Universidad Autónoma de Baja California

Facultad de Ingeniería, Arquitectura y Diseño

Maestría y Doctorado en Ciencias e Ingeniería

Análisis de mutaciones en genoma de *Mycobacterium tuberculosis* para la predicción de resistencia a fármacos utilizando métodos de *Machine Learning*

Tesis para la obtención de grado de:

Maestro en Ingeniería

Presenta:

Guillermo René Paredes Gutiérrez

Directora:

Dora Luz Flores Gutiérrez

Codirectora:

Raquel Muñiz Salazar

Ensenada, Baja California, a diciembre de 2023

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

FACULTAD DE INGENIERÍA, ARQUITECTURA Y DISEÑO

MAESTRÍA Y DOCTORADO EN CIENCIAS E INGENIERÍA

Análisis de mutaciones en genoma de *Mycobacterium tuberculosis* para la predicción de resistencia a fármacos utilizando métodos de *Machine Learning*

TESIS

Que para obtener el grado de maestro en ingeniería:

Guillermo René Paredes Gutiérrez

Aprobada por:

Dora Luz Flores Gutiérrez
Director de tesis

Raquel Muñiz Salazar
Miembro del comité

Dante Alberto Magdaleno Moncayo
Miembro del comité

Everardo Gutiérrez López
Miembro del comité

Dayanira Sheira Paniagua Meza
Miembro del comité

Ensenada Baja California, México. Noviembre 2023

Agradecimientos

A mis padres, Guillermo Paredes Acevedo y Angélica Gutiérrez Wenceslao, por su amor incondicional, su apoyo constante y su fe en mí. Gracias por enseñarme a ser una persona responsable y a luchar por mis sueños.

A mi novia, Andrea Manríquez, por su comprensión, paciencia y amor durante todo este proceso. Gracias por estar siempre ahí para mí y por ser mi roca en momentos de dificultad.

A mis amigos, por su amistad y por estar siempre dispuestos a ayudar y a brindarme su apoyo. Gracias por hacer que mi camino sea más llevadero y por compartir conmigo tantos buenos momentos.

A mi directora, Dora Luz Flores, y codirectora, Raquel Muñiz Salazar, por su apoyo y orientación durante la realización de mi trabajo de tesis. Su paciencia, disponibilidad y constante disposición para responder mis preguntas y brindarme su valiosa orientación, fueron esenciales para que pudiera completar este trabajo.

A Ricardo Perea, mi compañero de posgrado, por su amistad, colaboración y motivación durante este proceso. Agradezco sus comentarios, ideas y sugerencias que enriquecieron mi trabajo y lo hicieron más completo. Gracias por compartir este camino conmigo.

También quiero agradecer a todos los participantes de mi estudio, sin su colaboración este trabajo no hubiera sido posible. Y a todas las personas que, de alguna manera, me apoyaron y alentaron durante esta etapa de mi formación académica. Gracias por creer en mí y por motivarme a seguir adelante.

Por último y no menos importante, quisiera dedicar este logro a mi familia, a quienes tanto amo y admiro, y que han sido mi fuente de inspiración y motivación en la vida. Gracias por su amor incondicional y por ser mi mayor estímulo para superarme día a día.

Resumen

La tuberculosis (TB), causada por la bacteria *Mycobacterium tuberculosis*, es una de las principales causas de muerte por enfermedad infecciosa en el mundo. El tratamiento farmacológico para la TB farmacosensible es una combinación de cuatro fármacos durante seis meses (tratamiento de primera línea). En el caso de TB farmacorresistente, se utiliza otra combinación de fármacos (tratamiento de segunda línea), con una duración de nueve a 18 meses o, en algunos casos, hasta 24 meses.

La farmacorresistencia es un problema de salud global en la lucha contra la TB. El diagnóstico actual se realiza mediante pruebas de susceptibilidad a los fármacos por medio de cultivo microbiológico o pruebas rápidas moleculares. Recientemente, se han realizado diversos estudios de secuenciación de genoma completo y técnicas de inteligencia artificial, como *Machine Learning*, para lograr una clasificación del perfil de farmacorresistencia. En este estudio se realizó un preprocesamiento del archivo VCF realizado por CRyPTIC [45] y la información de farmacorresistencia de aislados clínicos de *M. tuberculosis.* Se implementaron cuatro modelos de clasificación en el conjunto de datos original y en el conjunto de datos con reducción de variantes por medio de un análisis de componentes principales. En el caso del modelo XGBC se aplicó una reducción arbitraria considerando las características más importantes para lograr la clasificación en el conjunto de datos original. Estos modelos fueron entrenados y puestos a prueba. Al realizar una comparación de los resultados, se observó que el modelo XGBC en el conjunto de datos original fue el que mejor desempeño obtuvo con una sensibilidad de 0.97, 0.90 y 0.94, una especificidad de 0.97, 0.99 y 0.96, y un F1-Score de 0.93, 0.94 y 0.92 para etambutol, isoniazida y rifampicina, respectivamente. Utilizar una representación binaria para denotar presencia o ausencia de mutaciones es un enfoque efectivo para entrenar modelos de *Machine Learning* capaces de clasificar satisfactoriamente la resistencia y/o susceptibilidad a etambutol, isoniazida y rifampicina en *M. tuberculosis*.

Palabras Clave: Tuberculosis, *Machine Learning*, Farmacorresistencia, Inteligencia Artificial, XGBC.

Abstract

Tuberculosis (TB) is one of the leading causes of death from infectious disease in the world, caused by the bacterium *Mycobacterium tuberculosis*. Drug treatment for drug-sensitive TB is a combination of four drugs for six months (first-line treatment). In the case of drug-resistant TB, another combination of drugs is used (second-line treatment), with a duration of nine to 18 months, or in some cases up to 24 months.

Drug resistance is a global health problem in the fight against TB, the current diagnosis is performed by drug susceptibility testing by microbiological culture or molecular rapid tests. Recently, several whole genome sequencing studies and artificial intelligence techniques such as Machine Learning have been performed to achieve a classification of the drug resistance profile. In this study, a preprocessing of the VCF file performed by CRyPTIC [45] and the drug resistance information of clinical isolates of *M. tuberculosis* was performed, and four classification models were implemented on the original dataset and the dataset with variant reduction by principal component analysis, and in the case of the XGBC model, an arbitrary reduction considering the most important features to achieve classification on the original dataset. These models were trained and tested, and a comparison of the results showed that the XGBC model on the original dataset performed best with a sensitivity of 0.97, 0.90 and 0.94, specificity of 0.97, 0.99 and 0.96 and an F1-Score of 0.93, 0.94 and 0.92 for ethambutol, isoniazid and rifampicin, respectively. Using a binary representation to denote presence or absence of mutations is an effective approach to train Machine Learning models capable of successfully classifying resistance and/or susceptibility to ethambutol, isoniazid and rifampicin in *M. tuberculosis.*

Keywords: Tuberculosis, Machine Learning, Drug resistance, Artificial Intelligence, XGBC.

Índice

Lista de Tablas

Lista de Figuras

1. INTRODUCCIÓN

La tuberculosis (TB) es una enfermedad infecciosa causada por bacterias del complejo *Mycobacterium tuberculosis*. Si bien la TB afecta principalmente los pulmones, también puede presentarse en el sistema nervioso, huesos, piel, intestinos, genitales y ganglios, entre otros. Se transmite de una persona a otra principalmente por el aire a través de gotículas expulsadas por pacientes con TB, de manera similar a la COVID-19. Aproximadamente un tercio de la población mundial está infectada por la bacteria *M. tuberculosis*. En 2021, 10.6 millones de personas enfermaron de TB y 1.6 millones murieron por esta enfermedad, lo que equivale a 4,500 decesos por día [1].

La Organización Mundial de la Salud (OMS) estimó que, a nivel global, en el 2021 se presentaron 558,000 casos de TB resistente a rifampicina (TB-RR), el fármaco de primera línea más efectivo. De éstos, el 82% era TB multi-farmacorresistente (TB-MDR), es decir, también presentaban resistencia, al menos, al segundo fármaco más importante, isoniacida. Los casos de TB resistente a fármacos se ha incrementado en los últimos 15 años, llegando a generar cepas de *M. tuberculosis* que son resistentes a todos los fármacos existentes para tratar la enfermedad, denominándose cepas extremadamente resistentes (TB-XDR). La TB farmacorresistente (TB-DR) a los medicamentos se transmite de la misma forma que la TB sensible a los medicamentos, sin embargo, es más complicada de tratar y curar, además de que es mucho más caro el tratamiento y su manejo inadecuado puede tener resultados potencialmente mortales [1].

La estrategia tradicional para el diagnóstico de la TB-DR se basa en las pruebas de sensibilidad farmacológica (PSF) a partir del cultivo microbiológico. La principal limitación de estas pruebas es que se requieren de 4 a 6 semanas para obtener resultados. Durante este tiempo, el paciente, al no tener el tratamiento farmacológico correcto, seguirá transmitiendo la bacteria entre la población.

A nivel mundial, sólo el 64% de los casos de TB son diagnosticados, es decir, de los 10 millones de nuevos casos, 3.6 millones de personas se encuentran sin tratamiento y, consecuentemente, contagiando a más personas [2]. Muchos países, incluyendo México, dependen de la baciloscopía para diagnosticar TB, prueba que viene utilizándose desde hace más de 100 años. No obstante, esta técnica no detecta a la TB-DR.

Actualmente, los métodos moleculares tienen ventajas considerables para escalar la gestión y vigilancia programática de la TB-DR, ofreciendo ensayos para un diagnóstico rápido y estandarizado con potencial para un alto rendimiento, así como menos requisitos de bioseguridad en el laboratorio [3]. Estos métodos ofrecen resultados en cuestión de horas con alta sensibilidad y especificidad. El ensayo molecular Xpert MTB/RIF (Cepheid, Sunnyvale, USA) se utiliza para la rápida identificación de *M. tuberculosis* y la detección de resistencia a rifampicina. Este método es rápido, requiriendo sólo dos horas. No obstante, sólo analiza la resistencia al fármaco de rifampicina. La nueva versión Xpert® MTB/XDR detecta mutaciones asociadas a la resistencia a isoniazida (INH), fluoroquinolonas (FLQ), fármacos inyectables de segunda línea (amikacina, kanamicina, capreomicina) y etionamida (ETH) en una sola prueba.

Por otro lado, los recientes avances en la secuenciación de próxima generación (NGS por sus siglas en inglés) de *M. tuberculosis* han permitido incrementar la rapidez del ensayo, de varias semanas a sólo 2-5 días, y reducir su costo. Esto le permitirá al médico tener un perfil de farmacorresistencia más completo en cuestión de horas, en vez de semanas o meses. De esta manera, se tendrá la información necesaria para implementar el tratamiento farmacológico de forma individualizada y oportuna. Consecuentemente, la cadena de transmisión de la enfermedad se cortará de manera más temprana [4].

La NGS es una alternativa para los métodos tradicionales de detección de mutaciones dirigidas. Es una técnica que permite identificar tanto mutaciones

comunes como mutaciones raras que podrían estar asociadas con la resistencia a fármacos antituberculosis [5]. Se han estudiado los genes involucrados en la resistencia a fármacos y se ha demostrado que los polimorfismos de un solo nucleótido (SNPs por sus siglas en inglés), así como las deleciones e inserciones (INDELS) en estos genes generan cepas farmacorresistentes [6].

Una de las formas más prometedoras de abordar el problema de la TB-DR es el uso de modelos de aprendizaje automático y aprendizaje profundo para clasificar muestras de laboratorio de *M. tuberculosis*, también conocidos como aislados de *M. tuberculosis.* Estos modelos se basan en algoritmos que pueden analizar grandes cantidades de datos y encontrar patrones complejos en ellos. Esto les permite identificar con mayor precisión los aislados resistentes a los medicamentos al analizar conjuntos de datos, como la NGS. En los últimos años, específicamente en el estudio de tuberculosis farmacorresistente, se han implementado distintos modelos de aprendizaje automático, desde redes neuronales y árboles de decisión hasta métodos de agrupación y regresión logística. Los modelos de aprendizaje profundo han mostrado resultados prometedores, sin embargo, al aumentar la complejidad de los modelos se requiere una representación más abstracta de las características biológicas [7].

El propósito de la presente investigación consiste en el uso de datos obtenidos a partir del análisis de NGS para el entrenamiento de modelos de *Machine Learning* (ML) con el objetivo de llevar a cabo la clasificación de TB-DR. Con la finalidad de clasificar la resistencia a fármacos como rifampicina (RIF), INH y etambutol (EMB) en aislados de *M. tuberculosis*, se busca lograr un rendimiento superior al obtenido en estudios de asociación de genoma completo (EAGC) a partir de la información correspondiente de mutaciones presentes en este genoma.

1.1. ANTECEDENTES

La detección de resistencia a fármacos antituberculosis, principalmente RIF e INH, ha mejorado significativamente con la introducción de herramientas de diagnóstico molecular, ya que son simples de utilizar y presentan resultados en menos tiempo que una prueba de cultivo convencional. Debido a esto, la recolección de información de TB en pacientes ha mejorado mucho en el transcurso del tiempo. Así, por ejemplo, de los pacientes que fueron confirmados con TB mundialmente en 2012, sólo al 7% se les hicieron pruebas de susceptibilidad a RIF, mientras que en 2019 estas pruebas fueron hechas al 61% de los pacientes [8].

A pesar de los avances tecnológicos y las nuevas pruebas de diagnóstico moleculares, la TB-MDR sigue siendo un problema de salud global. Las pruebas convencionales de diagnóstico y las pruebas de susceptibilidad a fármacos basadas en cultivo requieren semanas, incluso meses, para reportar resultados debido al lento crecimiento *in vitro* de *M. tuberculosis.* Las pruebas moleculares han demostrado susceptibilidad limitada y baja cantidad de medicamentos probados (cinco fármacos antituberculosis). Esto es debido a la poca cantidad de loci (ubicaciones genómicas específicas) detectados por prueba (entre uno y seis) y a que no detectan las variantes más raras de los genes en cada loci, especialmente en inserciones, deleciones y variantes en las regiones promotoras.

Los avances tecnológicos en NGS han permitido implementar nuevos estudios de caracterización algorítmica. Éstos se han realizado en el último par de décadas, dando inicio en el 2005 con un estudio de degeneración macular relacionada con la edad (AMD por sus siglas en inglés) [9]. Las NGS permiten analizar de forma exhaustiva las secuencias de ADN de un individuo y detectar de forma exhaustiva variaciones genéticas presentes en su genoma.

Uno de los enfoques más utilizados para analizar estas variaciones son los estudios de asociación del genoma completo (GWAS por sus siglas en inglés). Los

GWAS analizan miles y millones de variaciones genéticas en las NGS de pacientes que han sido diagnosticados con alguna enfermedad o anomalía, la cual se quiere estudiar más profundamente. Como resultado se obtienen variantes nuevas, genes relacionados con la causa de la enfermedad y nuevos objetivos farmacológicos, entre otros [10]. Estos estudios han tenido bastante éxito para diferentes enfermedades y afecciones como, por ejemplo, el cáncer de pulmón [11], el cáncer de piel [12] y la depresión [13], entre muchas otras.

En 2015 se realizó un estudio por medio de GWAS en *M. tuberculosis* [14], en el cual se analizaron 2,099 aislados y se consideraron 23 genes como candidatos relacionados a la resistencia a fármacos antituberculosis. Al realizar este estudio, fueron capaces de clasificar las distintas mutaciones encontradas como benignas (no confieren resistencia), determinantes de resistencia o sin categorizar. Posteriormente, utilizaron estas clasificaciones para predecir resistencia en aislados nuevos, obteniendo así resultados comprometedores en sensibilidad y especificidad.

En 2021, la OMS publicó el primer catálogo de mutaciones asociadas a la resistencia a fármacos antituberculosis, como RIF, INH, EMB y PZA, entre otros. Se analizaron 38,215 aislados de NGS de *M. tuberculosis*, tras descartar 2,922 secuencias después de pasar los controles de calidad. Dentro del catálogo, sólo el 16% de las mutaciones son INDELS. Se utilizaron muestras obtenidas por el sistema de secuenciación de ADN Illumina, ampliamente utilizado a nivel mundial, que genera NGS en el transcurso de tres a cinco días [15].

TBProfiler y *Mykrobe* son dos aplicaciones informáticas ampliamente reconocidas que permiten predecir la resistencia a fármacos en *M. tuberculosis*, tanto para la primera línea como la segunda. Estas aplicaciones utilizan la lógica empleada en los GWAS. El *TBProfiler* considera 1,325 SNPs e INDELS en 992 posiciones de 31 loci, seis promotores y 25 regiones codificantes [16]. Por otro lado, *Mykrobe* utiliza 1,920 aislados de *M. tuberculosis* y logra identificar la especie y la resistencia a 12 fármacos antituberculosis [17]. Se han realizado evaluaciones del potencial para predecir la

resistencia a fármacos en *M. tuberculosis* con estas dos aplicaciones [18]. Los resultados de las mismas pueden observarse en la Tabla I, en la cual se comparan los resultados de sensibilidad y especificidad obtenidos por *Mykrobe* [22], *TBProfile* [21, 22] y con modelos tanto de *Machine Learning* (ML) [19, 20] *como* de *Deep Learning* (DL) [19, 22].

Por otro lado, el término Inteligencia Artificial (AI por sus siglas en inglés) se utiliza para sistemas que ejecutan tareas que normalmente necesitan intervención humana, tales como toma de decisiones, percepción visual, reconocimiento de voz y traducción de idiomas, entre otras. Así, los métodos de ML son un conjunto de modelos utilizados en AI diseñados para encargarse de resolver tareas concretas, enfocándose en el desarrollo y aplicación de métodos analíticos computacionales para extraer información de conjuntos de datos complejos con énfasis en tareas involucradas en la predicción.

Con una cantidad cada vez mayor de aislados clínicos de *M. tuberculosis* y aislados clínicos de MDR-TB, los métodos de ML han demostrado una mejor adaptación para realizar trabajos de predicción con complejas bases de datos clínicos [23,24]. Los modelos de DL son una técnica avanzada de ML. Están basados en redes neuronales con múltiples capas y parámetros. En el contexto médico, estas redes neuronales profundas se utilizan para extraer y transformar características complejas de los datos médicos, lo que permite una mejor comprensión y análisis de los mismos. Las capas inferiores de la red neuronal aprenden características simples, como los valores de los signos vitales, mientras que las capas superiores aprenden características más complejas, como patrones de enfermedad o predicción de resultados a algún tratamiento. La arquitectura jerárquica de la red neuronal permite una representación potente de las características de los datos médicos, lo que permite la identificación de patrones y relaciones entre ellos [25].

