifier

Se evaluaron dos configuraciones de modelos de Stacking para la detección de MRA. El Stacking 3c, diseñado con tres clasificadores base (RL, RF y XGBoost), se entrenó con el 100% de datos sintéticos. Al aplicar un umbral ajustado a 0,20, obtuvo un alto recall (0,774), aunque con una precisión muy baja (0,228) y un F1-score de 0,352. Su AUC en test fue de 0,705, mostrando un sobreajuste moderado. En el análisis de importancia de variables del modelo 3c (Figura 5), se observa que el número de antibióticos previos fue la variable más influyente, seguida por edad y días de ingreso.

Por otro lado, el Stacking 4c buscó un equilibrio más optimizado, añadiendo un cuarto clasificador base (CatBoost) y usando un 40% de datos sintéticos. Aunque su recall fue menor (0,510), superó al modelo 3c en precisión (0,355) y F1-score (0,419), con un mejor AUC de 0,714. Su principal debilidad fue un sobreajuste mucho más acusado (AUC en entrenamiento 0,998). En su análisis de importancia de variables, el modelo 4c mostró una contribución más repartida entre otras variables como edad, frecuencia respiratoria y comorbilidades (Figura 5), lo que indica que sus decisiones se basan en un conjunto clínico más diverso.

La matriz de confusión muestra que el modelo Stacking 3c prioriza la sensibilidad, identificando más casos positivos (41 frente a 27), aunque con un mayor número de falsos positivos (139 frente a 49). En contraste, el modelo Stacking 4c presenta un enfoque más equilibrado, con menos falsos positivos y mayor precisión. La curva ROC (Figura 6) refleja una ligera ventaja del modelo 4c (AUC = 0,71 vs 0,70), lo que indica una capacidad discriminativa algo superior. No obstante, la curva de precisión-

recall revela un área casi comparable en ambos modelos, evidenciando las dificultades propias de trabajar con clases desbalanceadas. Finalmente, el modelo 3c asigna puntuaciones más altas a los casos positivos (MRA = 1), lo que también se traduce en una mejor separación entre clases (Figura 7).

Figura 5. Importancia de las variables en los modelos predictivos de MRA según Permutation Importance (basado en AUC ROC).
*Se presentan las 10 variables más importantes en cada uno de los modelos (Stacking 3c y Stacking 4c), calculadas mediante el método de Permutation Importance en el conjunto de test.
**La métrica empleada para cuantificar la pérdida de rendimiento fue el AUC ROC; las barras reflejan la contribución individual de cada variable al rendimiento del modelo.

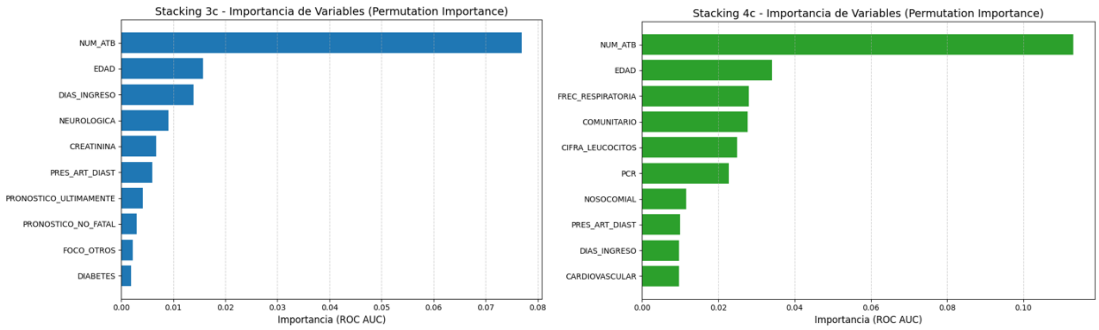


Figura 6. Comparación del desempeño de los modelos base (stacking 3c y stacking 4c) para la detección de MRA: curvas ROC, precision-recall y matrices de confusión

