

PREDICCIÓN DEL EFECTO INÓCULO A CEFAZOLINA EN *Staphylococcus aureus* SUSCEPTIBLE A METICILINA POR UN METODO DE APRENDIZAJE AUTOMÁTICO

Autores: Reyes-Manrique Lynda Jehny¹, Bermúdez-Munar José Alejandro¹, Martín-López Zaidy Ocnary¹, Quiroga-Calderón César Hobany¹, Matiz-González Juan Manuel², Carvajal-Ortiz Lina Paola², Duitama Leal Alejandro³, Reyes Jinnethe²

1. Maestría en Estadística Aplicada y Ciencia de Datos, Facultad de Ciencias, Universidad El Bosque.
2. Unidad de Genética y Resistencia Antimicrobiana, Universidad El Bosque.
3. Grupo Signos. Departamento de Matemáticas, Universidad El Bosque.

Directores de Tesis:

1. Alejandro Duitama, Grupo Signos. Departamento de Matemáticas, Universidad El Bosque. duitamaalejandro@unbosque.edu.co
2. Jinnethe Reyes, Unidad de Genética y Resistencia Antimicrobiana, Universidad El Bosque. reyesjinnethe@unbosque.edu.co

RESUMEN

Introducción. La resistencia a antibióticos constituye un desafío de importancia clínica, no solo en términos de tratamiento biológico y terapéutico de las infecciones, sino también debido a su impacto en la salud pública (1). El *Staphylococcus aureus*, es un agente bacteriano común en el microbioma humano. Sin embargo, también ocasiona gran variedad de entidades infecciosas, incluyendo, bacteriemia, endocarditis, así como infecciones osteoarticulares, cutáneas, de tejidos blandos, pleuropulmonares y relacionadas con dispositivos (2). La incidencia de bacteriemia por *Staphylococcus aureus* (SAB) en Estados Unidos oscila entre 20 y 50 casos por cada 100.000 habitantes al año, con una tasa de mortalidad entre el 10% y el 30%, superando en número de muertes combinadas al VIH/SIDA, la tuberculosis y la hepatitis viral, lo que representa un considerable costo en términos de salud pública (3,4). La Sociedad Americana de Enfermedades Infecciosas (IDSA) recomienda los antibióticos betalactámicos como tratamiento fundamental para infecciones causadas por *Staphylococcus aureus* susceptible a meticilina (SASM) (5,6). La cefazolina se ha convertido en una excelente alternativa de tratamiento por sus bajos efectos adversos y su costo (6). Sin embargo, ha surgido un fenómeno de resistencia conocido como el efecto inóculo a cefazolina (CzIE), asociado a la producción de la betalactamasa (BlaZ) (7), lo que plantea la necesidad de explorar alternativas terapéuticas. El uso de técnicas de aprendizaje automático (Machine Learning - ML) se presenta como una vía prometedora para evaluar la capacidad predictiva de modelos en este contexto, lo que podría tener implicaciones significativas en la práctica médica, permitiendo el uso adecuado de la cefazolina y por ende optimizando la toma de decisiones para el tratamiento antibiótico.

Objetivo. El propósito de este trabajo fue implementar diferentes modelos de clasificación de aprendizaje automático para encontrar aquel con el mejor desempeño, en métricas estándar como exactitud, precisión, sensibilidad, especificidad, puntuación F1 (f1-Score) y el área bajo la curva (AUC); que permita predecir el efecto inóculo a Cefazolina en SASM, basado en la secuencia de nucleótidos del operón *Bla* y sus componentes reguladores (*BlaZRI*).

Materiales y Métodos. Se tomaron dos grupos de aislamientos de SASM, usando la secuenciación del genoma completo donde se analizaron las secuencias de nucleótidos del operón *blaZRI* en 410 cepas de SASM recuperadas de sangre, infecciones de piel y tejidos blandos y neumonía en adultos para su entrenamiento, prueba y validación, en general este grupo presentó una prevalencia de CzIE del 49.6%. El preprocesamiento se realizó por medio de la metodología de K-meros, analizándose varios tamaños (5,7,12,17 y 23 meros), obteniendo las mejores métricas con 23 meros. Se incluyeron un total de 410 secuencias de genomas de SASM, donde en la depuración se eliminaron 24 secuencias por presentar delección en el operón. Se realizó el entrenamiento de los modelos utilizando el 60% de los datos (n=231), el conjunto de prueba con el 28% (n=108) y como validación 12% (n=47). Se exploraron diferentes métodos de clasificación de aprendizaje automático, entre los que están: Light Gradient Boosting Machine (lightGBM), K Neighbors Classifier (knn) y Ridge Classifier (ridge) siendo estos los que mostraron mejor desempeño. Se realizó un afinamiento o tuning, con el objetivo de estimar los mejores hiperparámetros que permitieran tener unas métricas de desempeño con un error de aprendizaje mínimo.

Resultados. La predicción con mayor exactitud para el CzIE fue del modelo lightgbm en 72% para el entrenamiento, 76% para la prueba y 80% para la validación. La precisión de 0.74 significa que el 74% de las predicciones de CzIE son correctas. Lo cual puede ser considerado aceptable, pero con opciones de mejora. Una sensibilidad (Recall) del 90% sugiere que el modelo es efectivo para detectar en esa medida los casos positivos, F1-Score del 0.81 indica un buen rendimiento del modelo y una especificidad del 0.72% refiere una buena identificación de verdaderos negativos, crucial para toma de decisiones médicas. En conclusión, el modelo lightgbm, se recomienda como el modelo de aprendizaje automático óptimo, desde la validación de métricas de evaluación para la predicción del efecto inóculo a cefazolina en SASM.

Conclusión. Este modelo puede ser considerado como una alternativa diagnóstica, a partir de herramientas genómicas como la secuenciación, la cual, en la actualidad, resulta ser una opción promisoriosa y accesible; siendo el inicio de un gran avance, con rapidez y eficiencia en ambientes clínicos, cuyo objetivo redundaría en el mejoramiento de la calidad de vida de los pacientes de nuestro país.