Los trabajos de ML y DL implementados en la investigación de MDR-TB en los últimos años han sido bastante diversos, tanto en los métodos implementados como

en el campo de la investigación; se ha trabajado con bases de datos de imágenes de radiografías de tórax [26, 27, 35] y con la caracterización de genes de farmacorresistencia, cambiando y/o comparando los distintos y variados métodos de ML, entre ellos, *support vector machine* (SVM) [27, 28, 30, 31, 32, 33], *random forest* (RF) [28, 29, 31, 33], *logistic regression* (LR) [28, 33, 34], *product-of-marginals* (PM) [31], *gradient boosting tree* (GBT) [28, 34], *class-conditional Bernoulli mixture model* (CBMM) [31], *k-nearest neighbot* (kNN) [32, 33], *artificial neural networks* (ANN) [27, 32], *neuronal network* (NN) [33] y *algoritmos naive Bayes* (NB) [32][35].

Tanto para *M. tuberculosis* [36] como para fármacos antimicrobianos utilizados en humanos y animales, *Antinobacillus pleuropneumoniae* (APP) [37] y VIH/SIDA [38], entre muchos otros, se han reportado resultados positivos en las predicciones de resistencia a fármacos al utilizar modelos de ML con conjuntos de datos de NGS y sus respectivas pruebas de resistencia a fármacos (DST).

Tabla I. Desempeño de modelos de ML y DL, así como de plataformas Mykrobe y TBProfiler, que predicen la farmacorresistencia molecular en *Mycobacterium tuberculosis*.

Año	Referencia		EMB		INH		RIF	
			S	**E**	**S**	**E**	**S**	**E**
2019	Machine Learning Predicts Accurately *M. tuberculosis* Drug Resistance from Whole Genome Sequencing Data		84.7	94.7	**91.1**	99.1	88.8	**99.6**
2019	Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in *M. tuberculosis* resistance prediction		90.6	85.6	90.3	96.4	**95.4**	97.8
2020	A large-scale evaluation of TBProfiler and Mykrobe for antibiotic resistance prediction in *M. tuberculosis*	Mykrobe	87.5	93.7	88.5	98.3	92.4	98.3
		TBProfiler	**92.5**	92.5	89.5	95.8	91.4	98.3
2021	GenTB: A user-friendly genome-based predictor for tuberculosis resistance powered by machine learning		86.0	92.0	91.0	98.0	93.0	98.0
2022	A modified decision tree approach to improve the prediction and mutation discovery for drug resistance in *M. tuberculosis*		57.1	**98.2**	84.2	**99.2**	86.1	98.5

S = Sensibilidad, E = Especificidad.
Mykrobe: https://www.mykrobe.com/.
TBProfiler: https://tbdr.lshtm.ac.uk/.
Los valores en negrita representan el valor más alto de sensibilidad y especificidad, respectivamente, para cada medicamento.

Los estudios de WGS en *M. tuberculosis* para predecir farmacorresistencia varían tanto en el tipo de análisis como en los modelos de ML utilizados y métricas de desempeño presentadas. El tipo de análisis se ha abordado desde una representación binaria para la presencia y ausencia de mutaciones en genes de resistencia o en genes de todo el genoma, y para las características fisicoquímicas de los aminoácidos resultantes por las bases en cada aislado analizado [39].

Existen técnicas de reducción de dimensionalidad, como el análisis de componentes principales (PCA por sus siglas en inglés), que permiten reducir la cantidad de variables en el conjunto de datos, lo cual facilita tanto su visualización como el análisis y disminuye el tiempo para la construcción de los modelos. El modelo PCA juega un papel muy importante al utilizar modelos de ML, en especial con conjuntos de datos que contengan una gran cantidad de características, como el caso de los datos genómicos, debido al tamaño de las muestras que se pueden llegar a generar analizando SNPs. Este modelo proyecta los datos en un espacio de menor dimensionalidad utilizando la descomposición en valores singulares (SVD por sus siglas en inglés). Los SVD descomponen la matriz de datos en tres matrices más simples: matriz de vectores singulares izquierdos, matriz de valores singulares y matriz de vectores singulares derechos, también conocidos como componentes principales. Estos componentes se utilizan para transformar los datos originales en un nuevo sistema de coordenadas que resalta las direcciones de mayor variabilidad de los datos. Al seleccionar un número reducido de componentes principales es posible reducir la dimensionalidad del conjunto de datos, mientras se retiene una gran cantidad de la información original [40]. A pesar de que el PCA es un método de reducción de dimensionalidad bastante utilizado, podemos encontrar otros métodos estadísticos como t-SNE o *random forest* (RF) e incluso reducir las variables genéticas con criterios no estandarizados al realizar comparaciones con catálogos de mutaciones relacionadas con la farmacorresistencia [7,15].

1.2. JUSTIFICACIÓN

Aunque la TB es tratable y curable en más del 80% de los casos, si no se recibe tratamiento, una gran parte de los casos termina en defunción. El uso de un tratamiento genérico sin realizar pruebas de susceptibilidad a fármacos puede complicar la situación al haber presencia de MDR-TB o RR-TB. En tales casos, es necesario implementar un tratamiento de segunda línea, siendo más costoso, extenso e irritable para el paciente. Además, existe la posibilidad de generar cepas de MDR-TB al utilizar fármacos que no logran curar completamente al paciente.

Conocer de antemano los fármacos eficientes para la TB de cada caso individual disminuiría los tiempos de tratamiento y los costos por medicamento, y reduciría el riesgo de generar cepas MDR-TB.

La prueba rápida molecular Xpert MTB/RIF presenta resultados comprometedores de sensibilidad y especificidad comparados con métodos de cultivo, DST e incluso pruebas del mismo tipo, como TB-LAMP. Esto se debe a que Xpert MTB/RIF se centra en el diagnóstico de la TB y la resistencia a RIF mediante la detección de mutaciones comunes en la región determinante de resistencia (RRDR) del gen *rpoB*. Sin embargo, esta prueba pasa por alto por completo mutaciones menos conocidas, tanto en el gen *rpoB* como en otras secciones del genoma que podrían también generar resistencia.

En los últimos años ha aumentado considerablemente la cantidad de aislados clínicos de *M. tuberculosis* sometidos a WGS. Gracias a esto, se ha enriquecido la información de loci implicados en la resistencia a fármacos antituberculosis. Los modelos de ML y DP ofrecen un enfoque con mayor profundidad que los estudios GWAS para realizar una clasificación de resistencia en nuevos aislados clínicos. La OMS, en su más reciente reporte de tuberculosis a nivel mundial [1], menciona que los estudios de diagnóstico de cultivo de hace más de 100 años se siguen utilizando alrededor del mundo, siendo reemplazados poco a poco por las pruebas rápidas

moleculares y, eventualmente, por estudios de secuenciación para realizar un perfil de resistencia más apropiado.

La implementación de modelos de ML y DP para clasificar farmacorresistencia en aislados de *M. tuberculosis* en México podría tener un impacto significativo en la lucha contra la TB. Además, el uso de estas técnicas puede ayudar a los investigadores a entender mejor la dinámica de la farmacorresistencia en México. La ausencia de estas tecnologías avanzadas de diagnóstico ha sido una limitación en la detección temprana y precisa de la resistencia a los fármacos en pacientes con tuberculosis. La introducción de modelos de ML y DP en este contexto permitiría superar las barreras tradicionales, agilizando el proceso de evaluación de farmacorresistencia y proporcionando a los médicos información valiosa en tiempo real para tomar decisiones informadas sobre el tratamiento.

Además de su impacto directo en la atención médica, la incorporación de estas técnicas también abriría las puertas a la generación de datos epidemiológicos más detallados y contextualizados. Al analizar grandes conjuntos de datos, los investigadores podrían identificar patrones de farmacorresistencia específicos de regiones, subgrupos de pacientes o incluso cepas de *M. tuberculosis*. Esta información podría dar como resultado estrategias preventivas más específicas y la optimización de protocolos de tratamiento adaptados a la realidad de México [41].

1.3. PLANTEAMIENTO DEL PROBLEMA

La TB es una enfermedad infecciosa grave que afecta principalmente a los pulmones, siendo una de las enfermedades infecciosas más mortales del mundo. Se sabe que varias cepas de *M. tuberculosis* son resistentes a los fármacos utilizados en el tratamiento, tanto de primera como de segunda línea, lo cual limita los tratamientos y los hace más pesados para el paciente, llegando a durar incluso 24 meses [42]. Esta situación reclama la importancia de detectar y prevenir una mayor resistencia a los fármacos y así reducir la tasa de mortalidad.

Hoy en día se cuenta con la secuenciación del genoma completo, la cual captura la mutación rara y conocida de los aislados de *M. tuberculosis* que pueden contribuir a la resistencia del fármaco. El diagnóstico de MDR-TB es una prioridad sanitaria mundial. La secuenciación del genoma completo de los aislados clínicos de *M. tuberculosis* promete eludir los largos tiempos de espera y el alcance limitado de la susceptibilidad a los fármacos convencionales [43].

Teniendo un conjunto de datos de *M. tuberculosis* con sus respectivas pruebas de resistencia a fármacos es posible implementar un modelo de ML que, al ser entrenado con datos de esta naturaleza, pueda clasificar nuevos conjuntos de datos y así predecir la resistencia a los fármacos. La capacidad de predecir la resistencia a los fármacos de manera oportuna es crucial para reducir la mortalidad asociada a la TB y prevenir la amplificación de la resistencia a los antibióticos existentes. Por esta razón, es fundamental emplear técnicas avanzadas de análisis de datos, como los modelos de ML y de reducción de dimensionalidad, para obtener resultados más precisos y eficientes en el menor tiempo posible [44].

1.4. OBJETIVO GENERAL

Desarrollar modelos de ML para la predicción de resistencia a fármacos antituberculosis de primera línea a partir del análisis del genoma completo de aislados clínicos de *M. tuberculosis*.

1.4.1. Objetivos específicos

a) Construir una matriz de entrenamiento que contenga la información de presencia o ausencia de mutaciones en el genoma completo de distintas secuencias de *M. tuberculosis* con su respectiva información de farmacorresistencia a, por lo menos, tres fármacos.
b) Diseñar modelos de ML con enfoque en predicción de resistencia a fármacos antituberculosis de primera línea, entrenados en secuencias de mutaciones en el genoma completo de aislados clínicos de *M. tuberculosis*.
c) Evaluar el efecto de la reducción de dimensionalidad en la precisión de los modelos de ML entrenados en el conjunto de datos original.
d) Analizar la correlación entre las mutaciones destacadas por los modelos ML y las mutaciones conocidas por la literatura.

1.5. HIPÓTESIS

Los modelos de ML muestran valores de especificidad y sensibilidad superiores a los estudios de asociación directa de genoma completo y modelos desarrollados en investigaciones recientes que utilizan ML y DL al predecir la farmacorresistencia en el genoma completo de *M. tuberculosis.*

2. METODOLOGÍA

2.1. Matriz de entrenamiento

2.1.1. Base de datos

La matriz de entrenamiento se generó a partir de secuencias de mutaciones de DNA de 12,289 genomas completos de aislados de *M. tuberculosis* con sus respectivas pruebas de susceptibilidad a 13 fármacos obtenidos del consorcio internacional de tuberculosis (CRyPTIC por sus siglas en inglés) [45]. En este estudio se realizó un procesamiento de datos genómicos en muestras clínicas de *M. tuberculosis*, generando así un archivo de llamado de variantes (VCF por sus siglas en inglés) para cada aislado.

2.1.2. Preprocesamiento

El archivo VCF original, realizado por CRyPTIC, contiene información de mutaciones para cada aislado. Sin embargo, su estructura no es óptima para crear una matriz de entrenamiento de manera eficiente. Es necesario realizar ciertas modificaciones para mejorar su utilidad en el proceso de entrenamiento de modelos de ML. Para ello, se llevó a cabo un preprocesamiento para extraer y organizar la información considerando cada mutación como una representación binaria de ausencia o presencia de ésta en el genoma completo de *M. tuberculosis*, en lugar de utilizar una representación de cambio de nucleótido (Figura 1).

El primer paso fue realizar una extracción de los aislados o identificaciones (ID) únicos del archivo VCF mencionado. Para el tipo de análisis realizado se obtuvo un total de 929 IDs. Posteriormente, se realizó la extracción de mutaciones únicas entre todos los IDs. Se consideraron mutaciones únicas a cualquier mutación que estuviera presente en, al menos, uno de los IDs únicos, representando la presencia de la mutación con un “1” y la ausencia con un “0”. De esta manera, se obtuvo un total de 79,256 mutaciones únicas, considerando SNPs e INDELS.

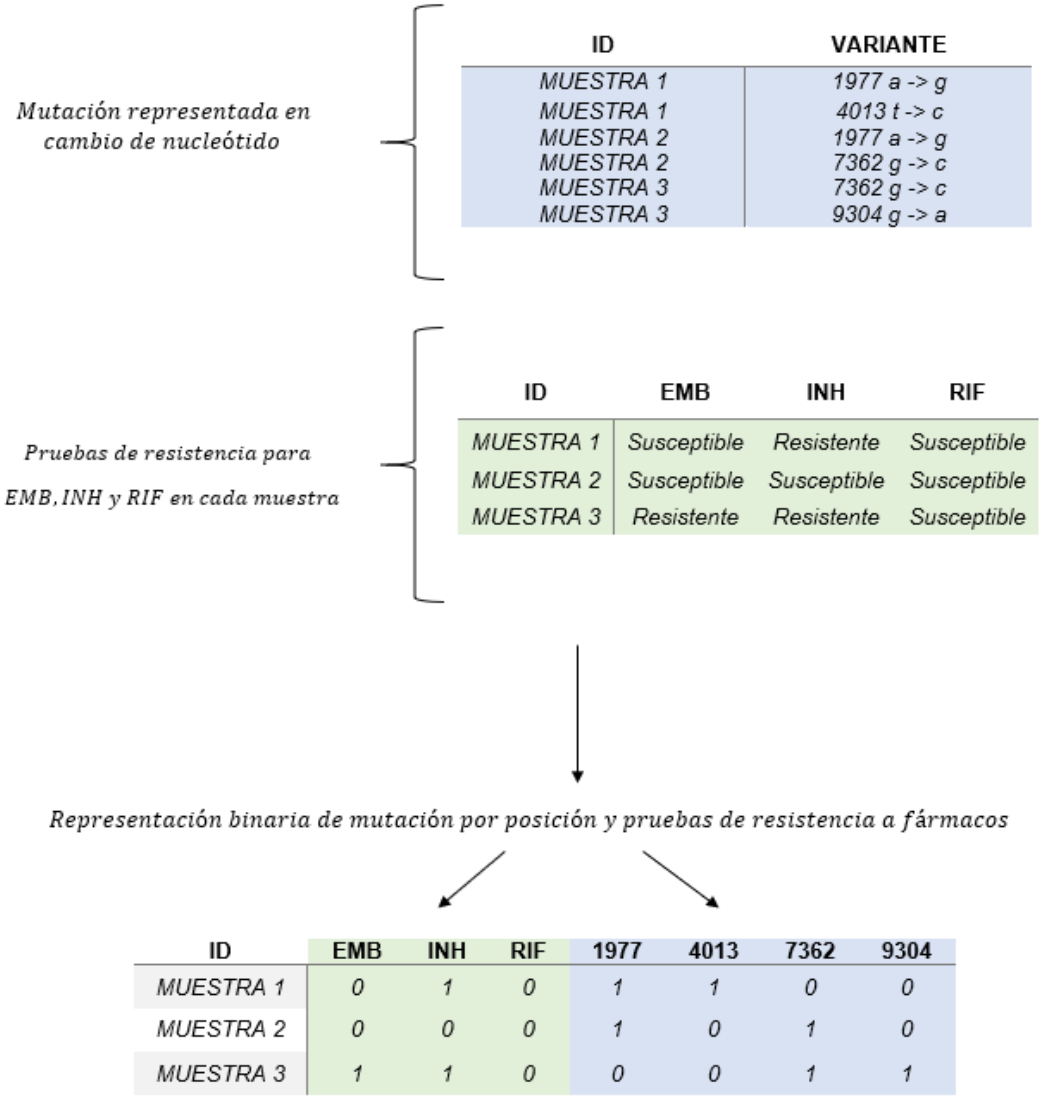


Figura 1: Representación binaria de mutaciones genéticas y perfiles de sensibilidad y resistencia a fármacos en muestras clínicas. EMB = Etambutol, INH = Isoniacida y RIF = Rifampicina.

Se extrajo la información de farmacorresistencia para EMB, INH y RIF para cada uno de los 929 IDs. Se etiquetaron como “0” y “1” para representar la susceptibilidad y resistencia al fármaco, respectivamente. Sin embargo, como algunos de los IDs no contaban con información de susceptibilidad para estos fármacos, se realizó un filtro para ignorar la información ausente, dando un total de 847 IDs. Es importante destacar que esta acción resultó en un porcentaje de eliminación del 8.8% del conjunto de datos inicial. Aunque esta reducción puede parecer significativa, es fundamental priorizar la calidad de los datos sobre la cantidad, ya que un modelo entrenado en datos incompletos o incorrectos podría generar predicciones poco confiables y potencialmente peligrosas en aplicaciones clínicas. El producto del preprocesamiento consistió en una matriz de entrenamiento que incluye 847 aislados (589 susceptibles y 258 resistentes), de los cuales 155 son resistentes a EMB, 244 a INH y 200 a RIF. También se consideraron 79,256 mutaciones entre los 847 aislados, con una media de 1,112 mutaciones por aislado. Esta matriz es referida como el conjunto de datos original (Tabla II).

En el apartado de Anexos del presente estudio se explica con detalle el procesamiento llevado a cabo para crear dos conjuntos de datos de reducción de variantes. Estos conjuntos de datos incluyen uno que utiliza la técnica de PCA y otro que utiliza una reducción arbitraria (RA) en función de las variantes más importantes identificadas durante el entrenamiento del modelo XGBC.

Tabla II: Dimensiones de las matrices de entrenamiento utilizadas para los modelos partiendo del conjunto de datos original, el conjunto de datos reducido con análisis de componentes principales (PCA) y el conjunto de datos con reducción arbitraria (RA).

Conjunto de datos	Muestras	Mutaciones
Original	847	79,259
PCA	847	473
RA	847	159

2.2. Entrenamiento y prueba

2.2.1. Validación cruzada

La validación cruzada o *cross-validation* es una técnica de evaluación de modelos que se utiliza para evaluar el rendimiento de un modelo de aprendizaje automático. En Scikit-learn, la biblioteca de aprendizaje automático de Python, se proporciona la función “cross_val_score ()” para realizar la validación cruzada. Esta validación implica dividir el conjunto de datos en múltiples conjuntos más pequeños llamados “pliegues”, y luego entrenar y evaluar el modelo en diferentes combinaciones de pliegues. Esto ayuda a evaluar la capacidad del modelo para generalizar a datos nuevos no vistos durante la fase de entrenamiento. Para optimizar los parámetros de los modelos se implementó validación cruzada de 10 pliegues (k=10).