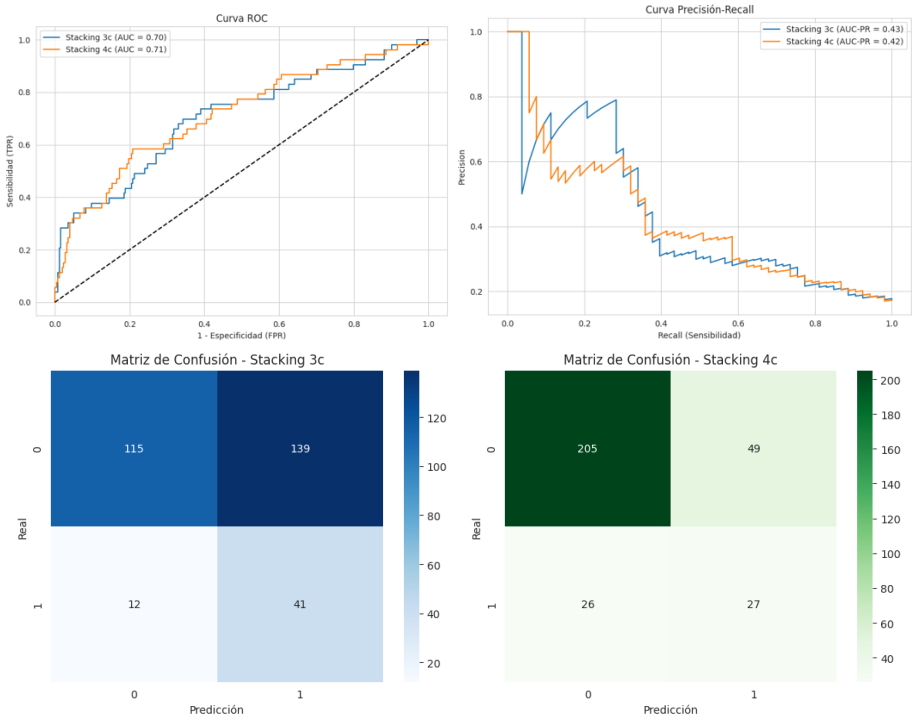


Figura 7. Comparación de la distribución de probabilidades predichas por los modelos Stacking 3c y 4c en la clasificación de MRA.

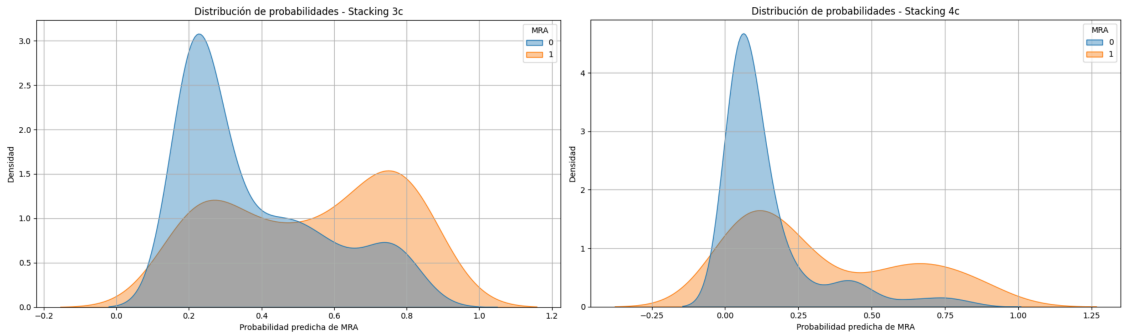


Tabla 1. Comparativa de métricas de modelos para la detección de MRA.

Modelo	% Datos Sintéticos	AUC CV	AUC Train	AUC Test	F1 Train	F1 Test	Recall Test	PPV Test	NPV Test
RL	100	0,767	0,793	0,702	0,716	0,438	0,660	0,327	0,910
RF	0	0,720	0,888	0,673	0,182	0,000	0,000	0,000	0,827
XGBoost	40	0,761	0,769	0,701	0,545	0,404	0,434	0,378	0,878
CatBoost	0	0,719	0,744	0,711	0,432	0,388	0,623	0,282	0,895
CatBoost	50	0,927	0,975	0,723	0,821	0,396	0,736	0,271	0,914
Stacking 3c	100	0,827	0,877	0,705	0,693	0,352	0,774	0,228	0,884
Stacking 4c	40	0,741	0,852	0,714	0,672	0,419	0,510	0,356	0,887

AUC: área bajo la curva ROC; CV: validación cruzada; F1: media armónica entre precisión y sensibilidad; PPV: valor predictivo positivo (precisión); NPV: valor predictivo negativo; RL: regresión logística; RF: random forest.
* Las métricas de *train* se calcularon sobre el conjunto de entrenamiento y las de *test* aplicando el umbral optimizado para sensibilidad (recall).