ABSTRACT

Background. Antibiotic resistance constitutes a challenge of clinical importance, not only in terms of biological and therapeutic treatment of infections, but also due to its

impact on public health (1). *Staphylococcus aureus* is common in the human microbiome. However, it also causes a wide variety of infections, including bacteremia, infective endocarditis, as well as osteoarticular, cutaneous, soft tissue, pleuropulmonary, and device-related infections (2). The incidence in U.S of *Staphylococcus aureus* bacteremia (SAB) ranges between 20 and 50 cases per 100,000 inhabitants per year, with a mortality rate between 10% and 30%, surpassing HIV/AIDS, tuberculosis in the number of combined deaths and viral hepatitis, which represents a considerable cost in terms of public health (3,4). The Infectious Diseases Society of America (IDSA) recommends beta-lactam antibiotics as the primary treatment for infections caused by methicillin-susceptible *Staphylococcus aureus* (MSSA) (5,6). Cefazolin has become an excellent treatment alternative due to its low adverse effects and cost (6). However, a resistance phenomenon known as the cefazolin inoculum effect (CzIE) has emerged, associated with the production of beta-lactamase (BlaZ) (7), which raises the need to explore therapeutic alternatives. The use of machine learning techniques (ML) is presented as a promising way to evaluate the predictive capacity of models in this context, which could have significant implications in medical practice, allowing the appropriate use of cefazolin and therefore optimizing decision making for antibiotic treatment.

Aim. To implement different machine learning classification models to find among the different models the one that performs best in standard metrics such as accuracy, precision, sensitivity, specificity, F1-Score and area under the curve (AUC), which allows predicting the inoculum effect of Cefazolin in MSSA, based on the nucleotide sequence of the Bla operon and its regulatory components (BlaZRI).

Materials and Methods. Two groups of MSSA isolates were taken, using whole genome sequencing where nucleotide sequences of the *blaZRI* operon were analyzed in 410 MSSA strains recovered from blood, skin and soft tissue infections and pneumonia in adults for training, testing and validation, in general this group presented a prevalence of CzIE of 49.6%. Preprocessing was carried out using the K-mers methodology, analyzing various sizes (5,7,12,17 and 23 mers), obtaining the best metrics with 23 mers and 24 sequences were eliminated due to deletion. The models were trained using 60% of the sequences (n=231), the test set was 28% (n=108), as validation and 12% (n=47). Different machine learning classification methods were explored, including Light Gradient Boosting Machine (lightgbm), K Neighbors Classifier (knn) and Ridge Classifier (ridge), these being the ones that showed the best performance, a refinement or tuning was performed, with the objective of estimating the best hyperparameters that allow having performance metrics with a minimum learning error.

Results. The prediction with the highest accuracy for the CzIE was from the lightgbm model 72% for training, 76% for testing and 80% for validation, precision 0.74 which means that 74% of the CzIE are correct, it can be considered good, but could be improved, a sensitivity (Recall) of 90% suggests that the model is effective in detecting positive cases, F1-Score of 0.81 indicates good performance of the model

and a specificity of 0.72% refers to a Good identification of true negatives crucial for medical decision making.

Conclusion. The lightgbm model is recommended as the most optimal ML model, from the validation of evaluation metrics for the prediction of the inoculum effect of cefazolin in MSSA. This model can be considered as a diagnostic alternative, based on genomic tools such as sequencing that is increasingly promising and accessible, being the beginning of a great advance in speed, efficiency, and precision in clinical environments to improve the quality of life of the patients of our country.

INTRODUCCIÓN

La Resistencia a los Antimicrobianos (RAM), como indica la Organización Mundial de la Salud, ha representado un desafío significativo en el ámbito de la Salud Pública durante décadas (8). Los microorganismos resistentes comprometen la eficacia de los tratamientos para enfermedades infecciosas comunes, lo que resulta en enfermedades prolongadas, discapacidad y en ocasiones, aumento en indicadores de mortalidad (8). La complejidad y gravedad de la RAM a nivel mundial generan preocupación, estimándose que para el año 2050 las muertes atribuibles a esta superarán los 10 millones anuales, sobrepasando las causadas por enfermedades graves como el cáncer, la diabetes, el cólera y el tétanos, entre otras (9). Además, la RAM ha tenido un impacto significativo en la economía global, proyectándose una reducción del 3.5% en el Producto Interno Bruto (PIB) mundial para ese mismo año, lo que implicaría pérdidas económicas de hasta 100 billones de dólares (8,9).

El *Staphylococcus aureus*, una bacteria habitual en el microbioma humano coloniza de forma asintomática entre el 20% y el 40% de la población (10). Sin embargo, su capacidad patógena puede desencadenar una amplia variedad de infecciones, desde leves como los carbúnculos, hasta graves, como las infecciones de piel y tejidos blandos, óseas, articulares, neumonía, bacteriemia y diversas enfermedades inducidas por toxinas, como el síndrome de shock tóxico (11). Es crucial subrayar que el *S. aureus* es el principal agente etiológico de las Infecciones Asociadas a la Atención en Salud (IAAS) y es la bacteria más frecuentemente aislada en pacientes atendidos en servicios de emergencia en los Estados Unidos (12). En Colombia, *S. aureus* se posiciona como el cuarto microorganismo responsable de infecciones en Unidades de Cuidados Intensivos (UCI) y el tercero en infecciones que provocan hospitalización (13). Las infecciones producidas por *Staphylococcus aureus* susceptible a metilicina (SASM) presentan una incidencia y complejidad equiparables a las causadas por cepas resistentes a metilicina (SARM) (12,14). De hecho, en nuestro país, el 60% de las infecciones atribuibles a *Staphylococcus aureus* son ocasionadas por cepas susceptibles a metilicina (SASM) (15,16).

La oxacilina es el antibiótico de primera línea para tratamiento por SASM. El uso de la cefazolina se ha convertido en una excelente alternativa para el tratamiento de infecciones por estos microorganismos debido a que tiene la misma efectividad que la oxacilina y posee menores efectos adversos para el paciente, menor mortalidad y es de menor costo (17). Se ha reportado un fenómeno de resistencia a cefazolina

por parte de SASM, denominado efecto inóculo a cefazolina (CzIE), el cual ocurre en infecciones con alto inóculo bacteriano como bacteriemia, endocarditis, osteomielitis, entre otras (7,18,19). Este fenómeno no es detectable en laboratorios clínicos y la prueba estándar para su detección es demasiado dispendiosa y costosa (19). Sin embargo, se ha llevado a cabo una caracterización genómica de los SASM que presentan este fenotipo de resistencia, y se encontró que la mayoría de ellos están asociados a la enzima BlaZ tipo A y C (7,18,19). También, se han identificado por otro tipo de clasificación, un total de 29 alotipos y 43 sustituciones diferentes en el gen BlaZ, donde notablemente, el alotipo BlaZ-2 mostró una asociación estadísticamente significativa con el fenómeno de resistencia a la cefazolina, mientras que los alotipos BlaZ-3 y BlaZ-5 se relacionaron con la ausencia del mismo (7,18,19).