La implementación de la validación cruzada para determinar los mejores parámetros con los cuales se entrenó el modelo de aprendizaje automático se describe en la Figura 2.

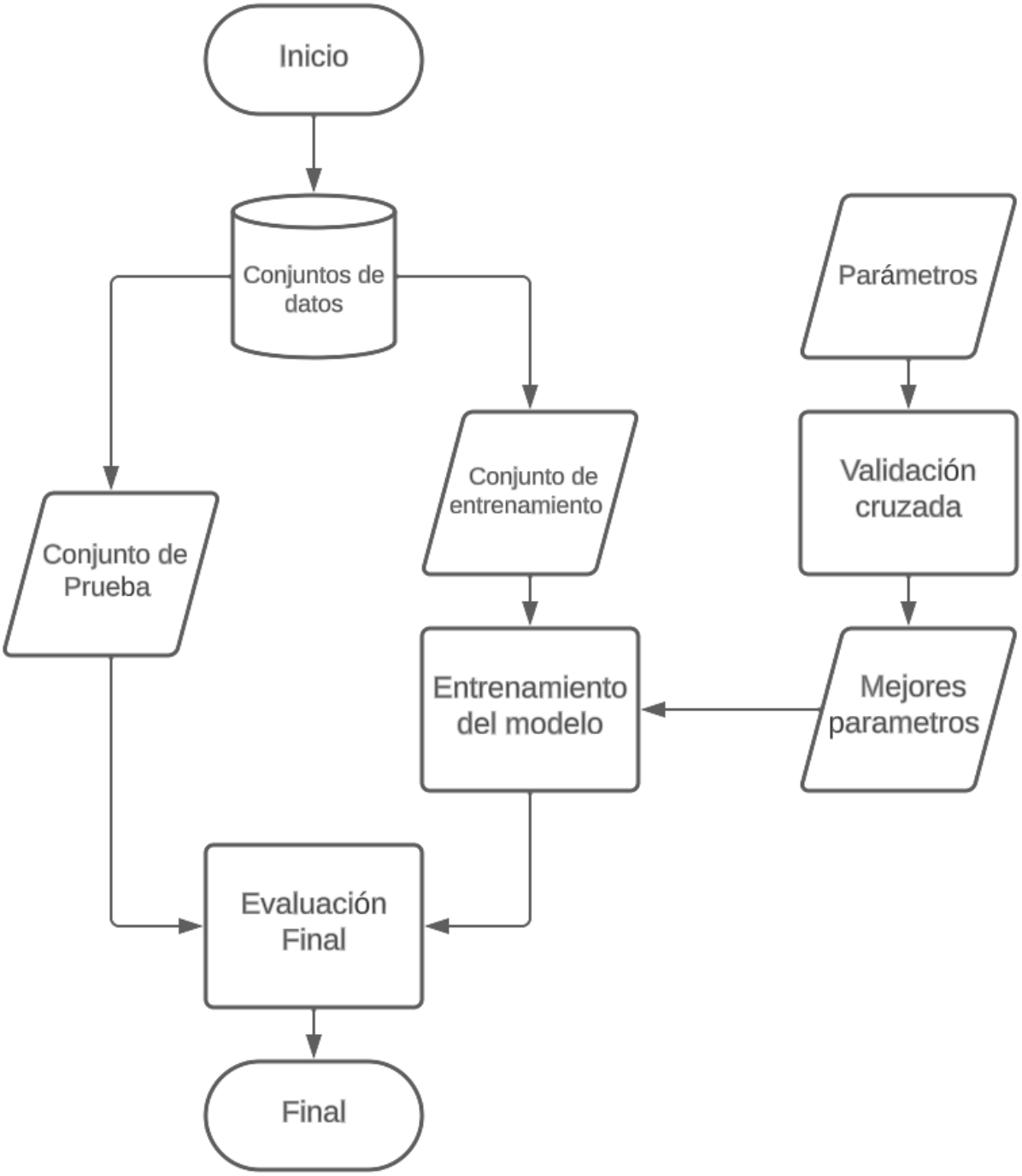


Figura 2: Diagrama de flujo para representar la implementación de validación cruzada para encontrar los mejores parámetros de entrenamiento.

2.2.2. Modelos de *Machine learning*

Los modelos utilizados son: XGBC, LGBC, GradientBoosting y MLPClassifier. El primero de ellos se utilizó para el conjunto de datos original y para la reducción de dimensionalidad con PCA y RA. Para los otros tres también se utilizó el conjunto de datos original pero sólo el PCA para la reducción de dimensionalidad. Los parámetros utilizados se muestran en la Tabla III. En el apartado de Anexos se encuentra una breve descripción de los modelos utilizados.

Tabla III: Parámetros de los modelos de *Machine learning* implementados en el estudio.

Clasificador	**Parámetros**
XGBC	objective = binary:logisctic max_depth = 6 alpha = 15 learning_rate = 0.001 n_estimators = 2000 random_state = 66 scale_pos_weight = 4 objective = binary:logisctic
LGBC	objective = binary:logisctic n_estimators = 2000 random_state = 66 learning_rate = 0.001 scale_pos_weight = 4
GradientBoosting	n_estimators = 2000 learning_rate = 0.001 max_depth = 6 random_state = 66
MLPC	hidden_layer_sizes = (100,100,100,3) activation="relu" solver="adam" alpha=0.001 batch_size = 'auto' learning_rate_init=0.001
	max_iter=1000 random-state=66

2.3. Salidas múltiples

MultiOutputClassifier

Es una clase de la librería de *Machine learning scikit-learn* que se utiliza para entrenar modelos de aprendizaje automático que tienen múltiples salidas. Esto es útil en situaciones en las que se desea predecir más de una etiqueta o variable de salida para cada ejemplo de entrada. En el presente trabajo se utilizó para obtener una predicción de resistencia a los tres fármacos distintos sin necesidad de realizar un entrenamiento distinto para cada fármaco y los distintos modelos de ML utilizados.

2.4. Reducción de dimensionalidad

Análisis de Componentes Principales (PCA)

Teóricamente, el análisis de componentes principales (PCA) es la transformación óptima de los datos dados en términos de mínimos cuadrados. Este análisis es utilizado principalmente para reducir la dimensionalidad de un conjunto de datos. Esto lo logra al concentrar aquellas características que representan una mayor varianza y manteniendo los componentes principales de bajo orden, los cuales pueden clasificarse como los más importantes en el conjunto de datos analizado. En el presente trabajo se utilizaron 470 componentes principales para el análisis, los cuales representan el 0.95 de la varianza acumulada del conjunto de datos (Figura 3).

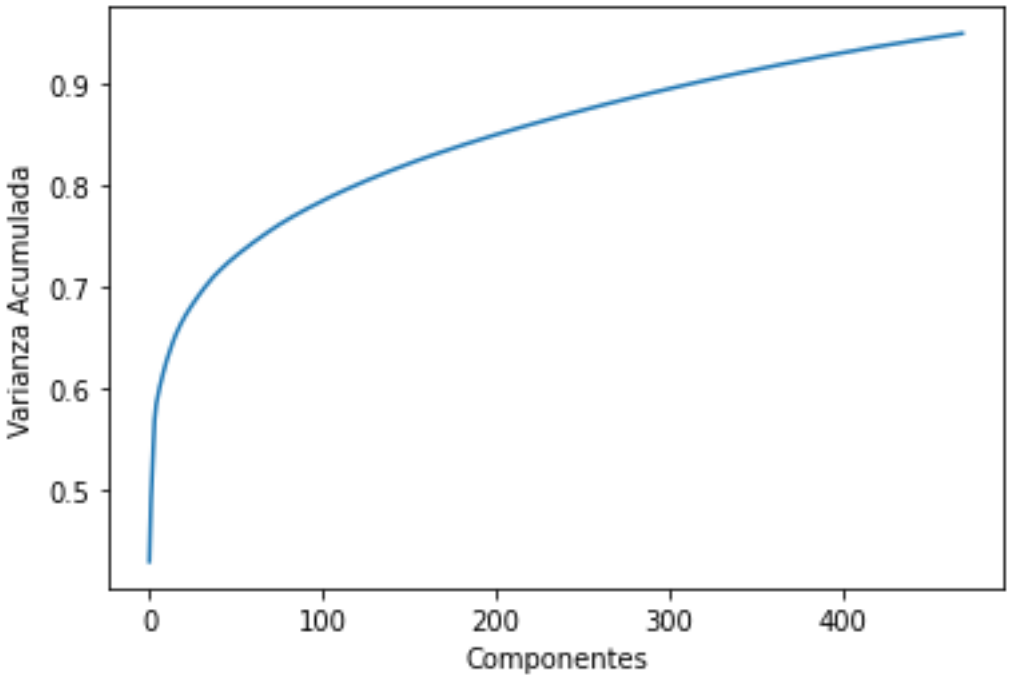


Figura 3: Aumento de varianza acumulada al utilizar más componentes principales. Se utilizaron un total de 470 para representar el 0.95 de varianza.

Reducción de variantes arbitraria

Los modelos utilizados basados en árboles de decisión, XGBC y LGBC, cuentan con un parámetro llamado *importance_type*, el cual puede ser definido de distintas maneras. En el presente trabajo se utilizó la definición *"weight"* que da como resultado el número de veces en el que la característica (mutación en este caso) se utilizó para dividir los datos entre los árboles de decisión. Este paso se utilizó principalmente para realizar un análisis de las mutaciones y una comparación entre las mutaciones más importantes contra las mutaciones de catálogos de resistencia. Esta información se utilizó como un método de reducción de características arbitrarias al realizar un nuevo conjunto de datos, el cual sólo tendrá en cuenta las mutaciones utilizadas en el modelo que presentó los mejores resultados, el XGBC (Tabla IV).

2.5. Métricas de desempeño

Matriz de confusión

La matriz de confusión es una herramienta fundamental en la evaluación del desempeño de los modelos de aprendizaje automático, especialmente en el campo de la medicina. Se utiliza para evaluar la eficacia de un modelo en la detección de enfermedades, para comparar diferentes modelos y, como en este caso, para detectar la resistencia a fármacos.

La matriz de confusión se divide en cuatro categorías: verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN). Los VP son aquellos pacientes (aislados en este caso) que el modelo ha identificado correctamente como enfermos (resistentes a los fármacos). Los VN son aquéllos que el modelo ha identificado correctamente como sanos (susceptibles a los fármacos). Los FP son los que el modelo ha identificado como resistentes, pero, en realidad, son susceptibles. Por último, los FN son aquellos pacientes que el modelo ha identificado como susceptibles, pero, en realidad, son resistentes.

Los resultados esperados para un rendimiento óptimo del modelo clasificador tendrán los resultados concentrados en la diagonal de VP y VN, tal como aparece representado en la Figura 4. Las matrices de confusión para cada uno de los modelos en los distintos conjuntos de datos se pueden consultar en los Anexos del presente trabajo.

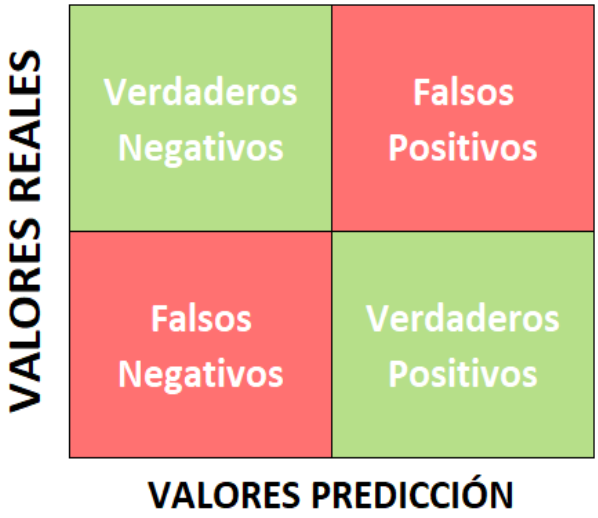


Figura 4: Representación de una matriz de confusión en la que se comparan los valores de predicción contra los valores reales. Se busca que la mayoría de los datos se encuentren en la diagonal verde.

Exactitud

Se utilizó la métrica de exactitud debido a que es comúnmente utilizada en modelos de ML. Sin embargo, dada la presencia de una base de datos desequilibrada entre "resistentes" y "susceptibles", también se emplearon otras métricas para abordar esta situación: sensibilidad, especificidad, precisión y F1-Score.

Sensibilidad

También conocida como tasa de verdaderos positivos, proporciona información sobre la capacidad del modelo para detectar correctamente los casos positivos.

Especificidad

Proporciona información sobre la capacidad del modelo para identificar correctamente los casos negativos.

Precisión

Proporciona información sobre la proporción de predicciones positivas que son verdaderamente positivas.

F1-Score

Combina la precisión y la sensibilidad en una sola medida, brindando una evaluación más equilibrada del rendimiento del modelo.

2.6. Especificaciones de hardware y software

Se utilizó una computadora con el sistema operativo Windows 10, con un Intel® Core™ i5-9300H (8 CPUs) 2.40 GHz, 16.0 GB RAM, NVIDIA GeForce GTX 1650. El lenguaje de programación utilizado fue Python 3.10.11. Los modelos fueron desarrollados en la plataforma *Google Colab (https://colab.research.google.com/).*

3. RESULTADOS

En este estudio se demostró la capacidad de predecir la farmacorresistencia en *M. tuberculosis* al utilizar una matriz de entrenamiento que contiene mutaciones en el genoma completo representadas de forma binaria (1 para presente y 0 para ausente) y la información de farmacorresistencia para cada aislado, también representada de forma binaria (1 para resistente y 0 para susceptible). Es importante destacar que las mutaciones incluidas en la matriz de entrenamiento se encontraron tanto en regiones intergénicas como en regiones codificantes previamente relacionadas con la resistencia a fármacos, así como en genes que no habían sido reportados anteriormente en este contexto. Esto amplía la exploración de las mutaciones y su potencial influencia en la farmacorresistencia, brindando una visión más completa y detallada de los factores genéticos involucrados.

Se realizaron cuatro modelos distintos implementados en ML, XGBC, LGBC, GBC y ANN, con enfoque en la predicción de farmacorresistencia. Estos modelos se han implementado desde distintos ángulos como se ha mencionado en el epígrafe 1.2. Los resultados obtenidos en este trabajo revelaron la capacidad de los modelos de ML para clasificar con precisión la resistencia a los fármacos EMB, INH y RIF. Específicamente, se destaca el rendimiento sobresaliente del modelo XGBC entrenado en el conjunto de datos original, obteniendo resultados prometedores en términos de sensibilidad (0.97, 0.90, 0.94) y especificidad (0.97, 0.99, 0.96) para EMB, INH y RIF, respectivamente.

Al ser un conjunto de datos desbalanceado con la mayoría de la información concentrada en aislados susceptibles a los fármacos, se analizó y comparó la sensibilidad, especificidad y F1-score entre los distintos modelos (Tabla IV). En este sentido, se llevó a cabo una primera comparación entre los cuatro modelos realizados en el conjunto de datos original frente al creado a partir del método de reducción de dimensionalidad PCA. Se puede observar que los modelos de clasificación obtienen mejores resultados al ser entrenados con un conjunto de datos sin reducción de variantes.

El modelo XGBC-RA, el cual considera sólo las características más importantes del modelo XGBC, obtuvo mejores resultados de sensibilidad, 0.95 y 0.94, para INH y RIF, respectivamente, que el modelo XGBC-PCA (0.70, 0.84). Sin embargo, obtuvo resultados bajos de especificidad (0.07, 0.10) lo que disminuye su precisión y potencial de predicción de susceptibilidad a los fármacos (Tabla IV).

Tabla IV: Resultados de los modelos entrenados con el conjunto de datos original, el conjunto de datos reducido por análisis de componentes principales (PCA) y el conjunto de datos con reducción arbitraria (RA) con las características más importantes en el modelo XGBC.

Modelos	Original			PCA			RA		
	EMB	**INH**	**RIF**	**EMB**	**INH**	**RIF**	**EMB**	**INH**	**RIF**
					XGBC				
Sensibilidad	**0.97**	0.90	**0.94**	0.60	0.70	0.84	0.08	**0.95**	**0.94**
Especificidad	0.97	**0.99**	0.96	0.94	0.84	0.93	0.99	0.07	0.10
Precisión	0.89	**0.97**	0.90	0.75	0.60	0.84	0.75	0.26	0.23
F1-Score	**0.93**	**0.94**	0.92	0.66	0.65	0.81	0.15	0.41	0.37
Exactitud	**0.97**	**0.97**	0.96	0.87	0.80	0.91	0.80	0.30	0.29
					LGBC				
Sensibilidad	0.85	0.90	**0.94**	0.42	0.61	0.68	-	-	-
Especificidad	0.98	0.98	0.94	0.97	0.94	0.98	-	-	-
Precisión	0.93	0.95	0.83	0.83	0.79	0.92	-	-	-
F1-Score	0.89	0.93	0.88	0.56	0.69	0.78	-	-	-
Exactitud	0.95	0.96	0.94	0.86	0.85	0.91	-	-	-
					GBC				
Sensibilidad	0.80	0.86	**0.94**	0.31	0.54	0.63	-	-	-
Especificidad	**1**	**0.99**	**0.98**	0.97	0.93	0.97	-	-	-
Precisión	**1**	**0.97**	**0.94**	0.78	0.75	0.88	-	-	-
F1-Score	0.88	0.91	**0.94**	0.44	0.63	0.73	-	-	-
Exactitud	0.95	0.95	**0.97**	0.84	0.83	0.90	-	-	-
					ANN				
Sensibilidad	0.54	0.38	0.44	0.51	0.56	0.52	-	-	-
Especificidad	0.92	0.92	0.92	0.98	0.88	0.97	-	-	-
Precisión	0.65	0.62	0.62	0.90	0.64	0.86	-	-	-
F1-Score	0.59	0.47	0.52	0.65	0.60	0.65	-	-	-
Exactitud	0.84	0.78	0.81	0.88	0.80	0.87	-	-	-

Los valores en negrita representan el valor más alto reportado para cada medicamento entre todos los modelos implementados.

Los datos nulos para los modelos LGBC, GBC y ANN se representan mediante la utilización del símbolo "-". Esto se debe a que sólo se llevaron a cabo pruebas utilizando el modelo XGBC con el conjunto de datos RA (Tabla V).

Uno de los principales motivos de extraer las características más importantes es realizar una comparativa con el catálogo de mutaciones que confieren resistencia a los fármacos. En la Tabla V se observa que una gran parte de las características más importantes para el modelo XGBC están relacionadas con el catálogo de genes de resistencia en *M. tuberculosis* reportados por la OMS [15].