DISCUSIÓN

La creciente amenaza de microorganismos resistentes a los antibióticos (MRA) complica seriamente la elección del tratamiento empírico en las infecciones intraabdominales (IIA). En este escenario, el uso de modelos predictivos de aprendizaje automático (ML) se presenta como una herramienta prometedora, ya que podrían identificar de manera temprana a los pacientes con alto riesgo de infecciones por MRA, fundamental para tomar decisiones terapéuticas adecuadas. No obstante, el desarrollo de estos modelos presenta limitaciones vinculadas a las características de los datos empleados para su entrenamiento, siendo especialmente relevante en este estudio el tamaño de la cohorte analizada, compuesta por 1.532 episodios. En general, los estudios que utilizan ML suelen beneficiarse de conjuntos de datos mucho más grandes. Un mayor volumen de datos permite a los modelos aprender patrones más complejos y robustos, mejorando su capacidad de generalización y su aplicabilidad. En investigaciones anteriores sobre la predicción de bacteriemia, se han utilizado bases de datos significativamente más grandes, algunas superando los 50.000 pacientes [20-21]. No obstante, es relevante señalar que también se han publicado modelos eficaces con tamaños de muestra comparables. Por ejemplo, Pan et al. desarrollaron un modelo con un rendimiento predictivo robusto (AUC de 0,865) utilizando 1.385 pacientes para predecir el shock séptico producido por *Klebsiella pneumoniae* multirresistente [22]. Esto indica que cohortes de este tamaño, como la utilizada en este estudio, pueden ser suficientes para entrenar modelos predictivos eficaces, siempre que el diseño sea adecuado y el problema esté bien definido.

Otra condición evidente en el presente estudio es el desequilibrio de clases. La prevalencia de infecciones por MRA es del 17,2%, lo que representa una clase minoritaria significativa. Este desequilibrio puede sesgar los algoritmos de ML hacia la clase mayoritaria (no-MRA), resultando en modelos con alta precisión general, pero un bajo rendimiento en la identificación de la clase minoritaria, que es la de mayor interés clínico. Para mitigar este problema se emplearon varios enfoques. Se asignaron pesos de clase (class_weight = balanced), lo que proporciona al modelo más ejemplos de la clase minoritaria durante el entrenamiento, mejorando así su capacidad para identificar correctamente los casos de MRA sin comprometer excesivamente la generalización. Otra estrategia utilizada fue el aumento de datos mediante la generación de datos sintéticos con la librería Synthetic Data Vault (SDV). Esta técnica permitió crear ejemplos sintéticos realistas de la clase minoritaria sin replicar las observaciones existentes. El aumento de datos es una estrategia empleada en estudios previos. Gupta et al. aplicaron SMOTE (una estrategia de sobremuestreo sintético basada en la interpolación de vecinos cercanos) para mejorar la detección de infecciones urinarias multirresistentes, lo que resultó en un aumento de la sensibilidad de los modelos [23]. En otro estudio, el uso de SMOTE no mejoró el rendimiento de los modelos para predecir bacteriemia, por lo que los autores optaron por la ponderación de clases y el MCC como técnicas para mitigar el desbalance [24]. Estas diferencias metodológicas demuestran que las estrategias deben adaptarse a las características específicas de cada base de datos.

La evaluación de modelos individuales fue un paso esencial para entender el comportamiento de distintos algoritmos frente a la complejidad de los datos y el objetivo de la predicción de MRA. Para seleccionar los mejores modelos, es crucial considerar un equilibrio entre el test_auc (la capacidad de discriminación general), las métricas de la clase minoritaria (test_f1, test_recall, test_precision) y la presencia de overfitting (la diferencia entre train_auc y test_auc). En este sentido, el modelo CatBoost sin datos sintéticos destaca por tener la menor brecha de overfitting (train_auc = 0,744) de todos los modelos individuales, a pesar de no alcanzar el test_auc más alto (0,711). Presenta un recall elevado (0,625) y una precisión baja, pero aceptable, en función de su buena generalización. Por otro lado, la versión de CatBoost con un 50% de datos sintéticos logra un test_auc aún mayor (hasta 0,723) y un recall superior (0,736), aunque con un overfitting más marcado (train_auc cercano a 0,97). Esta última opción es ideal si se prioriza la sensibilidad y la capacidad discriminativa, incluso a costa de una menor estabilidad. El modelo XGBoost con un 40% de datos sintéticos también consiguió un equilibrio sólido, con un test_auc de 0,701, un f1-score de 0,404 y la mayor precisión (0,38) entre los modelos evaluados. Con un overfitting moderado (train_auc = 0,769), podría ser una alternativa robusta cuando las predicciones positivas son la prioridad.