En este estudio, se evidenció que no existe una característica genética única en la secuencia genómica que identifique por sí sola el fenómeno del CzIE (7,18,19). Este hallazgo representa una novedad en este campo que evidencia la necesidad de estudiar este fenómeno bajo diferentes perspectivas. En el año 2021, se desarrolló una prueba de detección rápida a cefazolina, con una sensibilidad del 82.5% y especificidad del 88.9%; la cual constituye una herramienta diagnóstica útil, sin embargo aún no está disponible en el mercado (19). Aunque se han realizado numerosas investigaciones sobre este tipo de resistencia y se han explorado alternativas diagnósticas, hasta el momento no se ha desarrollado un modelo matemático capaz de prever este fenómeno de manera efectiva. Un algoritmo de este tipo sería invaluable para guiar una terapia más precisa, lo que podría asegurar la continuación del uso de la cefazolina como antibiótico de elección primaria. Esto es especialmente relevante considerando su accesibilidad y costo en comparación con la oxacilina.

El aprendizaje automático (Machine Learning - ML) está transformando la atención médica al ofrecer nuevas herramientas y perspectivas para la prevención, diagnóstico y tratamiento de enfermedades. Por medio de su implementación, se mejora la precisión, eficiencia y potencial predictivo de pruebas diagnósticas, contribuyendo así a una optimización del cuidado de la salud a nivel mundial. Sin embargo, es fundamental abordar cuestiones éticas y de privacidad de los datos, para garantizar un uso responsable de estas tecnologías. Resulta importante reconocer el papel de la ciencia de datos en la resolución de problemas biológicos y la determinación de pronósticos para múltiples enfermedades (20). El ML puede procesar grandes cantidades de datos e identificar patrones complejos que podrían ser imposibles de detectar manualmente. Es importante mencionar que la magnitud del problema y el impacto de la RAM a nivel mundial en la salud humana y en los costos del sector de la salud es amplio y todavía en gran parte desconocido (21).

Por lo tanto, nos hemos planteado como objetivo la implementación de un modelo de clasificación que podría establecer una relación entre las características (Identificador de la bacteria, Secuencias) y las etiquetas (CzIE positivo o negativo) con el fin de mejorar la capacidad predictiva sobre la respuesta de un aislamiento

de *Staphylococcus aureus* susceptible a meticilina (SASM) que conduciría a un diagnóstico más preciso, y que a su vez podría resultar en una terapia más efectiva.

MATERIALES y MÉTODOS

Recolección y obtención de datos de los aislamientos de SASM.

Las secuencias de los genomas completos de los aislamientos de SASM fueron tomados del repositorio perteneciente a la Unidad de Genética y Resistencia Antimicrobiana de la Universidad El Bosque (UGRA), las cuales se obtuvieron por medio de la secuenciación de genoma completo a través de la plataforma illumina. Se seleccionaron 410 secuencias del operón BlaZRI divididas en dos grupos: i) 369 secuencias de esta porción genética de aislamientos de SASM causantes de bacteriemia recolectados como parte de un estudio de vigilancia multicéntrica en hospitales de nueve países de América Latina de 2011 a 2019 (15), 41 secuencias de aislamientos de SASM colonizadores obtenidos de pacientes de UCI de 6 hospitales de alta complejidad obtenidos bajo la realización de un estudio realizado en Colombia (Minciencias 776-2018). Se incluyeron en el desarrollo de este modelo las siguientes variables: i) identificación única de cada cepa o bacteria, ii) la presencia (Positivo) o ausencia (negativo) del efecto inóculo a cefazolina, y iii) la secuencia del operón BlaZRI correspondiente a 3.083 nucleótidos.

Para el preprocesamiento de datos, se seleccionaron datos categóricos. Inicialmente, se verificaron datos faltantes en las 410 secuencias genómicas, identificándose en la columna 1.664 (n=22) y en la columna 2.334 (n=2), resultando en un total de 24 aislamientos de SASM que se eliminaron y no se imputaron debido a la presencia de deleciones (pérdida de segmentos de ADN) reflejadas en la secuencia. Es importante destacar que la eliminación de estos datos no afectó el análisis ni el balance de la base de datos. Tras esta depuración, se incluyeron 386 secuencias del operón BlaZRI para el análisis de entrenamiento, prueba y validación.

Para la definición de los K-meros, se generaron secuencias únicas a partir de diferentes agrupaciones pequeñas, identificadas con un número específico. Las copias adicionales fueron eliminadas y la secuencia completa de ADN se convirtió en una matriz numérica, donde todos los K-meros generados se concatenaron para formar grupos de letras, reduciendo así la dimensionalidad de los datos (22). En este estudio, se empleó la codificación de etiquetas correspondiente a la presencia del efecto inóculo a cefazolina como positivo = 1 y la ausencia del CzIE como negativo = 0. Cada secuencia de ADN se transformó en un K-mero de diferentes tamaños "m", desde un mínimo de 5 hasta un máximo de 27, donde los K-meros con mejor rendimiento fueron los de tamaños 5, 7, 12, 17 y 23. De acuerdo con este procesamiento, se seleccionó un tamaño de K-mero de 23, ya que este valor proporcionó las mejores métricas, como la exactitud, entre otras, en el modelo escogido.

Selección de la arquitectura del modelo adecuado para la realización del entrenamiento, prueba y validación.

Se llevó a cabo la asignación de etiquetas y características, junto con la partición de los datos en conjuntos de entrenamiento, utilizando el 60% de las secuencias (n=231), el 28% como conjunto de prueba (n=108) y el 12% como conjunto de validación (n=47).