Tras analizar la importancia de los PCA del modelo XGBC-PCA, se ha comprobado que los cinco componentes principales más significativos para la clasificación de cada fármaco se fundamentan en un número restringido de mutaciones de resistencia según lo establecido en el catálogo de la OMS [15]. Específicamente, únicamente se consideran los genes *katG*, *rpoB*, *rpsL* y *gyrA* (Tabla VI).

Tabla V: Características más importantes utilizadas en el modelo XGBC para los genes que confieren farmacorresistencia para etambutol (EMB), isoniacida (INH) y rifampicina (RIF).

XGBC														
EMB					INH					RIF				
Columna	Posición	Gen	Catálogo	Importancia	Columna	Posición	Gen	Catálogo	Importancia	Columna	Posición	Gen	Catálogo	Importancia
14116	761143	rpoB	☑	1729	37159	2155161	katG	☑	2000	14116	761143	rpoB	☑	2000
14105	761095	rpoB	☑	1729	14116	761143	rpoB	☑	1775	14105	761095	rpoB	☑	2000
33509	1955515	Rv1729c	☐	1601	14105	761095	rpoB	☑	1711	14110	761115	rpoB	☑	1737
203	7570	gyrA	☑	1450	29006	1673340	Intergénica	☐	1621	37159	2155161	katG	☑	1510
37159	2155161	katG	☑	1397	49083	2726131	Intergénica	☐	1516	1397	48443	Rv0044c	☐	1206
47752	2637194	Intergénica	☐	1382	14110	761115	rpoB	☑	1299	7817	386402	Rv0318c	☐	1018
14781	799050	Rv0698	☐	1077	8561	423722	hspR	☐	1092	1054	36346	bioF2	☐	922
55982	3096203	Rv2787	☐	1004	25877	1472189	rrs	☑	1031	50450	2801781	PE_PFRS43	☐	849
26321	1501584	Intergénica	☐	929	68243	3878152	rpoA	☐	1026	49083	2726131	Intergénica	☐	825
36805	2154002	katG	☑	883	29993	1728470	Intergénica	☐	838	6289	292515	fadA2	☐	721
50450	2801781	PE_PGRS43	☐	614	5708	262407	Rv0219	☐	717	14517	781662	rpsL	☑	490
74508	4247249	embB	☑	603	70630	4037025	PE_PGRS59	☐	541	14781	799050	Rv0698	☐	424
1054	36346	bioF2	☐	589	69123	3940775	PE_PGRS55	☐	445	50436	2801205	Intergénica	☐	398
14110	761115	rpoB	☑	550	14247	766466	rpoC	☐	423	26411	1506974	rphA	☐	324
3160	132551	PE_PGRS1	☐	504	59556	3347434	Rv2990c	☐	409	9557	485223	fadD30	☐	309
60495	3414761	nrdH	☐	496	21753	1224048	phoH2	☐	345	43282	2357046	PE_PFRS36	☐	294
15351	835654	Rv0745	☐	459	53893	3021823	Intergénica	☐	306	1609	57537	Rv0052	☐	281
7722	381610	Rv0312	☐	424	17431	968250	cspB	☐	217	2382	93274	hycD	☐	275
66257	3741177	PE_PGRS50	☐	347	78223	4377951	eccC2	☐	195	77648	4359079	espK	☐	263
17733	988380	Rv0888	☐	347	6358	295561	fadE5	☐	195	54150	3037139	fadE20	☐	263

Las columnas en el conjunto de datos original representan una posición en el genoma completo de M. tuberculosis.
El gen en el que se encuentran está descrito por su nombre o con la etiqueta "Intergénica" si se encuentran en una zona no codificante.
Se consideró el catálogo de la OMS [15] para comparar si el gen considerado por el modelo es un gen relacionado con resistencia.

Tabla VI: Características más importantes utilizadas en el modelo XGBC-PCA para etambutol (EMB), isoniacida (INH) y rifampicina (RIF). CP = Componente Principal.

XGBC - PCA														
EMB					INH					RIF				
CP	Posición	Gen	Catálogo	Importancia de CP	CP	Posición	Gen	Catálogo	Importancia de CP	CP	Posición	Gen	Catálogo	Importancia de CP
11	4046007	Rv3603c	☐	1715	46	2155168	katG	☑	1878	46	2155168	katG	☑	2242
	484005	fadD30	☐			761155	rpoB	☑			761155	rpoB	☑	
	1529133	Rv1358	☐			454295	Rv0376c	☐			454295	Rv0376c	☐	
	1446733	argS	☐			3173107	Intergénica	☐			3173107	Intergénica	☐	
	1151304	kdpD	☐			3187860	Rv2876	☐			3187860	Rv2876	☐	
0	2177366	fadD31	☐	959	39	1753519	Intergénica	☐	1028	11	4046007	Rv3603c	☐	1199
	3962187	Rv3525c	☐			3308310	Rv2955c	☐			484005	fadD30	☐	
	1849051	lysX	☐			3186860	dipZ	☐			1529133	Rv1358	☐	
	2354791	pafB	☐			2634282	Intergénica	☐			1446733	argS	☐	
	1285001	pimE	☐			1443428	Intergénica	☐			1151304	kdpD	☐	
14	2976564	Intergénica	☐	906	11	4046007	Rv3603c	☐	907	44	2155168	katG	☑	885
	3969423	PPE61	☐			484005	fadD30	☐			2220512	Rv1977	☐	
	2881382	Intergénica	☐			1529133	Rv1358	☐			3186860	dipZ	☐	
	3901234	Intergénica	☐			1446733	argS	☐			761155	rpoB	☑	
	1914779	tyrS	☐			1151304	kdpD	☐			4008747	hsaB	☐	
5	2288085	Rv2042c	☐	789	81	2155168	katG	☑	882	89	2155168	katG	☑	853
	2288844	pncA	☐			340372	PPE3	☐			2528971	Rv2254c	☐	
	2288883	pncA	☐			781687	rpsL	☑			1313337	Intergénica	☐	
	2288867	pncA	☐			2220512	Rv1977	☐			3878547	Intergénica	☐	
	2288856	pncA	☐			3186860	dipZ	☐			1313338	Intergénica	☐	
30	2155168	katG	☑	757	15	1164571	Intergénica	☐	858	30	2155168	katG	☑	790
	761155	rpoB	☑			2168742	PPE35	☐			761155	rpoB	☑	
	3189481	Intergénica	☐			2700	dnaN	☐			3189481	Intergénica	☐	
	4318188	Intergénica	☐			8147	gyrA	☑			4318188	Intergénica	☐	
	628174	galE3	☐			1444134	Rv1290c	☐			628174	galE3	☐	

Las columnas en el conjunto de datos original representan una posición en el genoma completo de M. tuberculosis.
El gen en el que se encuentran está descrito por su nombre o con la etiqueta "Intergénica" si se encuentran en una zona no codificante.
Se consideró el catálogo de la OMS [15] para comparar si el gen considerado por el modelo es un gen relacionado con resistencia

4. DISCUSIÓN

En este estudio se obtuvieron resultados significativos que respaldan la viabilidad de utilizar modelos de aprendizaje automático, como XGBC, para predecir la farmacorresistencia en *M. tuberculosis* basándose en información binaria de mutaciones en el genoma completo. Los valores obtenidos en este estudio, como la sensibilidad (0.91, 0.90, 0.94) y especificidad (0.97, 1.00, 0.96) para la resistencia a EMB, INH y RIF, respectivamente, son comparables a los obtenidos en estudios recientes de ML, DL y modelos basados en GWAS (Sensibilidad: 0.92, 0.91, 0.95, Especificidad: 0.98, 0.99, 0.99) [5, 18, 20, 21, 22]. Estos resultados resaltan la utilidad de los modelos de ML en este campo de investigación al igual que el utilizar matrices de entrenamiento conformadas por representación binaria de presencia o ausencia de mutación en genoma completo de *M. tuberculosis*.

Sin embargo, también se observó una disminución en el desempeño de los modelos entrenados con el conjunto de datos con reducción de dimensionalidad (Sensibilidad: 0.60, 0.70, 0.84; Especificidad: 0.94, 0.84, 0.93) en comparación con los entrenados con el conjunto de datos original. Se sugiere que esto puede atribuirse a limitaciones y complejidades asociadas con la reducción de dimensionalidad utilizando PCA, especialmente en datos binarios. Existe el riesgo de perder información importante durante este proceso, lo cual puede afectar la capacidad predictiva de los modelos. La combinación lineal de características maximizando la varianza de los datos obtenida mediante PCA puede no coincidir con las características más relevantes en el conjunto de datos original. Sin embargo, se han estudiado versiones distintas del PCA tal como Sparse PCA, el cual podría potencialmente proporcionar resultados más favorables en este contexto. Al considerar la naturaleza binaria de los datos y al buscar una combinación lineal de características que también sea dispersa, podría capturar de manera más efectiva las relaciones y patrones en los datos reducidos. Esta técnica podría mitigar algunas de las complicaciones encontradas con la versión estándar de PCA, al preservar

características más discriminativas y reducir el riesgo de pérdida de información para la capacidad predictiva de los modelos [46-48].

El modelo XGBC-RA, el cual se utilizó como un conjunto de datos con datos reducidos sin aplicar métodos estadísticos, al ser entrenado con las características más importantes del modelo XGBC, presentó resultados poco favorables. Esto puede deberse a un sobreajuste del modelo al utilizar un conjunto de datos reducido que no captura el espectro completo de la variabilidad presente. Es importante considerar que, aunque la función "importance_type" proporciona las características utilizadas en los árboles de decisión del modelo XGBC, otras funciones del modelo, como el "split finding", pueden verse influenciadas por características no incluidas en los árboles de decisión, lo que puede afectar el proceso de entrenamiento y predicción [49].

Se destaca la relevancia de las mutaciones en las regiones intergénicas para predecir la resistencia a fármacos en *M. tuberculosis*. Investigaciones previas han evidenciado la relación entre las mutaciones en la región intergénica embC-embA y la resistencia al fármaco EMB [50] y relación entre la región intergénica oxyR-ahpC y la resistencia al fármaco INH [51]. Al examinar aislados que presentan resistencia a EMB, se observaron coeficientes de correlación que confirman una asociación entre la presencia de mutaciones en zonas intergénicas y la resistencia al fármaco [50]. En este estudio, se observó que el modelo XGBC considera zonas intergénicas en *M. tuberculosis* para clasificar la farmacorresistencia a EMB, INH y RIF, lo cual es importante, ya que estas regiones pueden contener elementos reguladores, como promotores, que influyen en la expresión de los genes relacionados. Estos hallazgos destacan la importancia de analizar el genoma completo en la predicción de la resistencia a fármacos, más allá del análisis de genes conocidos.

Existe la posibilidad de que las muestras resistentes a los fármacos utilizadas para entrenar el modelo provengan de una zona geográfica diferente, donde *M. tuberculosis* presenta una mutación intergénica particular de manera más frecuente

o "común". En estudios previos se ha buscado la mejora de predicción de los linajes de *M. tuberculosis* analizando distintos métodos de genotipado molecular tales como la tipificación de oligonucleótidos espaciadores (spoligotyping) y el numero variable de repeticiones en tándem (MIRU-VNTR por sus siglas en ingles). Estos métodos se basan en la comparación de secuencias repetitivas en *M. tuberculosis*, denominadas microsatélites, para lograr la discriminación de los linajes. De igual forma se han estudiado los SNPs y mutaciones sinónimas como posibles marcadores distintivos [52-55]. Este aspecto puede dar una interpretación distinta de los resultados, ya que sugiere que la inclusión de mutaciones intergénicas en el modelo puede estar más relacionada con las características de las muestras de resistencia utilizadas en el entrenamiento que con una correlación intrínseca entre estas mutaciones y la resistencia a los fármacos.

5. CONCLUSIONES

Los resultados obtenidos en este estudio indican que es factible utilizar una matriz de entrenamiento conformada por información de resistencia a fármacos de primera línea y datos de mutaciones en genes y zonas intergénicas en el genoma completo de *M. tuberculosis* para entrenar modelos de ML. Además, los modelos entrenados con esta matriz de entrenamiento binaria lograron una alta precisión en la predicción de la resistencia a los fármacos de primera línea, lo que sugiere que la matriz fue efectiva en la identificación de patrones y características de las mutaciones asociadas con la resistencia a los fármacos de *M. tuberculosis*.

Los resultados demuestran que el modelo XGBC tiene un desempeño igual o superior al obtenido por métodos de asociación directa y modelos de ML reportados en investigaciones previas. Esto sugiere que el modelo XGBC es una opción prometedora para clasificar satisfactoriamente la resistencia a fármacos antituberculosis. No obstante, es importante destacar que existen distintos abordajes que podrían mejorar aún más el desempeño del modelo, lo que representa una oportunidad interesante para futuras investigaciones en este tema. Por lo tanto, se recomienda que los investigadores continúen explorando las posibilidades de mejora del modelo XGBC y de otros puntos de abordaje con respecto a las características de entrenamiento para mejorar la precisión de los resultados en este campo de estudio.

Este trabajo de investigación representa un avance significativo en la lucha contra la tuberculosis resistente a los fármacos. Los modelos de aprendizaje automático realizados son capaces de identificar patrones complejos y predecir resultados con alta precisión, lo que puede tener implicaciones importantes en la mejora de los tratamientos y en la prevención de la propagación de la TB-MDR.

A pesar de que este proceso conlleva un aumento en los costes de diagnóstico debido al proceso de secuenciación y obtención de variantes genéticas, sus

beneficios son notables al impulsar la precisión y sensibilidad, así como mejorar y disminuir el tiempo de tratamiento en el caso de TB-MDR. Al recaudar información genómica también se contribuye a la identificación de brotes epidémicos antes de que alcancen proporciones mayores. Al comparar esta información en distintos pacientes es posible determinar si los casos están vinculados a una fuente común y si representan un brote localizado.

6. REFERENCIAS

[1] World Health Organization (WHO). Global tuberculosis report 2022 [Internet]. Geneva: World Health Organization; 2022 [consultado febrero 10 del 2023]. Disponible en: https://www.who.int/publications/i/item/9789240061729

[2] Sistema de Información en Salud. DGIS [Internet]. Subsistema Epidemiológico y Estadístico de Defunciones; 2020 [consultado noviembre 10 del 2022]. Disponible en: http://www.dgis.salud.gob.mx/contenidos/sinais/s_seed.html

[3] World Health Organization (WHO). Global tuberculosis report 2021 [Internet]. Geneva: World Health Organization; 2021 [consultado octubre 10 del 2022] Disponible en: https://www.who.int/publications/i/item/9789240037021

[4] Unissa AN, Subbian S, Hanna LE, Selvakumar, N. Overview on mechanisms of isoniazid action and resistance in *Mycobacterium tuberculosis*. Infect Genet Evol [Internet], 2016 [consultado octubre 19 del 2022]; 45: p. 474-492. Disponible en: 10.1016/j.meegid.2016.09.004

[5] Murphy SG, Smith C, Lapierre P, Shea J, Patel K, Halse TA, Dickinson M, Escuyer V, Rowlinson MC, Musser KA. Direct detection of drug-resistance *Mycobacterium tuberculosis* using targeted next generation sequencing. Front Public Health [Internet], 2023 [consultado junio 15 del 2023]; 11(1): p. 1-11. Disponible en: 10.3389/fpubh.2023.1206056

[6] Palomino JC, Martin, A. Drug resistance mechanisms in *Mycobacterium tuberculosis*. Antibiotics [Internet] 2014 [consultado octubre 19 del 2022]; 3(3): p. 317–340. Disponible en: 10.3390/antibiotics3030317

[7] Perea-Jacobo R, Paredes-Gutiérrez GR, Guerrero-Chevannier MA, Flores DL, Muñiz-Salazar R. Machine Learning of the Whole Genome Sequence of *Mycobacterium tuberculosis*: A Scoping PRISMA-Based Review. Microorganisms [Internet], 2023 [consultado febrero 10 del 2023]; 11(8): p. 1872. Disponible en: 10.3390/microorganisms11081872

[8] World Health Organization (WHO). Global tuberculosis report 2019 [Internet]. Geneva: World Health Organization; 2019 [consultado noviembre 10 del 2022]. Disponible en: https://www.who.int/publications-detail-redirect/9789241565714

[9] Klein RJ, Ziess C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Barnstable C, Hoh J. Complement factor H polymorphism in age-related macular degeneration. Science [Internet], 2005 [consultado octubre 19 del 2022]; 308(5720): p. 385–389. Disponible en: 10.1126/science.1109557

[10] Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet [Internet], 2019 [consultado octubre 19 del 2022]; 20(8): p. 467–484. Disponible en: 10.1038/s41576-019-0127-1

[11] Weissfeld JL, Lin Y, Lin HM, Kurland BF, Wilson DO, Fuhrman CR, Pennathur A, Romkes M, Nukui T, Yuan JM, Siegfried JM, Diergaarde B. Lung cancer risk prediction using common SNPs located in GWAS-identified susceptibility regions. J Thorac Oncol [Internet], 2015 [consultado octubre 10 del 2022]; 10(11): p. 1538–1545. Disponible en: 10.1097/JTO.0000000000000666.