En comparación con otros estudios, los modelos individuales, con una AUC máxima de 0,70, no alcanzaron los mejores resultados en la predicción de resistencia antimicrobiana. Por ejemplo, Garcia-Vidal et al. [25] en su estudio sobre infecciones por bacilos gramnegativos multirresistentes, reportaron AUCs para modelos de RF, Gradient Boosting Machine y XGBoost que oscilaron entre 0,78 y 0,79. Zhao et al. [17] identificaron a RF como el modelo de mejor rendimiento (AUC = 0,83) para la predicción de infecciones por microorganismos resistentes en pacientes de UCI. En el presente estudio, RF fue el modelo con peor rendimiento debido a un sobreajuste marcado y a su incapacidad para detectar casos de MRA en el conjunto de prueba. Esta baja eficacia podría explicarse por la alta sensibilidad del algoritmo a datos desbalanceados y a la ausencia de mecanismos de regularización eficaces, lo que lo hizo menos adecuado que otros modelos como XGBoost o CatBoost en este contexto.

Los modelos de stacking ofrecieron una mayor robustez en la predicción de MRA. Aunque algoritmos individuales como CatBoost con un 50% de datos sintéticos alcanzaron un AUC test ligeramente superior (hasta 0,723), aunque con un grado elevado de sobreajuste. Comparando el

rendimiento global, el modelo Stacking 3c obtuvo un AUC de 0,705, frente a los 0,714 del modelo 4c. Sin embargo, esta ventaja en discriminación no se tradujo en una mejor sensibilidad: el Stacking 3c detectó el 77,4% de los casos positivos, mientras que el 4c solo detectó el 50,9%. Aunque el modelo 4c presentó una mayor precisión (35,5% vs 22,8%), en un contexto clínico donde es prioritario no pasar por alto casos de MRA, el modelo 3c resulta preferible por su mayor recall, alcanzado mediante un umbral ajustado a 0,2. Desde un punto de vista clínico, ante la sospecha de MRA, es más conveniente empezar con una cobertura antibiótica amplia y desescalar posteriormente, minimizando así el impacto de un posible falso negativo. Las curvas ROC y precision-recall generadas para ambos modelos respaldan esta elección. En particular, la curva precision-recall, que es más sensible al comportamiento ante clases desequilibradas como MRA, permite ajustar el umbral de decisión según los objetivos clínicos, priorizando la sensibilidad cuando sea necesario (Figura 6). Este enfoque de ensemble está cobrando relevancia en la investigación médica para abordar problemas complejos de predicción, incluso fuera del ámbito de las infecciones, como en oncología y neurología [26, 27], con AUC superiores a 0,80. Aunque los resultados del presente estudio fueron más modestos, debe tenerse en cuenta el desequilibrio de clases y la complejidad clínica del problema.

El análisis de importancia de variables por permutación (Figuras 2 y 5) identificó el número de antibióticos administrados previamente como una de las variables con mayor peso en todos los modelos de predicción de infecciones por MDR. Este hallazgo es consistente con los resultados de estudios previos, que demostraron que el uso previo de antibióticos es un factor independiente, tanto en el análisis multivariado [10] como en los modelos de ML aplicados [25]. Según el modelo, otras variables como la edad y los días de ingreso también resultaron relevantes, lo que apoya clínicamente las decisiones tomadas por el sistema [29, 30]. Por otro lado, la asociación observada entre la frecuencia respiratoria y MRA no tiene una base fisiopatológica clara ni respaldo en la literatura científica y probablemente refleja una correlación indirecta con la gravedad del paciente u otras variables no incluidas, por lo que este resultado debe interpretarse con cautela.