La selección de los modelos se realizó a través de PyCaret, una herramienta que facilita la construcción y despliegue de modelos de aprendizaje automático (ML) de manera rápida, eficiente y con código sintetizado para problemas de clasificación binaria. PyCaret incluye alternativas para la optimización automática de hiperparámetros, lo que permite encontrar la combinación óptima de parámetros de un modelo para mejorar su rendimiento.

Se exploraron diferentes metodologías de clasificación, es importante destacar únicamente aquellos modelos con mejor desempeño. Basándonos en las métricas proporcionadas, se analizaron y destacaron los tres primeros modelos con mejor funcionamiento en términos de exactitud, precisión, sensibilidad, puntuación F1 y área bajo la curva (AUC). Conforme a la capacidad de predecir correctamente cada etiqueta (exactitud), el modelo LightGBM mostró el mayor porcentaje con un 74% de exactitud, seguido por KNN con 72% y Ridge con 71%. En la proporción de predicciones positivas correctas entre todas las predicciones positivas realizadas por el modelo (precisión), el mejor fue LightGBM con 75%, seguido por KNN con 74% y Ridge con 68%. En términos de sensibilidad (recall), que mide los casos positivos correctamente identificados, el modelo Ridge obtuvo un 85%, LightGBM un 77% y KNN un 74%. En la puntuación F1, que combina tanto la precisión como la sensibilidad, LightGBM y Ridge alcanzaron un 76%, mientras que KNN obtuvo un 73%. Respecto a la capacidad para distinguir entre clases positivas y negativas (AUC), un AUC más alto indica un mejor rendimiento del modelo, siendo LightGBM el mejor con un 79%, seguido por KNN con 77% y Ridge con 76% (Tabla 1). Se tomo como guía la exploración de modelos previa como referencia para realizar entrenamiento, evaluación y validación del top 3 de los mejores modelos de forma individual.

Tabla 1. Comparación de métricas de los modelos de aprendizaje automático (ML) implementados con PyCaret.

Modelo	Abrev.	Exactitud	AUC	Sensibilidad	Prec.	F1	Kappa	MCC	TT (Sec)
Light Gradient Boosting Machine	lightgbm	0.7451	0.7926	0.7740	0.7598	0.7651	0.4855	0.4885	573.820
K Neighbors Classifier	knn	0.7234	0.7735	0.7417	0.7434	0.7397	0.4433	0.4487	0.2860
Ridge Classifier	ridge	0.7103	0.7633	0.8550	0.6899	0.7620	0.4034	0.4193	0.0860
Logistic Regression	lr	0.7059	0.7668	0.8470	0.6876	0.7576	0.3950	0.4091	0.6560
Naive Bayes	nb	0.7057	0.6868	0.9437	0.6582	0.7751	0.3866	0.4434	0.2840
Linear Discriminant Analysis	lda	0.6888	0.7142	0.8387	0.6706	0.7438	0.3595	0.3765	0.1780

Ada Boost Classifier	ada	0.6717	0.7531	0.6770	0.7034	0.6843	0.3399	0.3460	0.1800
Gradient Boosting Classifier	gbc	0.6716	0.7460	0.7333	0.6807	0.7033	0.3341	0.3396	0.5260
Decision Tree Classifier	dt	0.6714	0.6694	0.7020	0.6926	0.6968	0.3387	0.3392	0.2520
Random Forest Classifier	rf	0.6714	0.7407	0.6857	0.7002	0.6921	0.3403	0.3411	0.2080
Extra Trees Classifier	et	0.6713	0.6919	0.6853	0.7002	0.6917	0.3400	0.3411	0.2640
SVM - Linear Kernel	svm	0.6671	0.6988	0.9597	0.6281	0.7573	0.3012	0.3784	0.1000
Dummy Classifier	dummy	0.5368	0.5000	1.000	0.5368	0.6986	0.0000	0.0000	0.0740
Quadratic Discriminant Analysis	qda	0.4979	0.6332	0.1770	0.6746	0.2676	0.0417	0.0832	0.2600

Abrev. Abreviatura del nombre del modelo; **Exactitud:** proporción de todas las predicciones (tanto verdaderos positivos como verdaderos negativos) que son correctas en relación con el total de predicciones; **AUC (Área bajo la curva ROC):** Es el área bajo la curva ROC (Receiver Operating Characteristic). Esta métrica proporciona una medida de la capacidad de discriminación del modelo en distintos umbrales de clasificación; **Sensibilidad:** Sensibilidad o Recall es la proporción de verdaderos positivos sobre el total de predicciones positivas (verdaderos positivos más falsos positivos). La precisión mide cuántas de las instancias clasificadas como positivas son realmente positivas; **Prec. (Precision):** También conocida como precisión positiva, es lo mismo que la precisión mencionada anteriormente; **F1 (F1-score):** Es la media armónica entre precisión y sensibilidad. Proporciona un balance entre estas dos métricas; **Kappa (Coeficiente Kappa):** El coeficiente kappa es una métrica que evalúa la concordancia entre las predicciones del modelo y las clases reales, teniendo en cuenta la posibilidad de que las predicciones se produzcan al azar; **MCC (Coeficiente de correlación de Matthews):** El coeficiente de correlación de Matthews es otra medida de la concordancia entre las predicciones del modelo y las clases reales. Es particularmente útil en problemas de clasificación binaria desbalanceada; **TT (Tiempo de Entrenamiento):** Es el tiempo, generalmente en segundos, que el modelo tardó en entrenarse.

RESULTADOS

Conjunto de entrenamiento

Se emplearon las técnicas de búsqueda por grilla y validación cruzada para el ajuste de hiperparámetros, evaluación del rendimiento y reducción del riesgo de sobreajuste. La validación cruzada es una técnica de evaluación de modelos de aprendizaje automático que divide el conjunto de datos en múltiples subconjuntos o "folds". El modelo se entrena en algunos de estos subconjuntos y se prueba en los restantes, rotando el conjunto de prueba en cada iteración. Esto permite estimar el rendimiento del modelo de manera más fiable y reduce el riesgo de sobreajuste, asegurando que los resultados sean generalizables a datos no vistos. En el procesamiento individual de cada modelo, se analizaron las siguientes métricas: exactitud, sensibilidad, precisión, especificidad, matriz de confusión y área bajo la curva (AUC).