[12] Roberts MR, Asgari MM, Toland, AE. Genome-wide association studies and polygenic risk scores for skin cancer: clinically useful yet?. Br J Dermatol [Internet], 2019 [consultado octubre 10 del 2022]; 181(6): p. 1146–1155. Disponible en: 10.1111/bjd.17917

[13] Xie T, Stathopoupou MG, Andrés F, Siest G, Murray H, Martin M, Cobaleda J, Delgado A, Lamont J, Peñas E, LLerena A Visvikis S. VEGF-related polymorphisms identified by GWAS and risk for major depression. Transl. Psychiatry [Internet], 2017 [consultado octubre 10 del 2022] 7(3): p. 3. Disponible en: 10.1038/tp.2017.36

[14] Walker TM, Kohl TA, Omar SV, Hedge J, Elias CDO, Bradley P, Iqbal Z, Feuerriegel S, Niehaus KE, Wilson DJ, Clifton DA, Kapatai G, Cavalcanti CLC, Bowden R, Drobniewski FA, Allix C, Gaudin C, Parkhill J, Diel R, Supply P, Crook D, Smith EG, Walker AS, Ismail N, Niemann S, Peto TEA. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: A retrospective cohort study. Lancet Infect. Dis. [Internet], 2015 [consultado octubre 15 del 2022]; 15(10): p. 1193–1202. Disponible en: 10.1016/S1473-3099(15)00062-6

[15] World Health Organization. Catalogue of mutations in *Mycobacterium tuberculosis* complex and their association with drug resistance [Internet]. Geneva: World Health Organization; 2021 [consultado febrero 10 del 2021]. Disponible en: https://www.who.int/publications/i/item/9789240028173

[16] Coll F, McNerney R, Preston MD, Guerra JA, Warry A, Hill G, Mallard K, Nair Mridul, Miranda A, Alves A, Perdigão J, Viveiros M, Portugal I, Hasan Z, Hasan R, Glynn JR, Martin N, Pain A, Clark TG. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. [Internet], 2015 [consultado noviembre 18 del 2022]; 7(1): p. 51. Disponible en: 10.1186/s13073-015-0164-0

[17] Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, Earle S, Pankhurst LJ, Anson L, Cesare M, Piazza P, Votintseva AA, Golubchik T, Wilson DJ, Wyllie DH, Diel R, Niemann S, Feuerriegel S, Kohl TA, Ismael N, Omar SV, Smith EG, Buck D, McVean G, Walker AS, Peto TEA, Crook DW, Iqbal Z. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. Nat. Commun. [Internet], 2015 [consultado diciembre 5 del 2022]; 6: 1-14 Disponible en: 10.1038/ncomms10063

[18] Mahé P, El Azami M, Barlas P, Tournoud M. A large-scale evaluation of TBProfiler and Mykrobe for antibiotic resistance prediction in *Mycobacterium tuberculosis*. PeerJ [Internet], 2019 [consultado diciembre 10 del 2022]; 7(1): p. 1-21 Disponible en: 10.7717/peerj.6857

[19] Chen ML, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, Kohane IS, Beam A, Farhat M. Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in *Mycobacterium tuberculo*sis resistance prediction. EBioMedicine [Internet], 2019 [consultado diciembre 10 del 2022]; 43: p. 356–369. Disponible en: 10.1016/j.ebiom.2019.04.016

[20] Deelder W, Christakoudi S, Phelan J, Benavente ED, Campino S, McNerney R, Palla L, Clark TG. Machine learning predicts accurately *Mycobacterium tuberculosis* drug resistance from whole genome sequencing data. Front. Genet. [Internet], 2019 [consultado diciembre 10 del 2022]; 10: p. 922. Disponible en: 10.3389/fgene.2019.00922

[21] Deelder W, Napier G, Campino S, Palla L, Phelan J, Clark TG. A modified decision tree approach to improve the prediction and mutation discovery for drug resistance in *Mycobacterium tuberculosis*. BMC Genomics [Internet], 2022 [consultado diciembre 10 del 2022]; 23(1): p. 46. Disponible en: 10.1186/s12864-022-08291-4

[22] Gröschel MI, Owens M, Freschi L, Vargas R, Marin MG, Phelan J, Iqbal Z, Dixit A, Farhat MR. GenTB: A user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. Genome Med. [Internet], 2021 [consultado diciembre 10 del 2022]; 13(1): p. 138. Disponible en: 10.1186/s13073-021-00953-4

[23] Zabeti H, Dexter N, Safari AH, Sedaghat N, Libbrecht M, Chindelevitch L. INGOT-DR: an interpretable classifier for predicting drug resistance in *M. tuberculosis*. Algorithms Mol Biol [Internet], 2021 [consultado febrero 10 del 2023]; 16(1): p. 17. Disponible en: 10.1186/s13015-021-00198-1

[24] Dande P, Samant P. Acquaintance to Artificial Neural Networks and use of artificial intelligence as a diagnostic tool for tuberculosis: A review. Tuberculosis [Internet], 2018 [consultado febrero 15 del 2023]; 108: p. 1–9. Disponible en: 10.1016/j.tube.2017.09.006

[25] 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA 2018). Pune, India: IEEE; 2018 [consultado octubre 15 del 2022]; 3: p. 1811. Disponible en: 10.1109/ICCUBEA.2018.8697366.

[26] Heo SJ, Kim Y, Yun S, Lim SS, Kim J, Nam CM, Park EC, Junh I, Yoon JH. Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Radiographs in Annual Workers' Health Examination Data. Int. J. Environ. Res. Public. Health [Internet]. 2019 consultado 20 de noviembre del 2022]; 16(2): p. 250. Disponible en: 10.3390/ijerph16020250

[27] Jaeger S, Juarez OH, Candemir S, Poostchi M, Yang F, Kim L, Ding M, Folio LR, Antani S, Gabrielian A, Hurt D, Rosenthal A, Thoma G. Detecting drug-resistant tuberculosis in chest radiographs. Int. J. Comput. Assist. Radiol. Surg [internet]. 2018 [consultado 20 de noviembre del 2022]; 13(12): p. 1915-1925. Disponible en: 10.1007/s11548-018-1857-9

[28] Kouchaki S., Yang Y, Walker TM, Walker AS, Wilson DJ, Peto TEA, Crook DW, CRyPTIC Consortium, Clifton DA. Application of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics [Internet]. 2019 [consultado 20 de noviembre del 2022]; 35(13): p. 2276-2282. Disponible en: 10.1093/bioinformatics/bty949

[29] Kouchaki S, Yang Y, Lapachelle A, Walker TM, Walker AS, CRyPTIC Consortium, Peto TEA, Crrok DW, Clifton DA. Multi-label random forest model for tuberculosis drug resistance classification and mutation ranking. Front. Microbiol [Internet]. 2020 [consultado 20 de noviembre de 2022]; 11: p 667. Disponible en: 10.3389/fmicb.2020.00667

[30] Kavvas ES, Catiou E, Mih N, Yurkovich JT, Seif Y, Dillon N, Heckmann D, Anand A, Yang L, Nizet V, Monk JM, Palsson BO. Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. Nat. Commun [Internet]. 2018 [consultado 20 de noviembre del 2022]; 9(1):4306. Disponible en: https://doi.org/10.1038/s41467-018-06634-y

[31] Yang Y, Niehaus K, Walker TM, Iqbal Z, Walker AS, Wilson DJ, Peto TEA, Crook DW, Smith EG, Zhu T, Clifton DA. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. Bioinformatics [Internet]. 2018 [consultado 20 de noviembre del 2022]; 34(10): p. 1666-1671. Disponible en: 10.1093/bioinformatics/btx801

[32] Jamal S, Khubaib M, Gangwar R, Grover S, Grover A, Hasnain SE. Artificial Intelligence and Machine learning based prediction of resistant and susceptible mutations in *Mycobacterium tuberculosis*. Sci. Rep [Internet]. 2020 [consultado 20 de noviembre del 2022]; 10(1):5487. Disponible en: 10.1038/s41598-020-62368-2

[33] Duffy FJ, Thompson EG, Scriba TJ, Zak DE. Multinomial modelling of TB/HIV co-infection yields a robust predictive signature and generates hypotheses about the HIV+TB+ disease state. PLoS One [Internet]. 2019 [consultado 20 de noviembre del 2022]; 14(7): p. 1-17. Disponible en: https://doi.org/10.1371/journal.pone.0219322

[34] Deelder W, Christakoudi S, Phelan J, Benavente ED, Campino S, McNerney R, Palla L, Clark TG. Machine learning predicts accurately *Mycobacterium tuberculosis* drug resistance from whole genome sequencing data. Front. Gen. 2019 [consultado 20 de noviembre del 2022]; 10(1): p. 922. Disponible en: https://doi.org/10.3389/fgene.2019.00922

[35] Guerrero-Chevannier MA, Perea-Jacobo R, Flores DL, Muñiz-Salazar R. Uso de machine learning como apoyo al diagnóstico del complejo *Mycobacterium tuberculosis*: Una revisión sistemática [Internet] 2020 [consultado 20 de noviembre de 2022]; 7[1]: p 424-430. Disponible en: 10.24254/CNIB.20.54

[36] Safari AH, Sedaghat N, Zabeti H, Forna A, Chindelevitch L, Libbrecht M. Predicting drug resistance in *M. tuberculosis* using a long-term recurrent convolutional network. En: Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. Gainesville, Florida: Association for Computing Machinery; 2021. P. 1-10 Disponible en: 10.1145/3459930.3469534.

[37] Liu Z, Deng D, Lu H, Sun J, Lv L, Li S, Peng G, Ma X, Li J, Li Z, Rong T, Wang G. Evaluation of Machine Learning Models for Predicting Antimicrobial Resistance of *Actinobacillus pleuropneumoniae* From Whole Genome Sequences. Front. Microbiol. [Internet] 2020 [consultado octubre 10 del 2022]; 11: p. 1-7 Disponible en: 10.3389/fmicb.2020.00048

[38] Singh Y. Machine learning to improve the effectiveness of ANRS in predicting HIV drug resistance. J. Healthc. Inform. Res. [Internet] 2017 [consultado octubre 10 del 2022]; 23(4): p. 271–276. Disponible en: 10.4258/hir.2017.23.4.271

[39] Chowdhury AS, Khaledian E, Broschat SL. Capreomycin resistance prediction in two species of *Mycobacterium* using a stacked ensemble method. J. Appl. Microbiol. [Internet] 2019 [consultado octubre 10 del 2022]; 127(6): p. 1656–1664. Disponible en: 10.1111/jam.14413

[40] Abraham G, Michael I. Fast Principal Component Analysis of Large-Scale Genome-Wide Data. *PLoS One* [Internet] 2014 [consultado noviembre 5 del 2022]; 9(4): p. 1-5 Disponible en: 10.1371/journal.pone.0093766

[41] Asare P, Asante-Poku A, Osei-Wusu S, Otchere ID, Yeboah-Manu D. The Relevance of Genomic Epidemiology for Control of Tuberculosis in West Africa. Front Public Health [Internet] 2021 [consultado febrero 15 del 2023]; 9: 1-17 Disponible en: 10.3389/fpubh.2021.706651

[42] Gygli SM, Borrell S, Trauner A, Gagneux S. Antimicrobial resistance in *Mycobacterium tuberculosis*: Mechanistic and evolutionary perspectives. FEMS Microbiol. Rev. [Internet] 2017 [consultado octubre 10 del 2022]; 41(3): p. 354–373. Disponible en: 10.1093/femsre/fux011

[43] Chen ML, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, Kohane IS, Beam A, Farhat M. Deep learning predicts tuberculosis drug resistance status from genome sequencing data. EBioMedicine [Preprint] 2019 [consultado febrero 10 del 2023]; p. 1-20 Disponible en: 10.1101/275628

[44] Müller SJ, Meraba RL, Dlamini GS, Mapiye DS. First-line drug resistance profiling of Mycobacterium tuberculosis: a machine learning approach. AMIA Annu Symp Proc [Internet] 2022 [consultado junio 10 del 2023]; 2021: p. 891-899. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8861754/

[45] The CRyPTIC Consortium. A data compendium associating the genomes of 12,289 *Mycobacterium tuberculosis* isolates with quantitative resistance phenotypes to 13

antibiotics. PLoS Biol [Internet] 2022 [consultado octubre 10 del 2022]; 20(8): p. 25. Disponible en: 10.1371/journal.pbio.3001721

[46] Lee S, Huang JZ, Hu J. Sparse logistic principal components analysis for binary data. Ann Appl Stat [Internet] 2010 [consultado octubre 10 del 2022]; 4(3): p. 1579-1601. Disponible en: 10.1214/10-AOAS327SUPP

[47] Landgraf AJ, Lee Y. Dimensionality reduction for binary data through the projection of natural parameters. J. Multivariate Anal. [Internet] 2020 [consultado octubre 10 del 2022]; 180: p. 18. Disponible en: 10.1016/j.jmva.2020.104668

[48] De Leeuw J. Principal component analysis of binary data by iterated singular value decomposition. Comput. Stat. Data Anal. [Internet] 2006 [consultado noviembre 14 del 2022]; 50(1): p. 21–39. Disponible en: 10.1016/j.csda.2004.07.010

[49] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. En: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 [consultado noviembre 10 del 2022]; 10: p. 785–794. Disponible en: doi:10.1145/2939672.2939785.

[50] Cui Z., Li Y, Cheng S, Yang H, Lu J, Hu Z, Ge B. Mutations in the embC-embA intergenic region contribute to *Mycobacterium tuberculosis* resistance to ethambutol. Antimicrob Agents Chemother. [Internet] 2014 [consultado octubre 10 del 2022]; 58(11): p. 6837–6843. Disponible en: 10.1128/AAC.03285-14

[51] Safi H, Lingaraju S, Amin A, Kim S, Jones M, Holmes M, McNeil M, Peterson SN, Chatterjee D, Fleischmann R, Alland D. Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-β-D-Arabinose biosynthetic and utilization pathway genes. Nat. Genet. [Internet] 2013 [consultado octubre 10 del 2022]; 45(10): p. 1190–1197. Disponible en: 10.1038/ng.2743

[52] Wan L, Liu H, Li M, Jiang Y, Zhao X, Liu Z, Wan K, Li G, Guan CX. Genomic Analysis Identifies Mutations Concerning Drug-Resistance and Beijing Genotype in Multidrug-Resistant *Mycobacterium tuberculosis* Isolated from China. Front. Microbiol. [Internet] 2020 [consultado febrero 10 del 2023]; 11: p. 1444. Disponible en: 10.3389/fmicb.2020.01444

[53] Dou HY, Lin CH, Chen YY, Yang SJ, Chang JR, Wu KM, Chen YT, Chin PJ, Liu YM, Su IJ, Tsai SF. Lineage-specific SNPs for genotyping of *Mycobacterium tuberculosis* clinical isolates. Sci. Rep. [Internet] 2017 [consultado marzo 3 del 2023]; 7(1): p. 1424. Disponible en: 10.1038/s41598-017-01580-z

[54] Sann WWM, Namwat W, Faksri K, Swe TL, Swe KK, Thwin T, Sangka A. Genetic diversity of *Mycobacterium tuberculosis* using 24-locus MIRU-VNTR typing and Spoligotyping in Upper Myanmar. J Infect Dev Ctries [Internet] 2020 [consultado marzo 10 del 2023]; 14(11): p. 1296–1305. Disponible en: 10.3855/jidc.12998

[55] Thain N, Le C, Crossa A, Ahuja SD, Meissner JS, Mathema B, Kreiswirth B, Kurepina N, Cohen T, Chindelevitch L. Towards better prediction of *Mycobacterium tuberculosis* lineages from MIRU-VNTR data. Infect Genet Evol [Internet] 2019 [consultado junio 19 del 2023]; 72: p. 59–66. Disponible en: 10.1016/j.meegid.2018.06.029

7. ANEXOS

El código implementado en este proyecto, así como los archivos necesarios, se encuentran en el siguiente repositorio de GitHub, el cual contiene toda la información necesaria para ejecutar el proyecto y hacer las modificaciones necesarias.

https://github.com/paredesgr/FirstLine_TB_ML_DrugClassifier

XGBC

El modelo de ML XGBC *(eXtreme Gradient Boosting Classifier)* es un clasificador basado en árboles de decisiones que utiliza el algoritmo de *boosting* de gradiente para mejorar la precisión de sus predicciones. El *boosting* de gradiente es un método que combina varios árboles de decisiones débiles para crear un árbol de decisiones más fuerte. Los árboles de decisiones débiles son aquellos que por sí solos tienen una precisión reducida, pero, cuando se combinan, pueden llegar a tener una precisión muy alta.

En cada iteración, el algoritmo calcula la pérdida (también conocida como función de costo) de cada nodo del árbol de decisión y selecciona el nodo que produzca la mayor reducción en la pérdida para dividir el árbol. Este proceso se repite hasta que se alcance un número predefinido de árboles o hasta que la reducción de la pérdida sea poco significativa.

Además, XGBC utiliza una técnica llamada *residual learning* que implica que, en cada iteración, el algoritmo aprende un nuevo modelo para predecir el residual de la predicción anterior. Esto lo logra mediante la adición de un nuevo árbol que se entrena para ajustar los residuos del modelo anterior. El nuevo árbol se agrega al modelo existente, lo que mejora gradualmente la capacidad del modelo para ajustar los datos.

Una de las ventajas de XGBC es que es muy rápido y puede manejar grandes conjuntos de datos sin sacrificar la precisión. También es capaz de manejar múltiples

tipos de variables (numéricas y categóricas) y tiene muchas opciones para ajustar y mejorar el rendimiento del modelo.

LGBC

El clasificador de aumento de gradiente ligero (LGBC, por sus siglas en inglés) emplea el mismo método de aumento que el XGBC. No obstante, una de las principales diferencias entre ambos radica en el tamaño de los árboles de decisión que se usan. El LGBC utiliza árboles de decisión más pequeños, lo que permite que se ejecute más rápido y se implemente con mayor facilidad en grandes conjuntos de datos. Por su parte, el XGBC emplea árboles de decisión más grandes y complejos, lo que puede generar un desempeño ligeramente superior en algunos casos, pero también puede ser más lento y difícil de implementar.

GradientBoosting

El algoritmo de *Gradient Boosting* funciona iterativamente, agregando un nuevo modelo al conjunto existente en cada iteración. Cada nuevo modelo trata de corregir los errores del conjunto anterior de modelos, lo que lleva a una mejora gradual de la precisión del modelo final.

Gradient Boosting es muy eficiente y capaz de manejar grandes conjuntos de datos sin sacrificar la precisión. También es capaz de manejar múltiples tipos de variables (numéricas y categóricas) y tiene muchas opciones para ajustar y mejorar el rendimiento del modelo. Uno de los principales inconvenientes de *Gradient Boosting* es que puede ser lento de entrenar y requerir muchos recursos computacionales, especialmente para conjuntos de datos muy grandes.

MLPClassifier

MLPClassifier es una clase de la librería de ML *scikit-learn* que se utiliza para entrenar modelos de redes neuronales artificiales, también conocidos como redes neuronales *multilayer perceptron* (MLP). Una red MLP es un tipo de red neuronal

feedforward que consta de una o varias capas ocultas de neuronas interconectadas entre sí y una capa de salida.

Exactitud

La exactitud mide la fracción de predicciones correctas que un modelo ha realizado sobre un conjunto de datos. Es una métrica útil cuando el número de casos positivos y negativos es equilibrado en el conjunto de datos. Sin embargo, si el número de casos positivos y negativos es desequilibrado, la exactitud puede ser engañosa y no reflejar el verdadero desempeño del modelo. En estos casos se recomiendan otras métricas, como la sensibilidad y la especificidad, para tener una evaluación más completa del modelo. La fórmula para calcular la exactitud es la siguiente:

$$\frac{Verdaderos\ Positivos + Verdaderos\ Negativos}{Total\ de\ muestras}$$

Sensibilidad

La sensibilidad, también conocida como tasa de verdaderos positivos, mide la fracción de casos positivos que un modelo ha identificado correctamente en un conjunto de datos. Es una métrica importante en el campo de la medicina para evaluar el desempeño de los modelos de clasificación, en este caso, para evaluar la detección de resistencia a los fármacos antituberculosis. La fórmula para calcular la sensibilidad es la siguiente:

$$\frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Negativos}$$

Especificidad

La especificidad es el contrapuesto de la sensibilidad, también conocida como tasa de verdaderos negativos. En el presente trabajo se utilizó para medir y comparar la fracción de casos negativos (susceptibles a los fármacos antituberculosis) que los

modelos lograron identificar en el conjunto de datos. Un modelo con una alta especificidad es capaz de identificar la mayoría de los aislados susceptibles a los fármacos. Sin embargo, una alta especificidad también puede tener un número elevado de falsos negativos. Para tener una evaluación completa del desempeño del modelo es importante comparar la especificidad con otras métricas como, por ejemplo, la sensibilidad. La fórmula para medir la especificidad es la siguiente:

$$\frac{Verdaderos\ Negativos}{Verdaderos\ Negativos + Falsos\ Positivos}$$

Precisión

La precisión se refiere a la fracción de predicciones positivas que un modelo ha identificado correctamente en un conjunto de datos. La precisión es útil para evaluar la capacidad de un modelo para hacer predicciones precisas.