Este estudio presenta varias limitaciones que deben tenerse en cuenta al interpretar sus resultados. En primer lugar, el número de casos correspondientes a la clase de interés (MRA) fue relativamente bajo. Este desequilibrio puede haber limitado la capacidad de algunos modelos para aprender patrones robustos y generalizables. Además, la utilidad de los datos sintéticos depende de qué tan bien el generador reproduce las características reales del conjunto original. Por otra parte, al tratarse de un estudio retrospectivo con datos de historias clínicas, puede haber variaciones en la calidad y la integridad de los datos recogidos. Los datos provienen de un único centro hospitalario, lo que podría afectar la generalización del modelo a otros entornos clínicos con distinta prevalencia de resistencias o prácticas asistenciales. Otra limitación es la presencia de datos faltantes en algunas variables. Aunque se aplicaron estrategias de imputación clínicamente razonables (como regresión para variables numéricas o moda para categóricas), esta aproximación puede introducir un sesgo o reducir la precisión del modelo. Asimismo, se descartaron variables con un porcentaje elevado de valores ausentes, limitando así el número de variables predictoras posibles. Por último, aunque se emplearon técnicas como permutation importance para interpretar los modelos, algunos algoritmos, como stacking o boosting, tienen limitaciones en términos de interpretabilidad, lo que podría dificultar su adopción en la práctica clínica diaria.

En conclusión, el modelo Stacking 3c, que se basa en las predicciones de tres modelos individuales (regresión logística, random forest y XGBoost), demostró ser la herramienta más aplicable en nuestro contexto clínico. Este modelo prioriza la identificación de casos de MRA (con un recall del 77%), lo que contribuye a evitar retrasos en el tratamiento de infecciones resistentes y a mejorar la adecuación terapéutica empírica. Aunque su AUC o F1-score difieran de los de la literatura, su

principal valor radica en su aplicabilidad práctica, siempre y cuando el número de falsos positivos se pueda manejar clínicamente. Para futuras mejoras y una mayor generalización, será clave aumentar el tamaño del conjunto de datos y la representación de la clase minoritaria, así como explorar arquitecturas más avanzadas.

REFERENCIAS

1. Krobot K, Yin D, Zhang Q, Sen S, Altendorf-Hofmann A, Scheele J, et al. Effect of inappropriate initial empiric antibiotic therapy on outcome of patients with community-acquired intra-abdominal infections requiring surgery. *Eur J Clin Microbiol Infect Dis*. 2004;23:682-7.
2. Sartelli M. A focus on intra-abdominal infections. *World J Emerg Surg*. 2010;5:9.
3. Labricciosa FM, Sartelli M, Abbo LM, et al. Epidemiology and risk factors for isolation of multi-drug-resistant organisms in patients with complicated intra-abdominal infections. *Surg Infect (Larchmt)*. 2018;19(3):264-72.
4. Seguin P, Laviolle B, Chanavaz C, Donnio PY, Gautier-Lerestif AL, Campion JP, et al. Factors associated with multidrug-resistant bacteria in secondary peritonitis: impact on antibiotic therapy. *Clin Microbiol Infect*. 2006;12:980-5.
5. Paul M, Shani V, Muchtar E, Kariv G, Robenshtok E, Leibovici L. Incidence and outcome of inappropriate in-hospital empiric antibiotics for severe infection: a systematic review and meta-analysis. *Crit Care*. 2015;19:63.
6. Seguin P, Fédun Y, Laviolle B, Nesseler N, Donnio PY, Mallédant Y. Risk factors for multidrug-resistant bacteria in patients with post-operative peritonitis requiring intensive care. *J Antimicrob Chemother*. 2010;65:342-6.
7. Ortega M, Marco F, Soriano A, Almela M, Martínez JA, López J, et al. Epidemiology and prognosis determinants of bacteraemic biliary tract infection. *J Antimicrob Chemother*. 2012;67:1508-13.
8. Reuken PA, Torres D, Baier M, Löffler B, Lübbert C, Lippmann N, et al. Risk factors for multi-drug resistant pathogens and failure of empiric first-line therapy in acute cholangitis. *PLoS One*. 2017;12:e0169900.
9. Montravers P, Dufour G, Guglielminotti J, Desmard M, Muller C, Houissa H, et al. Dynamic changes of microbial flora and therapeutic consequences in persistent peritonitis. *Crit Care*. 2015;19:70.
10. Rodríguez Núñez O, Agüero DL, Morata L, Puerta Alcalde P, Cardozo C, Rico V, et al. Antibiotic-resistant microorganisms in patients with bloodstream infection of intra-abdominal origin: risk factors and impact on mortality. *Infection*. 2021;49(4):693-702.
11. Ripoli A, Sozio E, Sbrana F, et al. Personalized machine learning approach to predict candidemia in medical wards. *Infection*. 2020;48(5):641-51.
12. Hu X, Zhi S, Li Y, et al. Development and application of an early prediction model for risk of bloodstream infection based on real-world study. *BMC Med Inform Decis Mak*. 2025;25(1):186.
13. McCabe WR, Jackson GG. Classification of acute bacterial infections by prognostic criteria. *Arch Intern Med*. 1962;110:856-64.
14. Friedman ND, Kaye KS, Stout JE, McGarry SA, Trivette SL, Briggs JP, et al. Health care–associated bloodstream infections in adults: a reason to change the accepted definition of community-acquired infections. *Ann Intern Med*. 2002;137:791-7.
15. Murri R, De Angelis G, Antenucci L, Fiori B, Rinaldi R, Fantoni M, et al. A machine learning predictive model of bloodstream infection in hospitalized patients. *Diagnostics (Basel)*. 2024;14(4):445.