Los modelos Knn y Ridge, mostraron una exactitud del 71% y una AUC del 88% y 85% respectivamente, mientras el modelo LightGBM alcanzó la mayor exactitud con un 72%, lo que indica que el modelo está aprendiendo correctamente de los datos y es capaz de generalizar adecuadamente para realizar predicciones precisas en datos nuevos. Un AUC del 90% refleja un rendimiento excelente del modelo en términos de su capacidad para distinguir entre clases positivas y negativas (Tabla 2).

Tabla 2. Comparación de métricas de modelos de ML implementados en el conjunto de entrenamiento.

MÉTRICAS ENTRENAMIENTO		
Modelo	Exactitud	Curva AUC
Light Gradient Boosting Machine	72%	90%
K Neighbors Classifier	71%	88%
Ridge Classifier	71%	85%

Conjunto de prueba

En el conjunto de prueba, se destacan las métricas del modelo LightGBM, que presenta una exactitud de predicción del 76%. Por su parte, el modelo Ridge exhibe una sensibilidad del 95%, indicando una alta eficacia para identificar correctamente los casos positivos. Además, cuenta con una especificidad del 94%, lo que demuestra su efectividad para identificar correctamente los casos negativos, y una precisión del 91%, reflejando una alta proporción de predicciones positivas correctas entre todas las predicciones positivas realizadas. Con un AUC del 86%, el modelo Ridge Classifier demuestra una notable capacidad para distinguir entre clases, sugiriendo que es capaz de realizar predicciones precisas y confiables en los datos de prueba (Tabla 3).

Tabla 3. Comparación de métricas de modelos de ML implementados en el conjunto de prueba.

MÉTRICAS PRUEBA						
Modelo	Exactitud	Sensibilidad	Especificidad	Precisión	AUC	F1
Light Gradient Boosting	76%	75%	75%	78%	85%	77%
K Neighbors Classifier	73%	75%	75%	74%	81%	75%
Ridge Classifier	83%	95%	94%	91%	86%	83%

Según los resultados relacionados con las métricas de desempeño mostrados en la Tabla 3 podemos concluir que el modelo Light Gradient Boosting Machine (LightGBM) es el mejor modelo de ML para la predicción del efecto del inóculo a la metilicina (CzIE). LightGBM alcanza las mejores métricas utilizando los hiperparámetros óptimos, como el número de vecinos considerado para la predicción, que es de 5, y el uso de un esquema de ponderación uniforme, lo que indica que todos los vecinos tienen el mismo peso en la predicción. Este modelo muestra una exactitud de predicción del 76%, una precisión del 78%, y aunque no tiene la sensibilidad más alta, su valor sigue siendo competitivo con un 75%, indicando su capacidad para identificar correctamente una alta proporción de casos positivos. Además, presenta una puntuación F1 del 77% y un área bajo la curva

(AUC) del 85%, lo que lo destaca en comparación con los otros modelos evaluados (Tabla 3).

Conjunto de Validación

En cuanto a la validación podemos decir que el modelo LightGBM tiene la mayor exactitud (80%), seguido por K Neighbors (74%) y Ridge (72%), tanto LightGBM como Ridge tienen la misma sensibilidad del (90%), lo que los hace muy efectivos en la detección de verdaderos positivos, LightGBM también tiene la mayor especificidad (72%), lo que significa que maneja mejor la identificación de verdaderos negativos en comparación con los otros dos modelos y tiene la mayor precisión (74%), lo que indica que es el mejor en términos de evitar falsos positivos. LightGBM lidera con un AUC del 89%, lo que muestra su superioridad en la capacidad general de discriminación entre las clases.

Tabla 4. Comparación de métricas de modelos de ML implementados en el conjunto de validación.

MÉTRICAS VALIDACIÓN					
Modelo	Exactitud	Sensibilidad	Especificidad	Precisión	AUC
Light Gradient Boosting	80%	90%	72%	74%	89%
K Neighbors Classifier	74%	86%	64%	67%	83%
Ridge Classifier	72%	90%	56%	64%	83%

LightGBM es el modelo más robusto en términos de balance entre sensibilidad, especificidad, precisión y AUC, sin embargo, si la prioridad es detectar la mayor cantidad de verdaderos positivos (sensibilidad alta), tanto LightGBM como Ridge Classifier son buenas opciones. LightGBM sigue siendo preferible debido a su mejor precisión y especificidad comparativa.

Matriz de confusión

Las matrices de confusión ofrecen una visión detallada del rendimiento de los modelos de clasificación (Tabla 5) al mostrar el número de predicciones correctas e incorrectas para cada clase. Aquí hay algunas observaciones sobre las matrices de confusión de los modelos LightGBM, KNN y Ridge:

Tabla 5. Comparación de Matrices de confusión de modelos de ML implementados en el conjunto de validación.

MÁTRICES DE CONFUSIÓN								
LightGBM			Knn			Ridge		
	Pos (+)	Neg (-)		Pos (+)	Neg (-)		Pos (+)	Neg (-)
Pos (+)	VP=18	FN=7	Pos (+)	VP=16	FN=9	Pos (+)	VP=14	FN=11
Neg (-)	FP=2	VP=20	Neg (-)	FP=3	VP=19	Neg (-)	FP=2	VP=20

Pos (+)= Resultado Positivo; Neg (-)= Resultado Negativo, VP = Verdadero Positivo; VN = Verdadero Negativo; FP = Falso Positivo; FN = Falso Negativo

LightGBM tiene un buen rendimiento y de los 3, es que el mejor predice los verdaderos positivos (casos correctamente identificados). El modelo Knn y Ridge también tienen un buen rendimiento, pero evidencian más falsos negativos en comparación con LightGBM.

Finalmente, las curvas ROC mostradas en las figura 1, 2 y 3 tienen AUC relativamente altos, lo que sugiere que todos son capaces de realizar predicciones discriminativas en el problema dado. Sin embargo, LightGBM tiene el mejor rendimiento con un AUC del 89%.