En términos generales, un modelo con una alta precisión es capaz de identificar correctamente la mayoría de los pacientes que están realmente enfermos, pero puede tener un número elevado de falsos negativos. Es importante evaluar tanto la precisión como la sensibilidad para tener una evaluación completa del desempeño del modelo y tomar decisiones informadas sobre su uso en la práctica clínica. La fórmula para medir la precisión es la siguiente:

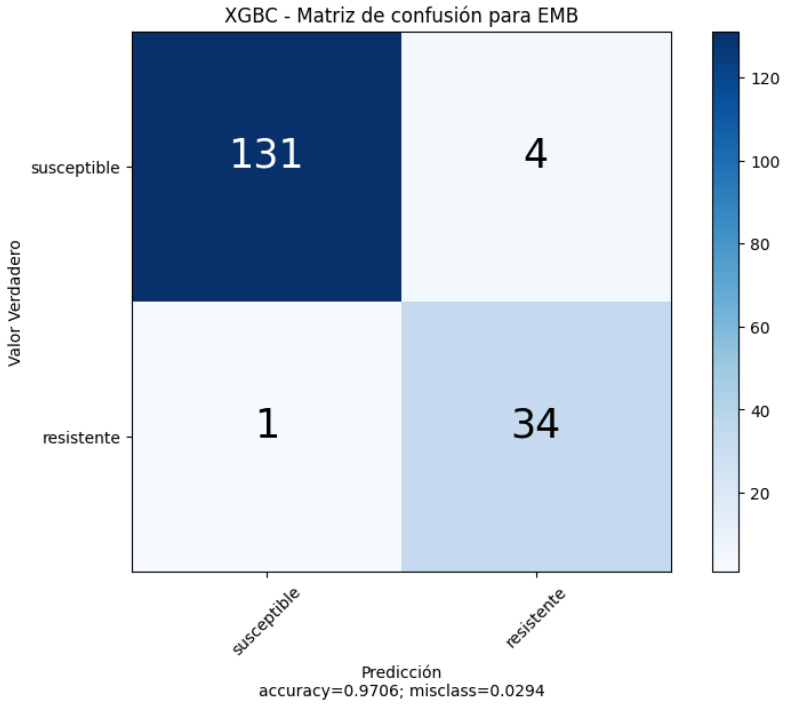
$$\frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Positivos}$$

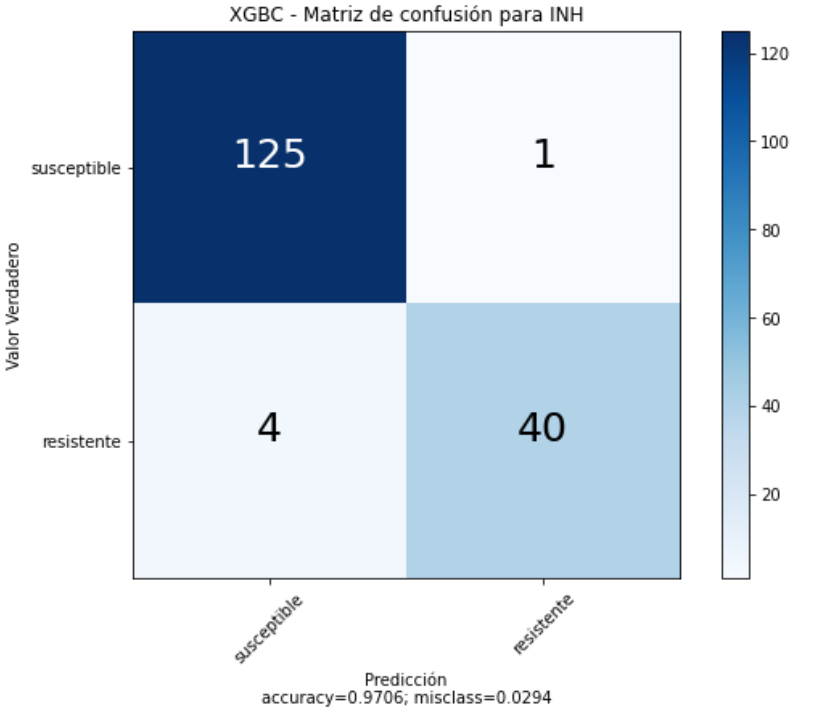
F1-Score

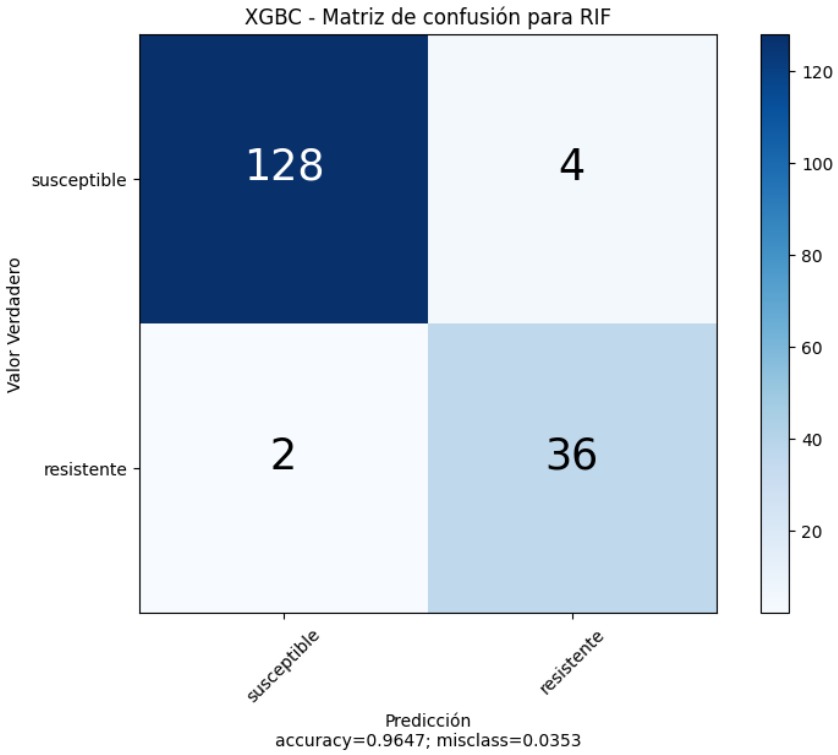
El F1-Score es una métrica que combina la precisión y la sensibilidad en un solo número que refleja la eficacia global de un modelo de aprendizaje automático en un problema de clasificación binaria. Un modelo con un alto F1-Score es capaz de identificar correctamente a la mayoría de los pacientes que están enfermos, mientras que también minimiza el número de falsos positivos y falsos negativos. Analizar el

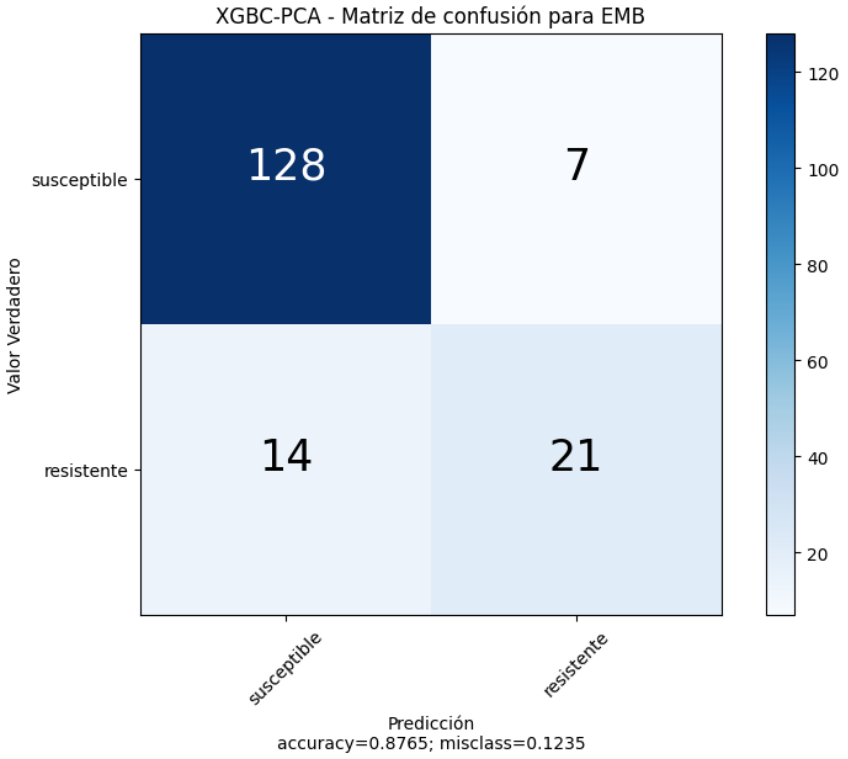
F1-Score es muy importante al comparar diferentes modelos ya que se obtiene un equilibrio entre la precisión y la sensibilidad.

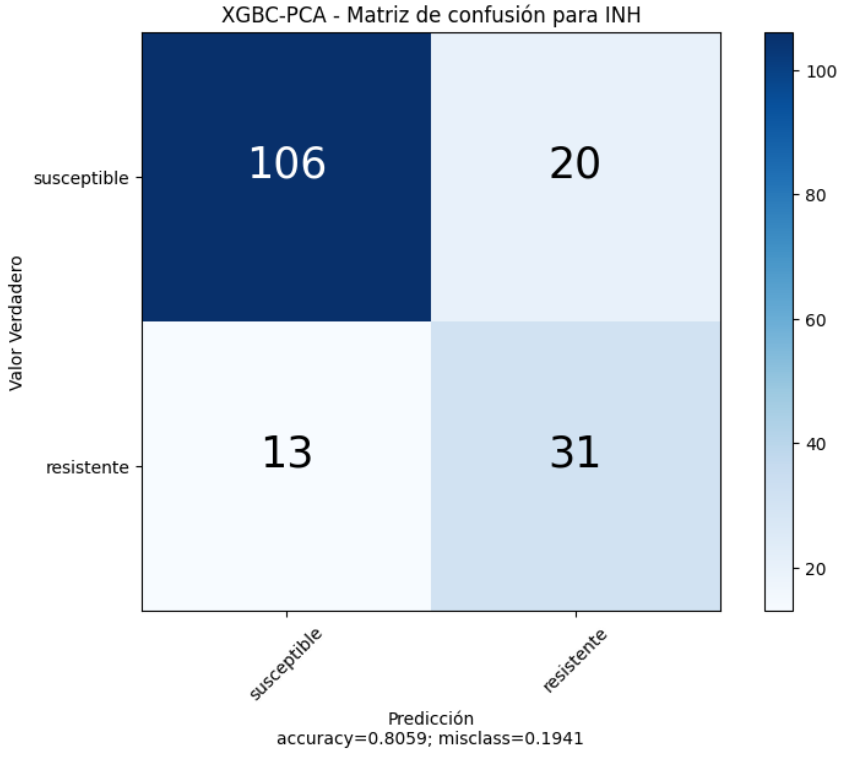
$$\frac{2 * (Precisión * Sensibilidad)}{Precisión + Sensibilidad}$$

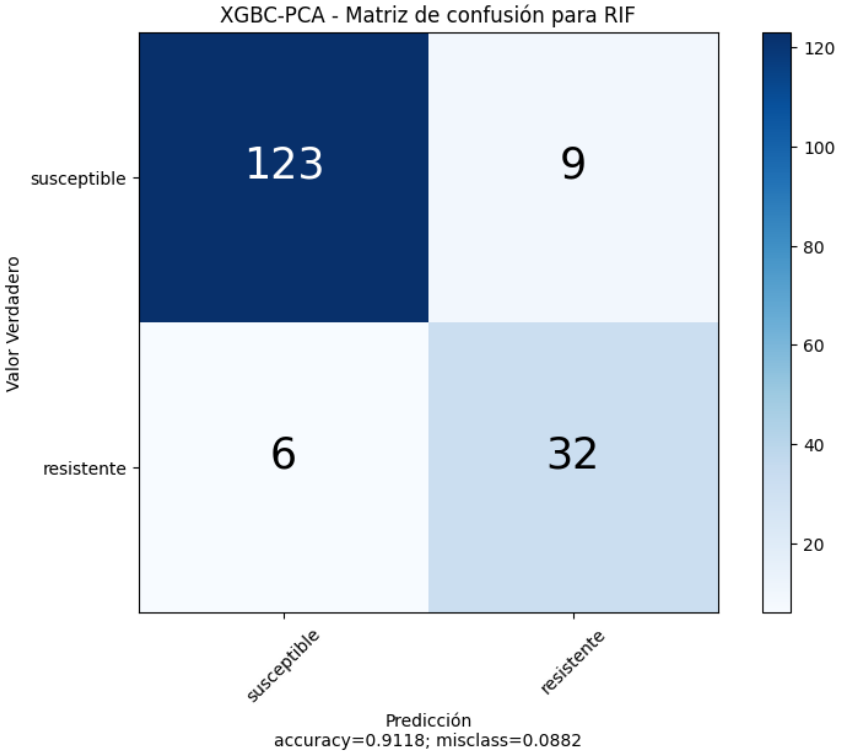


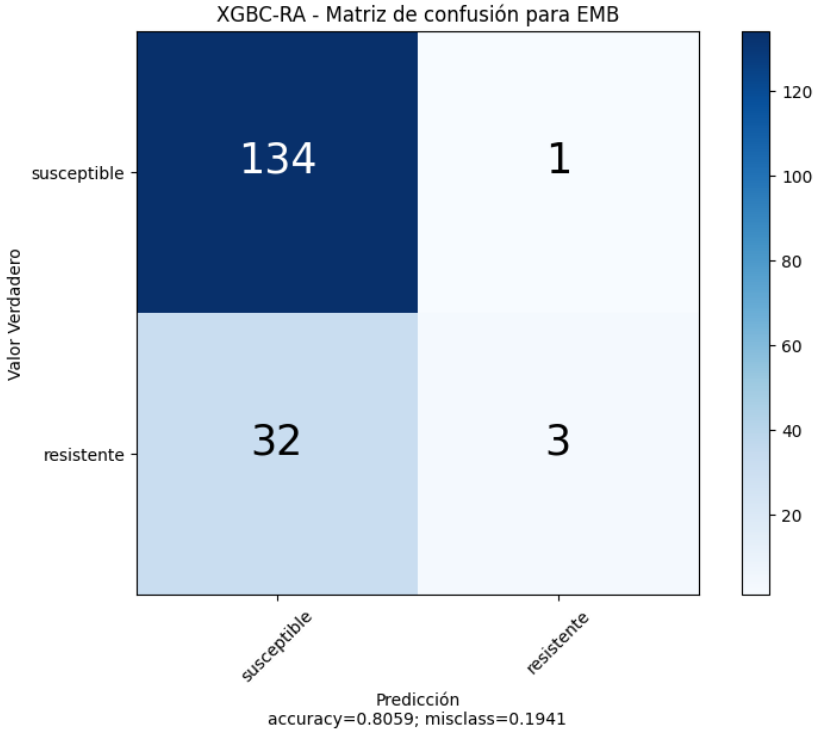

XGBC - Matriz de confusión para INH
125
1
4
40
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.9706; misclass=0.0294
120
100
80
60
40
20

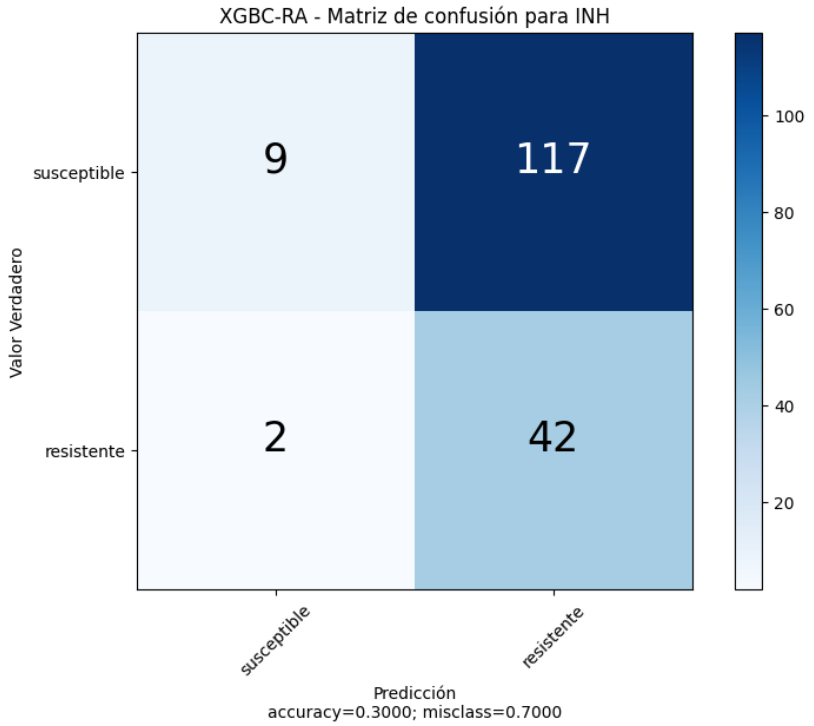

XGBC - Matriz de confusión para RIF
128
4
2
36
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.9647; misclass=0.0353
120
100
80
60
40
20