16. Liu Y, Zhang X, Chen Y, et al. Development, comparison, and validation of four intelligent, practical machine learning models for predicting bloodstream infections. *BMC Infect Dis*. 2023;23:345.
17. Zhao X, Wang Y, Li J, et al. Machine learning-based prediction model for multidrug-resistant organism infections in hospitalized patients. *Front Public Health*. 2025;13:1150023.
18. Hu H, Liu Y, Wang R, et al. Interpretable machine learning for early prediction of prognosis in sepsis. *J Transl Med*. 2022;20:320.
19. Duan M, Shu T, Zhao B, Xiang T, Wang J, Huang H, Zhang Y, et al. Explainable machine learning models for predicting 30-day readmission in pediatric pulmonary hypertension: a multicenter, retrospective study. *Front Cardiovasc Med*. 2022;9:919224.
20. Zoabi Y, Kehat O, Lahav D, Weiss-Meilik A, Adler A, Shomron N. Predicting bloodstream infection outcome using machine learning. *Sci Rep*. 2021;11:22725.
21. Zargari Marandi R, Andreasen SH, Nielsen MJ, Iversen K. Prediction of bloodstream infection using machine learning based primarily on biochemical data. *Sci Rep*. 2025;15:6841.
22. Pan S, Shi T, Ji J, Wang K, Jiang K, Yu Y, et al. Developing and validating a machine learning model to predict multidrug-resistant Klebsiella pneumoniae-related septic shock. *Front Immunol*. 2025;15:e1539465.
23. Gupta R, Wang W, Dee EC, Yang E, Bhambhvani HP, Chang SL, et al. Machine learning models predicting multidrug resistant urinary tract infections using "DsaaS". Sci Rep. 2022;12(1):20363
24. Marandi RZ, Hertz FB, Thomassen JQ, Rasmussen SC, Frikke-Schmidt R, Frimodt-Møller N, et al. Prediction of bloodstream infection using machine learning based primarily on biochemical data. Sci Rep. 2025;15:17478.
25. Garcia Vidal C, Puerta Alcalde P, Cardozo C, Orellana MA, Besanson G, Lagunas J, et al. Machine learning to assess the risk of multidrug resistant Gram negative bacilli infections in febrile neutropenic hematological patients. *Infect Dis Ther*. 2021;10(2):971-83.
26. Hao L, Zhang J, Di Y, Qi Z, Zhang P. Predicting a failure of postoperative thromboprophylaxis in non-small cell lung cancer: a stacking machine learning approach. *PLoS One*. 2025;20(4):e0320674.
27. Luo W, Li Q, Wang H, et al. Accelerated functional brain aging in major depressive disorder: a machine learning approach. *Neuroimage Clin*. 2022;35:103133.
28. Zhao Y, Zhang H, Yin J, Li W, Wang J, Guo X, et al. Risk factors and predictive model for multidrug-resistant organism infection in patients with open injuries: a retrospective case-control study. Sci Rep. 2023;13:1459.
29. Chen Y-P, Tasi X-W, Chang K, Cao X-D, Chen J-R, Liao C-S, et al. Multi-Drug Resistant Organisms Infection Impact on Patients Length of Stay in Respiratory Care Ward. Antibiotics (Basel). 2021;10(5):608.