Figura 1. Curva ROC (Receiver Operating Characteristic) para el modelo **LightGBM**

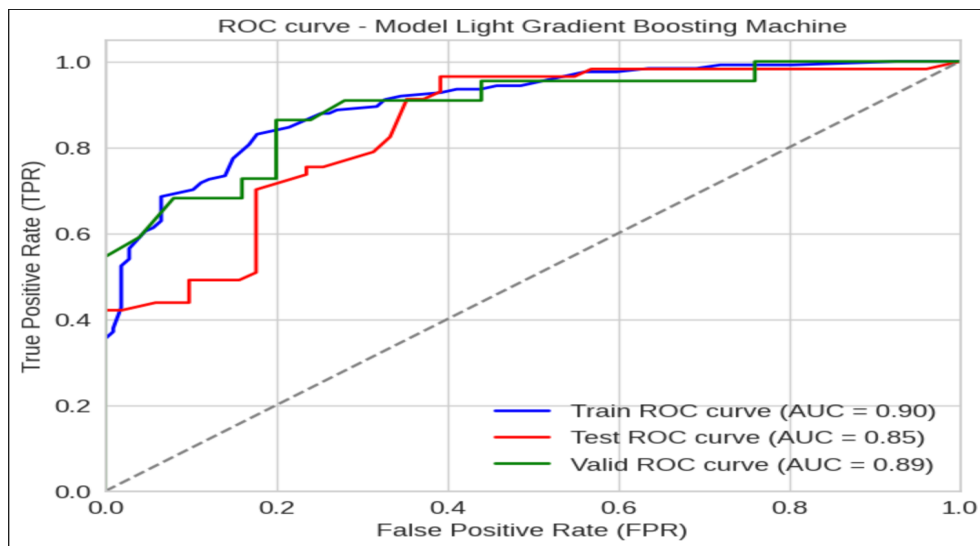


Figura 2. Curva ROC (Receiver Operating Characteristic) para el modelo **Knn**

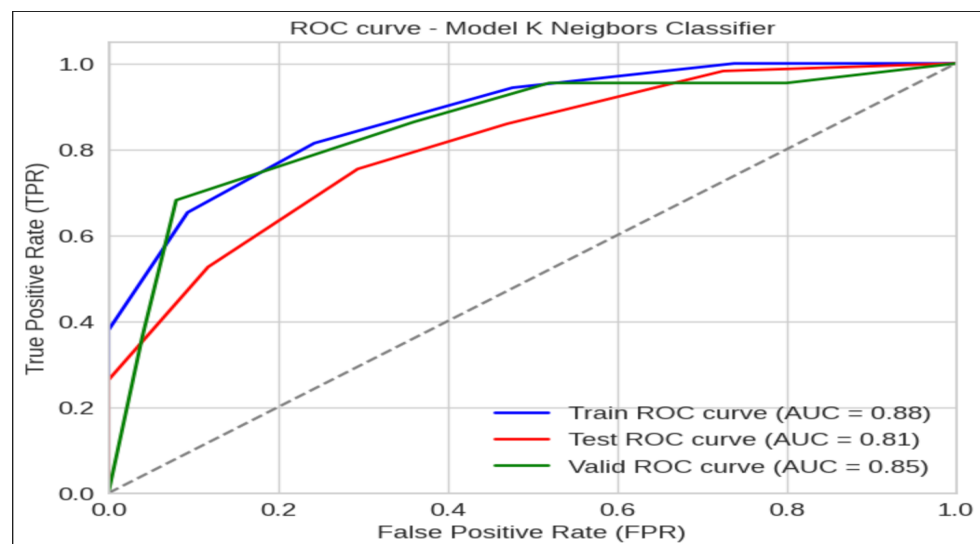
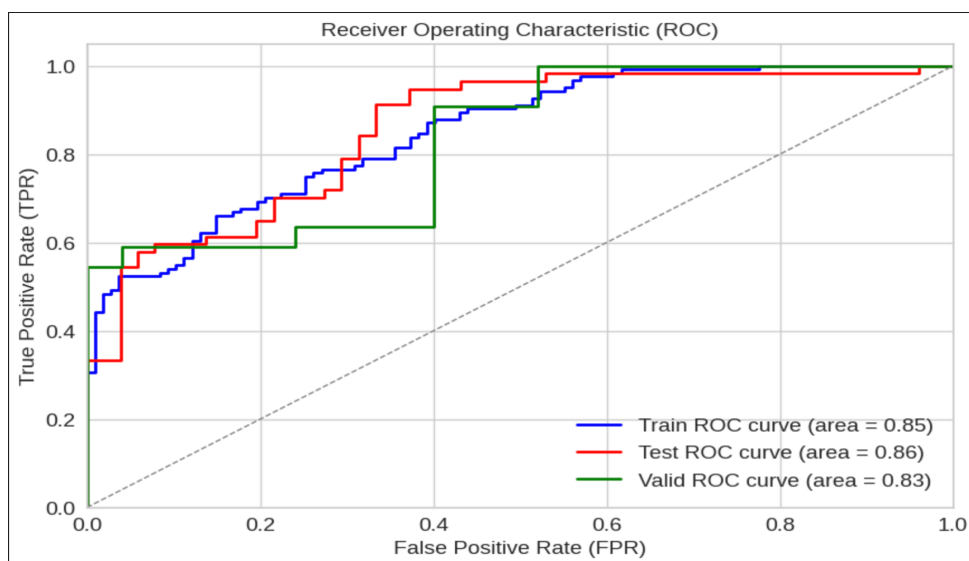


Figura 3. Curva ROC (Receiver Operating Characteristic) para el modelo **Ridge**



El modelo Light Gradient Boosting Machine (LightGBM) se destaca como el mejor modelo para predecir el efecto del inóculo a la metilina (CzIE) debido a su alto rendimiento en términos de sensibilidad, especificidad, precisión y exactitud.

DISCUSIÓN

El modelo LightGBM fue identificado, como el más eficaz para predecir el efecto inóculo a Cefazolina en *Staphylococcus aureus* susceptible a metilina. Las métricas obtenidas, pueden ser consideradas con alta representatividad matemática, y es sugerido su uso como una aproximación estadística útil en escenarios diagnósticos. LightGBM predice la presencia o ausencia del efecto con una sensibilidad del 90% mediante la utilización de técnicas de procesamiento de lenguaje natural, tokenización y agrupación por K-meros = 23. En este contexto, se obtuvo una especificidad del 72%, una exactitud del 80% y una precisión del 74%. Prospectivamente, se espera que este modelo produzca un efecto estadístico confiable, incluso en escenarios de implementación limitados. LightGBM se elige sobre KNN y RC debido a su mayor exactitud arrojando predicciones correctas en general, aunque RC iguala a LightGBM en sensibilidad, LightGBM mantiene altas otras métricas como especificidad, precisión y AUC.