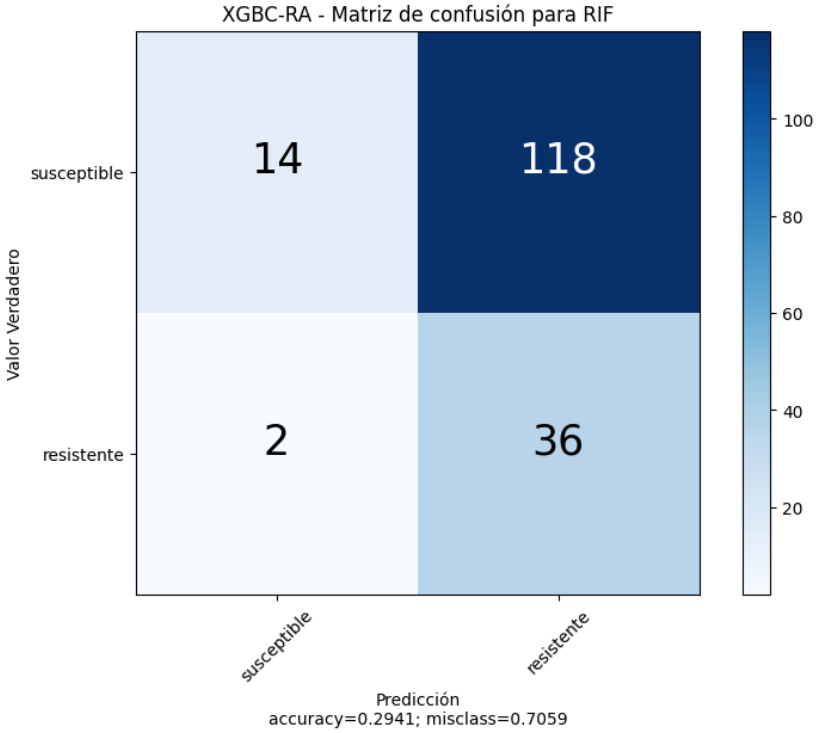

XGBC-PCA - Matriz de confusión para EMB
128
7
14
21
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.8765; misclass=0.1235
120
100
80
60
40
20

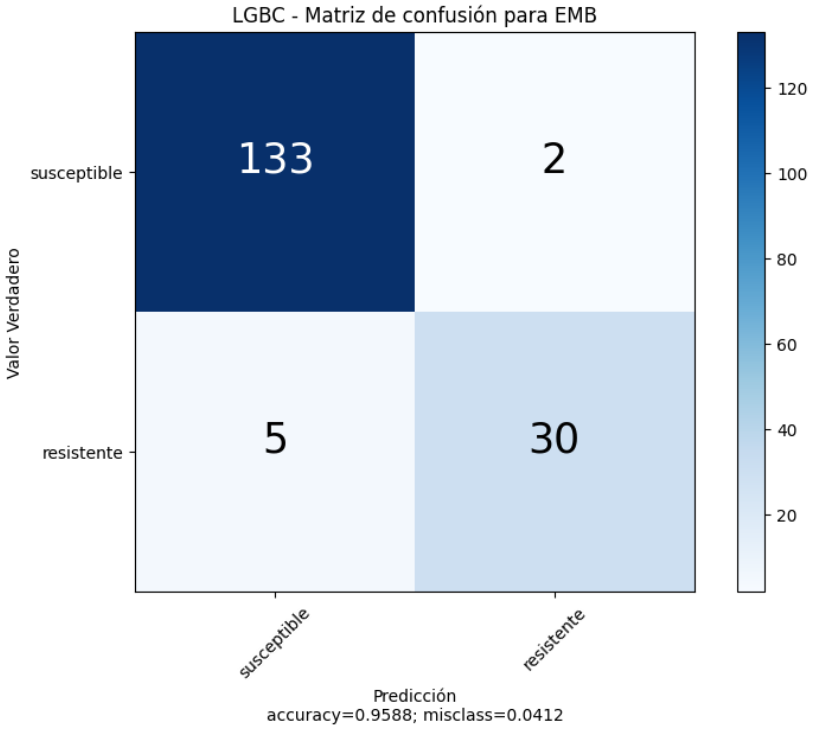

XGBC-PCA - Matriz de confusión para INH
106
20
13
31
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.8059; misclass=0.1941
100
80
60
40
20

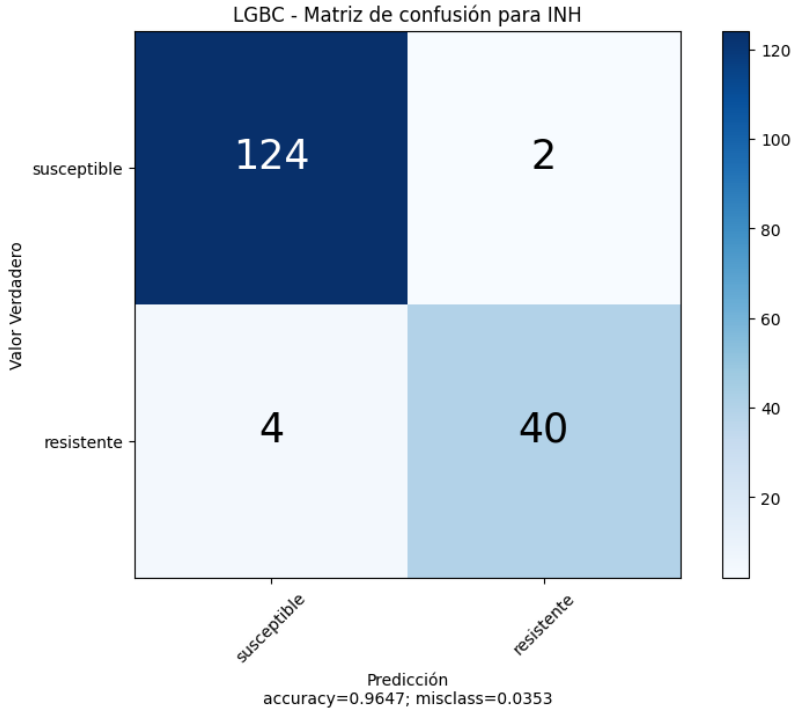

XGBC-PCA - Matriz de confusión para RIF
123
9
6
32
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.9118; misclass=0.0882
120
100
80
60
40
20

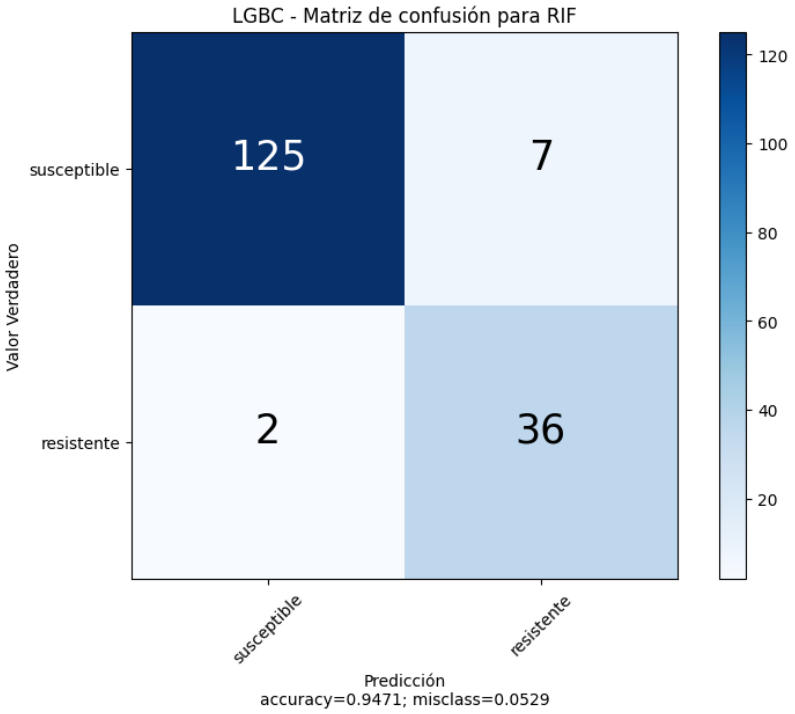

XGBC-RA - Matriz de confusión para EMB
134
1
32
3
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.8059; misclass=0.1941
120
100
80
60
40
20

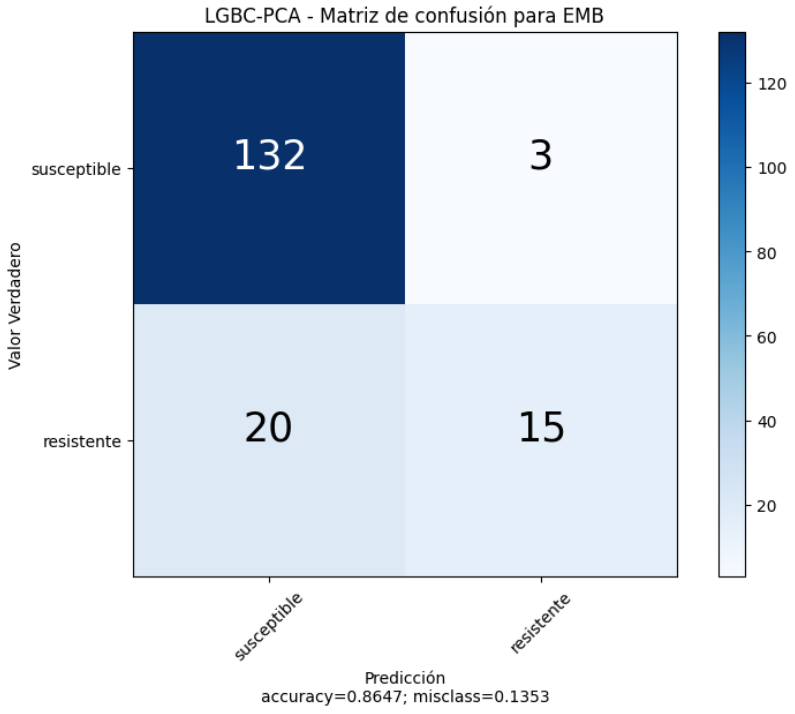

XGBC-RA - Matriz de confusión para INH
9
117
2
42
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.3000; misclass=0.7000
100
80
60
40
20

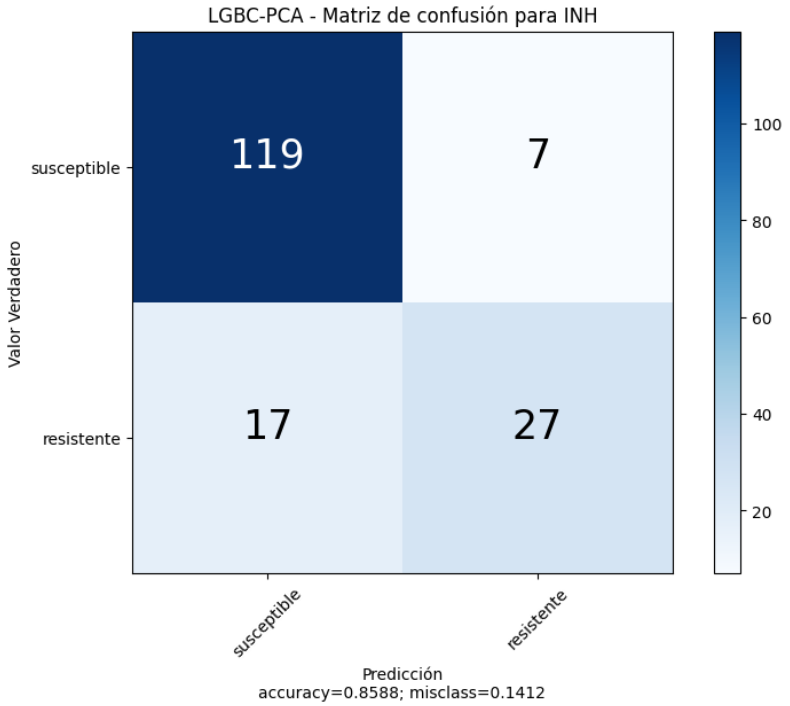

XGBC-RA - Matriz de confusión para RIF
14
118
2
36
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.2941; misclass=0.7059
100
80
60
40
20

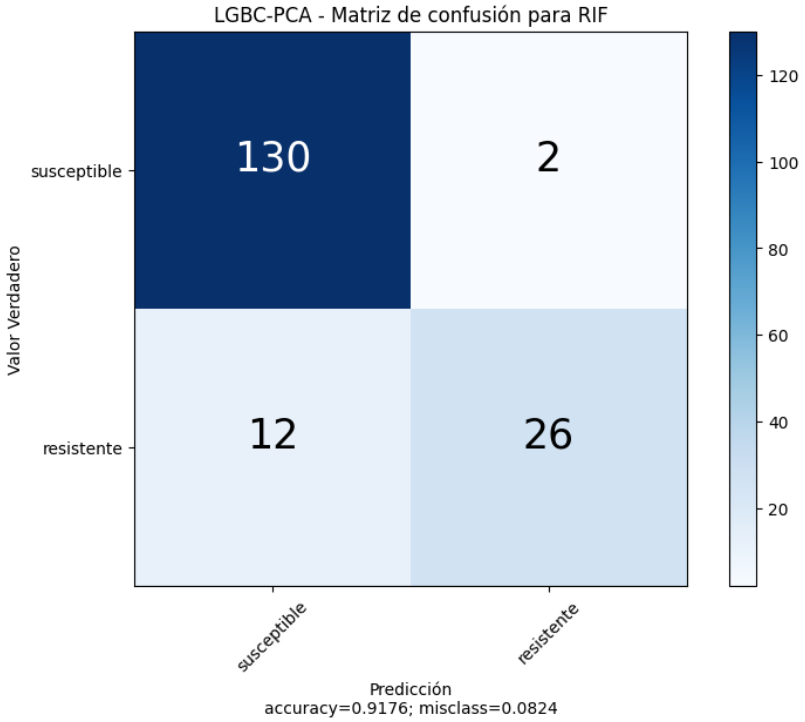

LGBC - Matriz de confusión para EMB
133
2
5
30
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.9588; misclass=0.0412
120
100
80
60
40
20

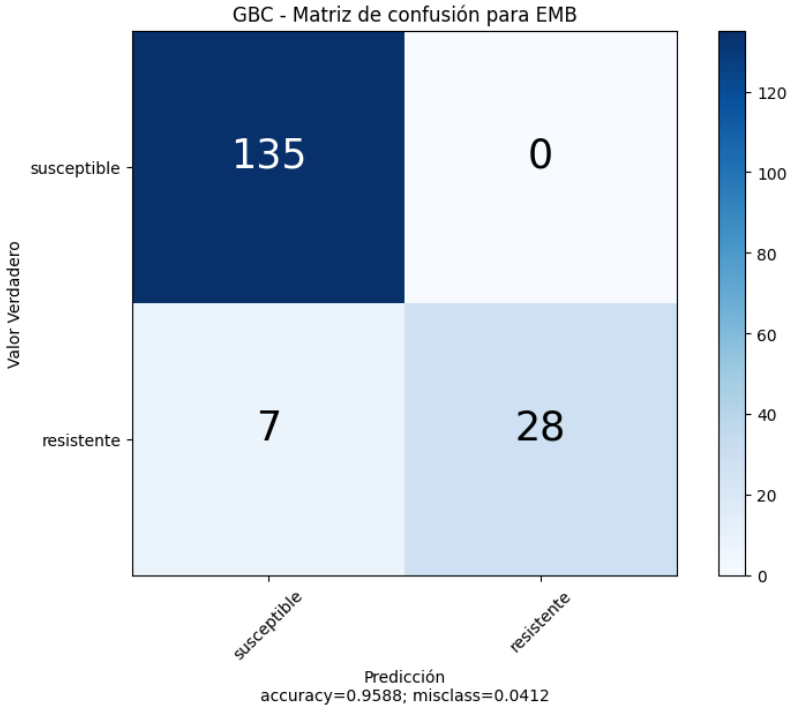

LGBC - Matriz de confusión para INH
124
2
4
40
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.9647; misclass=0.0353
120
100
80
60
40
20

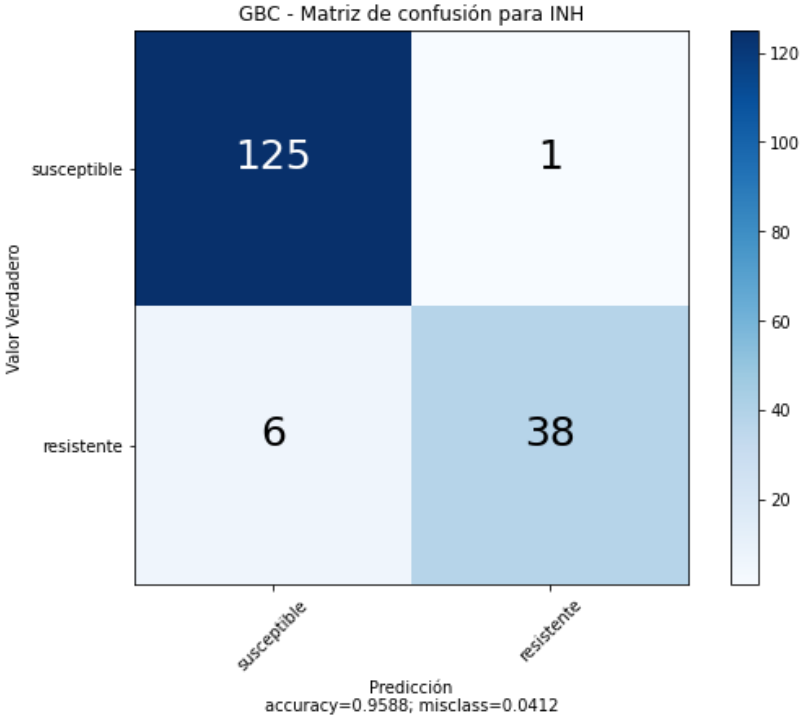

LGBC - Matriz de confusión para RIF
125
7
2
36
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.9471; misclass=0.0529
120
100
80
60
40
20

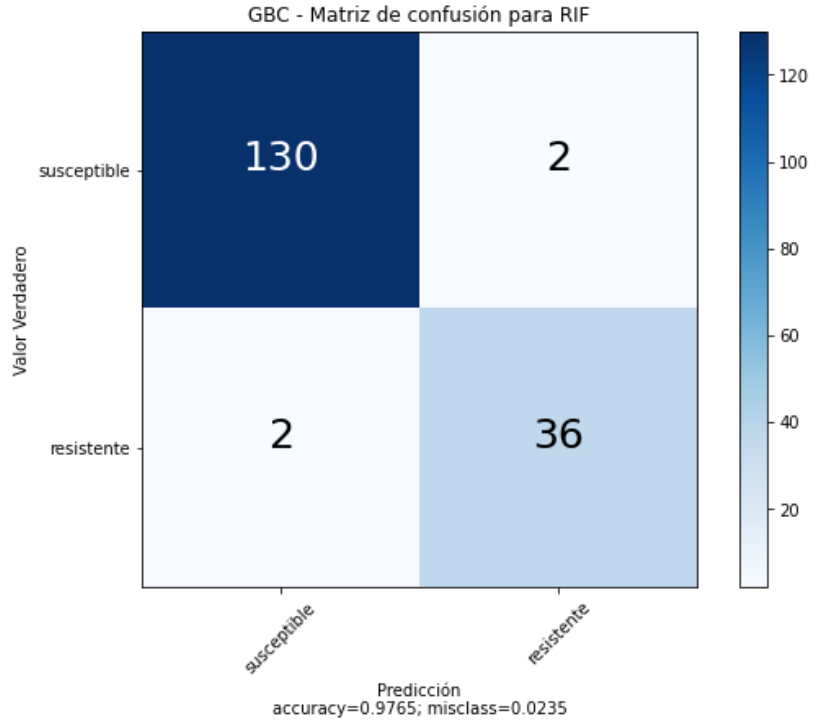

LGBC-PCA - Matriz de confusión para EMB
132
3
20
15
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.8647; misclass=0.1353
120
100
80
60
40
20