De acuerdo a lo mencionado anteriormente, estas ventajas hacen de LightGBM el modelo más robusto y fiable frente a algoritmos alternativos. Adicionalmente y si nos basamos en la literatura como reportan Fernández y colaboradores (21), la aplicación de métodos de ML son opciones viables para respaldar la detección de compuestos con actividad antibacteriana contra *S. aureus*, estos pueden ayudar hacia una mayor optimización de la actividad biológica, con el objetivo de desarrollar posibles agente terapéuticos para el tratamiento de infecciones bacterianas especialmente aquellas causadas por MRSA (21). Dentro de los modelos usados como K vecinos más cercanos (Knn), Análisis de componentes principales (PCA), Máquinas de soporte vectorial (SVM), Random Forest (RF), Perceptrón Multicapa

(MLP), Naive bayes (NB) y arboles de decisión (DT), emplearon al igual que el presente estudio técnicas de validación cruzada y evaluados con métricas como el coeficiente de correlación de Matthews (MCC), la puntuación F1, el área bajo la curva (AUC), la tasa de verdaderos positivos (TPR) y la tasa de verdaderos negativos (TNR). El algoritmo con mejores métricas fue el Knn siendo el más recomendado para clasificar compuestos como antibacterianos activos y/o inactivos (21).

Recientemente Wang y colaboradores (23), han publicado modelos que relacionan la resistencia a *Staphylococcus aureus* para clindamicina, cefoxitina, trimetoprim-sulfametoxazol demostrando que es importante vincular directamente el genotipo y el fenotipo de las bacterias. En este estudio se predijeron con éxito las características de resistencia de *S. aureus* a los principales agentes antimicrobianos (concentraciones inhibitorias mínimas de 10 agentes antimicrobianos) usando 466 aislamientos mediante la extracción de k-meros (K=11) de datos de secuenciación del genoma completo evaluando tres algoritmos de ML, Random forest (RF), Maquinas de soporte vectorial (SVM) y XGBoost, ellos para evaluar la confiabilidad del modelo se obtuvieron las curvas ROC con los mejores resultados de validación cruzada, dando como resultado que más del 90% de las características antimicrobianas podrían proporcionar información importante para el tratamiento clínico (23).

Se encontró en una revisión sistemática de 25 estudios (24) donde se implementó aprendizaje automático (ML) con una puntuación de riesgo como herramienta para predecir la RAM, el *S. aureus* resistente a la meticilina (MRSA) y la resistencia a los carbapenémicos fueron los resultados más comunes en los estudios con un patrón de resistencia específico. Los algoritmos más comunes en la predicción de ML fueron la regresión logística (n = 14 estudios), el árbol de decisión (n = 14) y el bosque aleatorio (n = 7). El rango del área bajo la curva (AUC) para la predicción de ML fue de 0,48 a 0,93. El AUC combinado para la predicción de ML fue 0,82 (0,78–0,85). En comparación con la puntuación de riesgo, se indicó una mayor especificidad [87% (82–91) frente a 37% (25–51)] para la predicción de ML, pero no sensibilidad [67% (62–72) frente a 73% (67–79).] (24). Ríos y colaboradores (25), procesaron un modelo de aprendizaje automático para predecir el efecto inoculo a Cefazolina encontrando una notable cantidad de falsos negativos (52%), contrario a lo que se identificó en nuestras métricas de evaluación; atribuyendo este resultado a la heterogeneidad de la población incluida en esa aproximación diagnóstica (25).

La secuencia analizada en nuestro trabajo mostró una precisión del 74%, menor en comparación con el 90% reportado por Wang y colaboradores (23). Es importante mencionar que el procesamiento de los datos fue diferente y el tamaño de las secuencias de estos dos últimos estudios es significativamente mayor, considerando que se utilizó únicamente el operón relacionado con el gen. El modelo Light Gradient Boosting Machine (LightGBM) evidencia diferentes ventajas significativas sobre otros métodos diagnósticos tradicionales y algunos métodos de aprendizaje automático tales como un alto rendimiento en métricas claves proporcionando un desempeño confiable, manejo eficaz de datos de gran dimensión

y complejidad como los genómicos. Ha demostrado su capacidad para evitar el sobreajuste con un buen desempeño en datos de entrenamiento como en datos nuevos.

CONCLUSIÓN

Los enfoques para los modelos de predicción en aprendizaje automático que se han desarrollado a partir de secuencias genómicas para *S. aureus* y para muchas otras especies de bacterias, suelen estar optimizados para maximizar las métricas de medición respecto a determinados genes, resaltando intencionalmente sus capacidades de diagnóstico, por encima de su utilidad para descubrir mecanismos genéticos de resistencia. La evaluación comparativa de los modelos de aprendizaje automático para predecir el efecto inóculo a cefazolina en *S. aureus* susceptible a meticilina demuestra que el modelo Light Gradient Boosting Machine (LightGBM) es superior en términos de exactitud, sensibilidad, especificidad, precisión y área bajo la curva ROC (AUC), no solo se destaca como el modelo más preciso y equilibrado, sino que también ofrece una robusta capacidad para generalizar nuevos datos, convirtiéndolo en una herramienta invaluable para mejorar el diagnóstico y por ende la terapia antimicrobiana. La aplicación del modelo LightGBM puede revitalizar el uso de cefazolina, un antibiótico eficaz, seguro y económico, lo cual en la actualidad con la utilización en los hospitales del denominado “Antimicrobial Stewardship” apoya totalmente la utilización de este tipo herramientas en el mejoramiento del gerenciamiento de los medicamentos, en esta caso, el rescate de un antibiótico. Esto no solo mejorará los resultados de salud de los pacientes, reduciendo la incidencia de efectos adversos y la resistencia antibiótica, sino que también tendrá un impacto positivo en los costos de atención médica. Por tanto, la implementación de este modelo representa una estrategia efectiva para optimizar el tratamiento de infecciones por *Staphylococcus aureus*, beneficiando tanto la salud pública como en la economía del sistema de salud.

REFERENCIAS

1. Aguado JM, San-Juan R, Lalueza A, Sanz F, Rodríguez-Otero J, Gómez-Gonzalez C, Chaves F. High vancomycin MIC and complicated methicillin-susceptible *Staphylococcus aureus* bacteremia. *Emerg Infect Dis*. 2011;17(6):1099–102
2. Tong SY, Davis JS, Eichenberger E, Holland TL, Fowler VG, Jr. *Staphylococcus aureus* infections: epidemiology, pathophysiology, clinical manifestations, and management. *Clin Microbiol Rev* 2015 28:603– 661.
3. van Hal SJ, Jensen SO, Vaska VL, Espedido BA, Paterson DL, Gosbell IB. Predictors of mortality in *Staphylococcus aureus* Bacteremia *Clin Microbiol Rev* 2012 25:362-86.
4. Kainer MA, Lynfield R, Greissman S, et al. Changes in prevalence of health care-associated infections in U.S. hospitals. *N Engl J Med* 379: 1732–1744.
5. Liu C, Bayer A, Cosgrove SE, et al. Clinical practice guidelines by the infectious diseases society of America for the treatment of methicillin-resistant *Staphylococcus aureus* infections in adults and childrens. *Clin Infect Dis* 2011 52:e18 – e55

6. Minter DJ, Appa A, Chambers HF, Doernberg SB. Contemporary Management of *Staphylococcus aureus* Bacteremia—Controversies in Clinical Practice. *Clin Infect Dis* 2023 77:e57–e68
7. Rincon S, Reyes J, Carvajal LP, et al. Cefazolin high-inoculum effect in methicillin-susceptible *Staphylococcus aureus* from South American hospitals. *J Antimicrob Chemother* 2013 68:2773-8.
8. OMS. Resistencia a los antimicrobianos. Noviembre 17 de 2021.
9. O’neill J. Antimicrobial Resistance: Tackling a crisis for the health and wealth of nations. Review on Antimicrobial Resistance. HM Government – The Wellcome Trust. December 2014:1-16.
10. Piewngam P, Otto M. *Staphylococcus aureus* colonisation and strategies for decolonisation. *Lancet Microbe*. 2024 Mar 19:S2666-5247(24)00040-5.
11. Liu A, Garrett S, Hong W, Zhang J. *Staphylococcus aureus* Infections and Human Intestinal Microbiota. *Pathogens* 2024 24;13:276.
12. Inagaki K, Lucar J, Blackshear C, Hobbs C V. Methicillin-susceptible and Methicillin-resistant *Staphylococcus aureus* Bacteremia: Nationwide Estimates of 30-Day Readmission, In-hospital Mortality, Length of Stay, and Cost in the United States. *Clin Infect Dis* 2019;69:2112–8.
13. INS. Informe De Resultados De La Vigilancia Por Laboratorio De Resistencia Antimicrobiana En Infecciones Asociadas a La Atención En Salud (IAAS) 2018. Dirección Epidemiol las Infecc Asoc a la atención en salud. 2019;29–44.
14. Gould IM, Reilly J, Bunyan D, Walker A. Costs of healthcare-associated methicillin-resistant *Staphylococcus aureus* and its control. Vol. 16, *Clinical Microbiology and Infection*. Blackwell Publishing Ltd; 2010. p. 1721–8.
15. Arias CA, Reyes J, Carvajal LP, et al. A prospective cohort multicenter study of molecular epidemiology and phylogenomics of *Staphylococcus aureus* bacteremia in nine Latin American countries. *Antimicrob Agents Chemother* 2017 61:e00816-17.
16. Reyes J, Rincón S, Díaz L, Panesso D, Contreras GA, Zurita J, et al. Dissemination of Methicillin-Resistant *Staphylococcus aureus* USA300 Sequence Type 8 Lineage in Latin America. *Clin Infect Dis*. 2009 Dec;49(12):1861–7.
17. Li J, Echevarria KL, Hughes DW, Cadena J, Bowling JE, Lewis JS. 2nd. Comparison of cefazolin versus oxacillin for treatment of complicated bacteremia caused by methicillin-susceptible *Staphylococcus aureus*. *Antimicrob Agents Chemother* 2014 58:5117–5124.
18. Carvajal LP, Rincon S, Echeverri AM, et al. Novel Insights into the Classification of *Staphylococcal* β -Lactamases in Relation to the Cefazolin Inoculum Effect. *Antimicrob Agents Chemother* 2020 64:e02511-19.
19. Rincon S, Carvajal LP, Gomez-Villegas SI, et al. A Test for The Rapid Detection of the Cefazolin Inoculum Effect in Methicillin-Susceptible *Staphylococcus aureus*. *J Clin Microbiol* 2021 59:e01938-20.
20. Wang W, Baker M, Hu Y, et al. Whole-Genome Sequencing and Machine Learning Analysis of *Staphylococcus aureus* from Multiple Heterogeneous Sources in China Reveals Common Genetic Traits of Antimicrobial Resistance. *mSystems* 2021 6:e0118520.

21. Fernandes PO, Dias ALT, Dos Santos Júnior VS, et al. Machine Learning-Based Virtual Screening of Antibacterial Agents against Methicillin-Susceptible and Resistant *Staphylococcus aureus*. *J Chem Inf Model* 2024 64:1932-1944.
22. Ge J, Meng J, Guo N, Wei Y, Balaji P, Feng S. Counting Kmers for Biological Sequences at Large Scale. *Interdiscip Sci* 2020 12:99-108.
23. Wang S, Zhao C, Yin Y, Chen F, Chen H, Wang H. A Practical Approach for Predicting Antimicrobial Phenotype Resistance in *Staphylococcus aureus* Through Machine Learning Analysis of Genome Data. *Front Microbiol* 2022 13:841289.
24. Tang R, Luo R, Tang S, Song H, Chen X. Machine learning in predicting antimicrobial resistance: a systematic review and meta-analysis. *Int J Antimicrob Agents* 2022 60:106684.
25. Rios R, Gomez-Villegas SI, McNeil JC, et al. A Machine-Learning Approach to Predict the Cefazolin Inoculum Effect in Methicillin-Susceptible *Staphylococcus aureus*. *Open Forum Infectious Diseases* 2021 8:S712–S713. P1248.