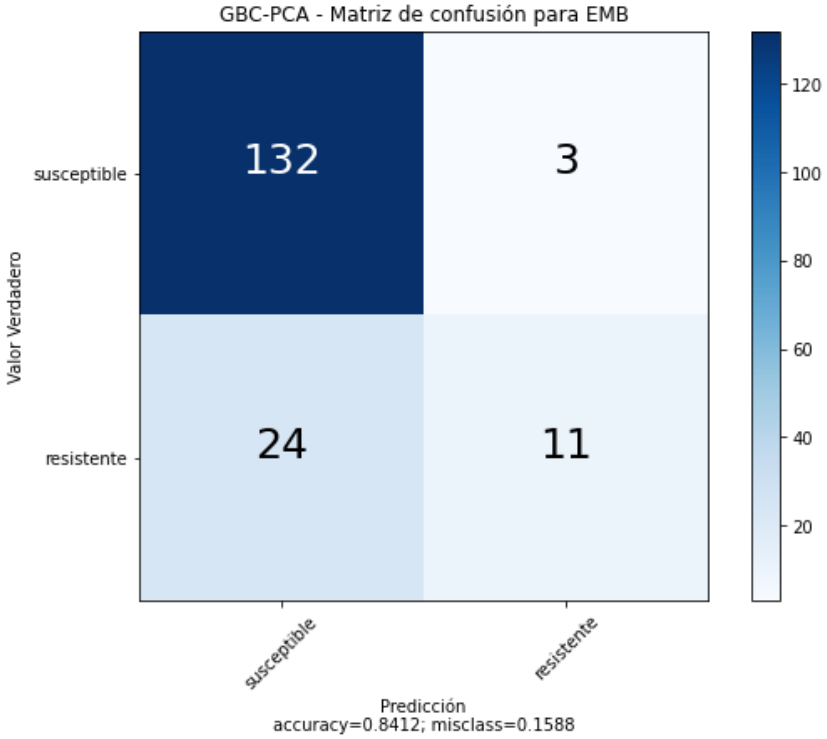

LGBC-PCA - Matriz de confusión para INH
119
7
17
27
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.8588; misclass=0.1412
100
80
60
40
20

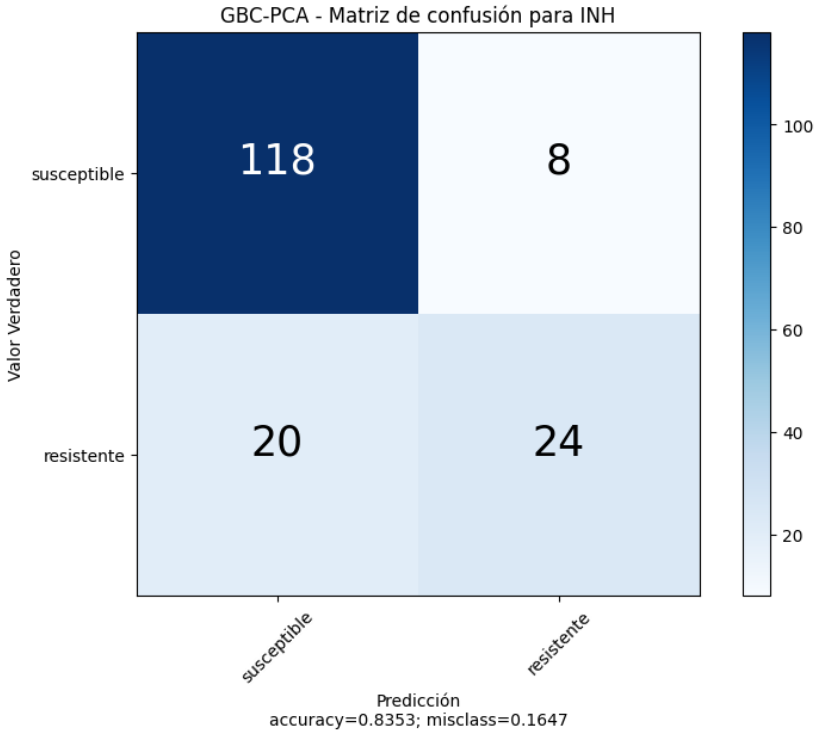

LGBC-PCA - Matriz de confusión para RIF
130
2
12
26
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.9176; misclass=0.0824
120
100
80
60
40
20

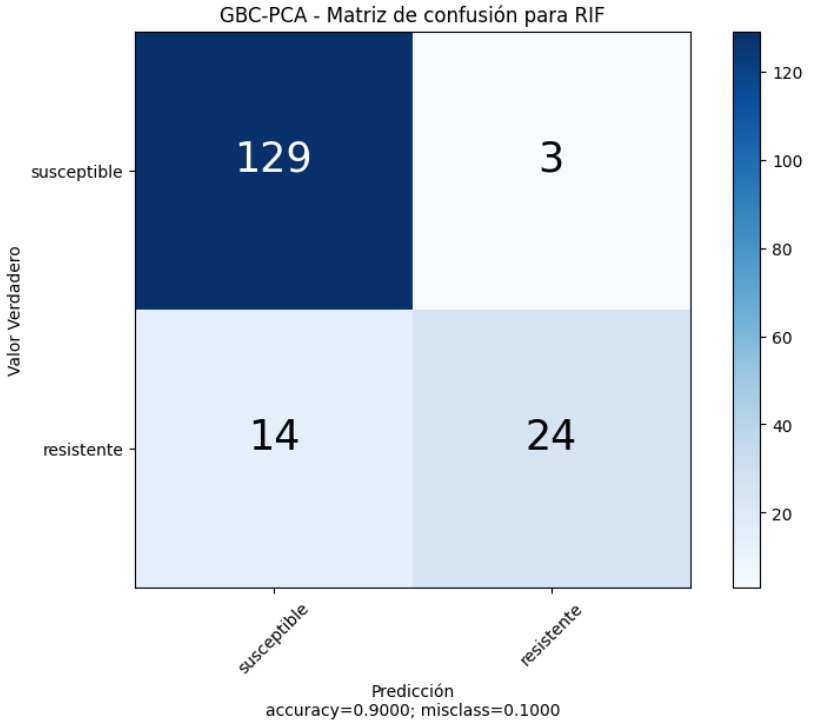

GBC - Matriz de confusión para EMB
susceptible
resistente
Valor Verdadero
135
0
7
28
120
100
80
60
40
20
0
susceptible
resistente
Predicción
accuracy=0.9588; misclass=0.0412

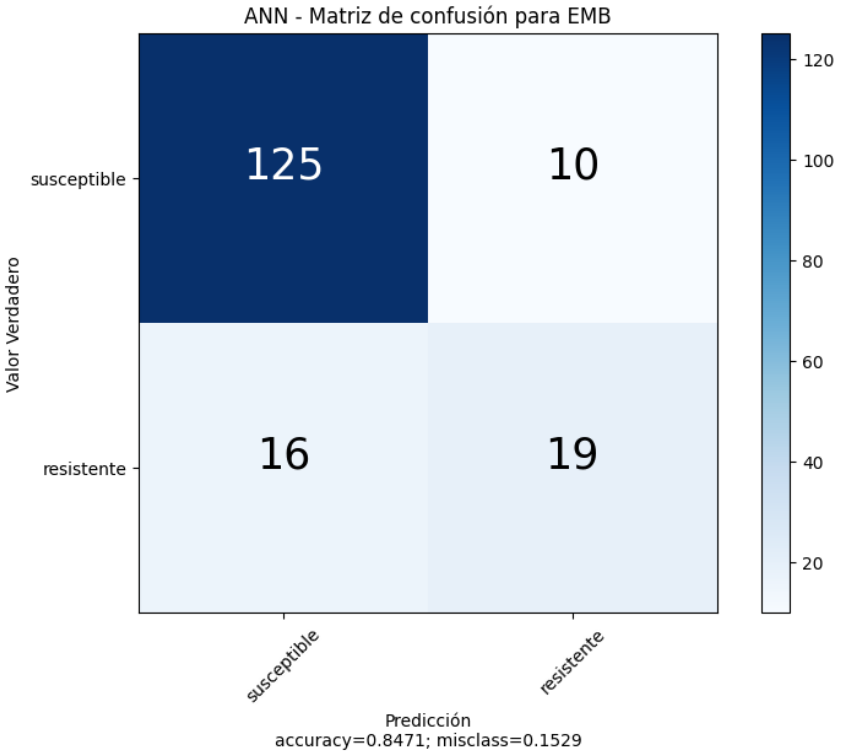

GBC - Matriz de confusión para INH
susceptible
resistente
Valor Verdadero
125
1
6
38
120
100
80
60
40
20
susceptible
resistente
Predicción
accuracy=0.9588; misclass=0.0412

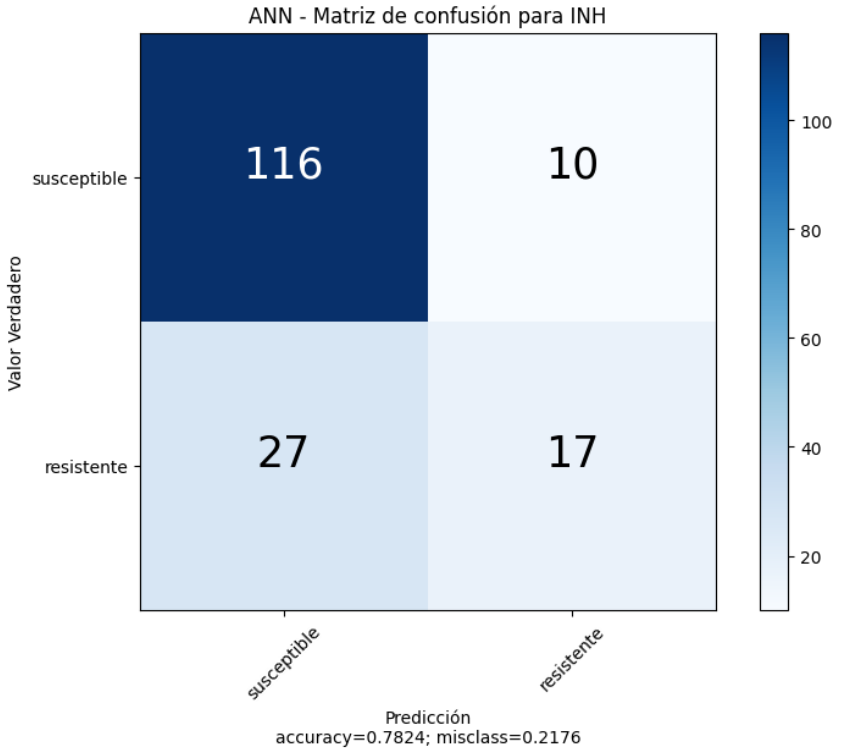

GBC - Matriz de confusión para RIF
130
2
2
36
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.9765; misclass=0.0235
120
100
80
60
40
20

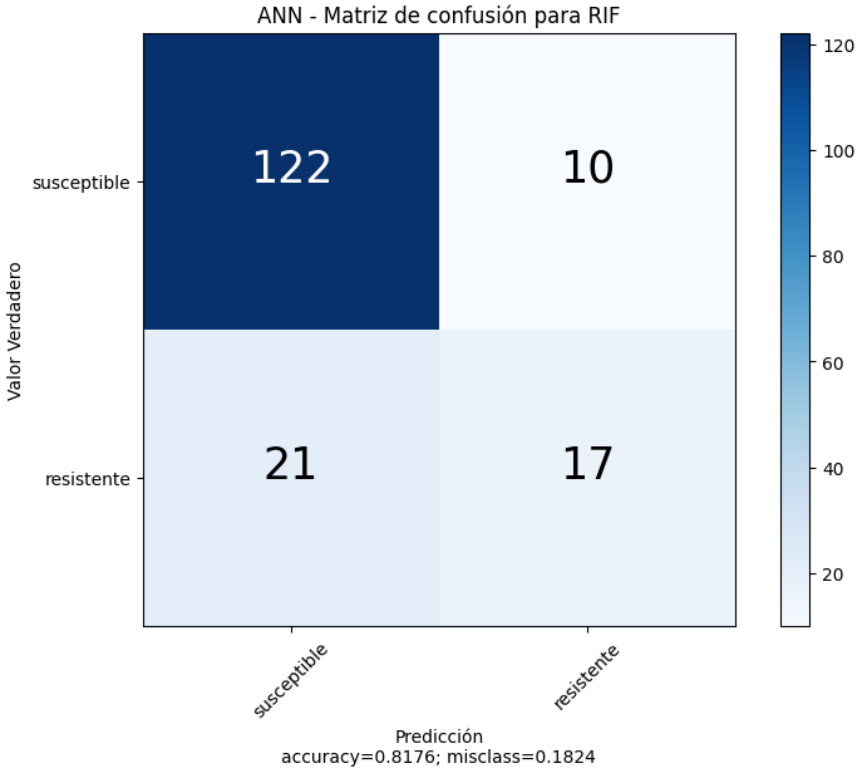

GBC-PCA - Matriz de confusión para EMB
132
3
24
11
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.8412; misclass=0.1588
120
100
80
60
40
20

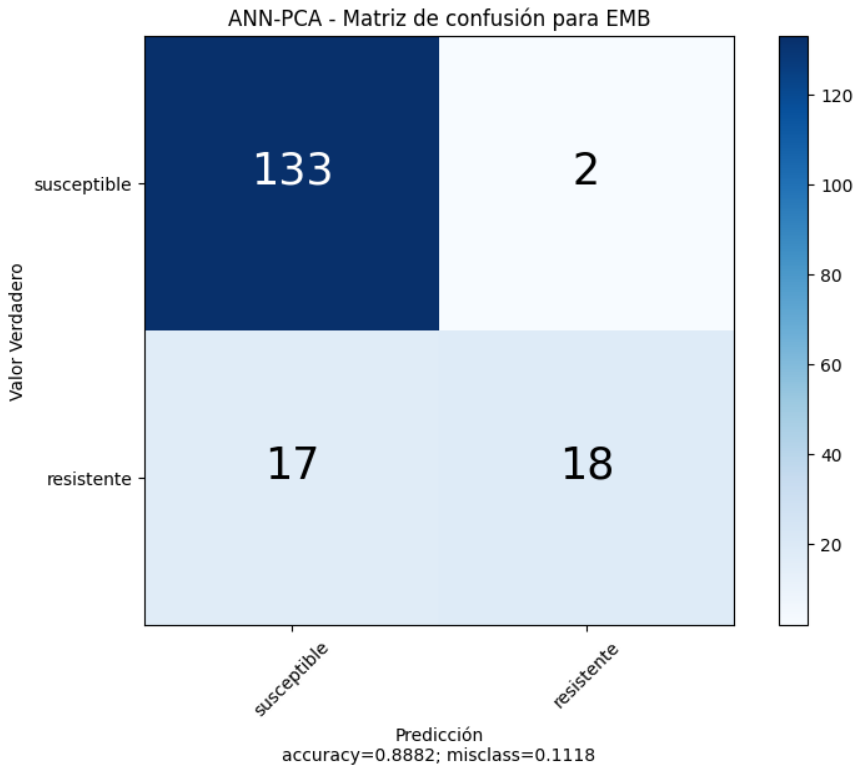

GBC-PCA - Matriz de confusión para INH
118
8
20
24
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.8353; misclass=0.1647
100
80
60
40
20

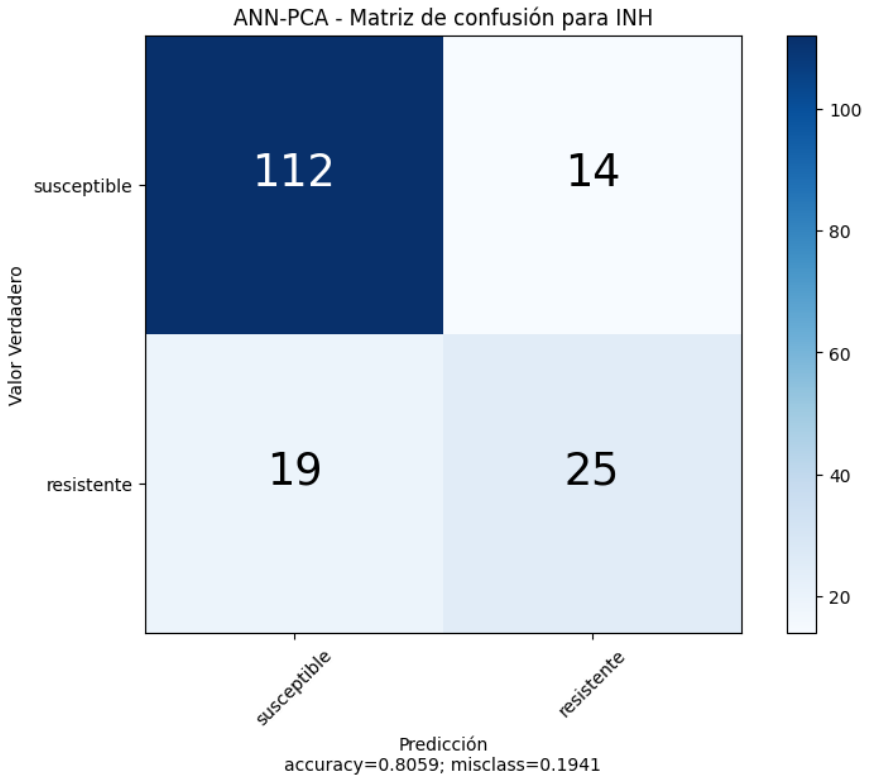

GBC-PCA - Matriz de confusión para RIF
129
3
14
24
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.9000; misclass=0.1000
120
100
80
60
40
20


ANN - Matriz de confusión para EMB
125
10
16
19
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.8471; misclass=0.1529
120
100
80
60
40
20


ANN - Matriz de confusión para INH
116
10
27
17
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.7824; misclass=0.2176
100
80
60
40
20


ANN - Matriz de confusión para RIF
122
10
21
17
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.8176; misclass=0.1824
120
100
80
60
40
20


ANN-PCA - Matriz de confusión para EMB
133
2
17
18
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.8882; misclass=0.1118
120
100
80
60
40
20


ANN-PCA - Matriz de confusión para INH
112
14
19
25
susceptible
resistente
Valor Verdadero
susceptible
resistente
Predicción
accuracy=0.8059; misclass=0.1941
100
80
60
40
20