Bionatura Journal
Ibero-American Journal of Biotechnology and Life Sciences

Article

Modelos impulsados por inteligencia artificial para la predicción de la respuesta al tratamiento antituberculoso: estudio de cohorte retrospectivo en el subtrópico ecuatoriano

AI-Driven Prediction of Tuberculosis Treatment Failure in Subtropical Ecuador: A Retrospective Cohort Study Comparing Logistic Regression, Random Forest, and Artificial Neural Networks

Alex Armijos-Hernández [1]*, *Nadia N. Sánchez -Pozo* [1,2]

[1] *Universidad Politécnica Estatal del Carchi, Tulcán 040102, Ecuador; alex.armijos, nadia.sanchez}@upec.edu.ec*
[2] *Mondragón Universidad, Mondragón 20500, España; n.sanchez@mondragon.edu*
* *Correspondence: alexmedicomauricio65@gmail.com*

RESUMEN

La tuberculosis (TB) sigue siendo un desafío global, cada año aumentan la tasa de nuevos infectados y de forma preocupante los pacientes que fracasan en su tratamiento.
Se evaluaron predictores clínicos y demográficos altamente relacionados con el fracaso de la terapia en una cohorte de adultos (18-64 años) con tuberculosis (n: 922, Hombres: 538 Mujeres:384). Se emplearon y evaluaron con un enfoque comparativo, modelos estadísticos partiendo de la regresión logística binaria (RLB), y modelos más complejos como Random Forest (RF) y redes neuronales artificiales con 2 capas ocultas (RNA) para capturar relaciones e interacciones entre variables. Además, se evaluaron dos escenarios: datos originales desbalanceados y datos balanceados a través de la técnica SMOTE, como un medio para abordar el desbalance entre fracasos y éxitos. Mientras que la RLB delineó una interpretabilidad clara, tuvo el mejor rendimiento en los escenarios sobre la RF y RNA ((Accuracy: 0,7111, Sensitivity: 0,44, Specificity: 0,7548 Precision: 0,2245 F1-Score: 0,2973 AUC-ROC:0,6591 Kappa: 0,1389), destacando sus capacidades de detección de patrones en edad, peso e historia de abandono inter variable. RLB logró el mejor equilibrio entre sensibilidad y especificidad con moderado rendimiento general para lograr la predicción de casos con posible fracaso terapéutico.
Keywords: Inteligencia Artificial; Random Forest; Tuberculosis; Smote; redes neuronas artificiales; regresión logística binaria.

ABSTRACT

Tuberculosis (TB) remains a major global health challenge, and treatment failure continues to undermine control efforts by prolonging transmission and increasing the risk of drug resistance. This study evaluated clinical and demographic predictors of TB treatment outcomes and compared statistical and machine-learning models to predict treatment failure in a subtropical Ecuadorian setting. We conducted a retrospective cohort analysis using routinely collected program data from a Ministry of Health primary care facility (Augusto Egas Type C Health Center, Santo Domingo de los Tsáchilas, Ecuador) covering TB cases treated from 2002 to 2024. Adults aged 18–64 years were analyzed (n=922). Candidate predictors included age, sex, baseline weight, HIV screening status, TB type (pulmonary/extrapulmonary), employment status, and prior TB history (relapse, previous abandonment, previous failure, and loss to follow-up), among others. Models were trained

and evaluated under two scenarios: (i) the original imbalanced dataset and (ii) a class-balanced training set generated with SMOTE to mitigate the minority-class (failure) underrepresentation. We compared binary logistic regression, random forest, and a two-hidden-layer artificial neural network using accuracy, sensitivity, specificity, precision, F1-score, balanced accuracy, Cohen's kappa, and AUC-ROC. In the imbalanced scenario, models showed strong bias toward treatment success, yielding very low or null sensitivity for failure detection. After SMOTE balancing, logistic regression achieved the most balanced performance (accuracy 0.711; sensitivity 0.44; specificity 0.7548; precision 0.2245; F1-score 0.2973; AUC-ROC 0.6591; kappa 0.1389), outperforming random forests and neural networks in identifying failures. Overall discrimination remained moderate, indicating that additional predictive variables and external validation are needed before clinical deployment. These findings support interpretable baseline modeling as a practical foundation for risk stratification in TB programs in similar resource-constrained contexts.

Keywords. Tuberculosis; Treatment failure; Treatment outcomes; Risk prediction; Machine learning; Logistic regression; Random forest; Artificial neural network; SMOTE; Retrospective cohort; Ecuador; Public health surveillance

INTRODUCCIÓN

La tuberculosis (TB) es una enfermedad infecciosa histórica con amplia distribución mundial que aún está lejos de erradicarse. Dentro de la meta 3.3 de los Objetivos de Desarrollo Sostenible de las Américas (ODS) está poner fin a la epidemia de tuberculosis, reduciendo su incidencia en un 80% y el número de muertes por TB en un 90% para el año 2030.[1]

Según la Organización Mundial de la Salud (OMS), se estima que en todo el mundo las personas que contrajeron TB fueron 10,6 millones hasta el 2022, de estos el 55% corresponden a hombres, el 33% a mujeres y el 12% a niños menores de 15 años. Las regiones más afectadas corresponden a África (23%), Asia Sudoriental (46%) y el Pacífico Occidental (18%). Asimismo, se estimó que hasta el 2022, el número de muertes por esta causa sería de 1,3 millones de personas.[1]

En Ecuador, según datos de la OMS, hasta el 2022 la tasa de incidencia fue de 45 casos por 100.000 habitantes, pero hasta enero del 2024 mostró un descenso a 23 casos por 100.000 habitantes.[2]

Los medicamentos recomendados para tratar la tuberculosis consisten en un combinado de antibióticos: isoniazida (H), rifampicina (R), pirazinamida (Z) y etambutol (E), que administrados en dosis fijas combinadas mejoran la aceptabilidad por parte del paciente, reducen el riesgo de dosis inadecuadas y mejoran la adherencia.[3]

Un problema grave para la efectividad de los tratamientos es el abandono terapéutico. La inasistencia a recibir terapia es un problema frecuente que impide en algunos casos lograr la curación completa, debido al alto riesgo de disminución de la eficacia, un periodo de contagio más prolongado, resistencia bacteriana, reactivación de la enfermedad y mayor mortalidad.[4] En diversos estudios se han identificado factores de riesgo para el abandono terapéutico entre los que destacan, el consumo de drogas recreativas, pobreza, antecedentes de abandonos previos, sentir malestar frecuente durante su tratamiento, así como se ha asociado fuertemente el ser de sexo masculino.[5]

La OMS recomienda que las tasas de curación de esta enfermedad deben estar sobre el 85% de los casos, y menos del 5% del abandono del tratamiento. Sin embargo, en países como Brasil se ha evidenciado una tasa de abandono del 12% y de curación del 70.1%.[4]

Lograr determinar que pacientes tienen una alta probabilidad de fracaso terapéutico desde el inicio de la terapia antituberculosa surge como una necesidad imperiosa, dado que, detectar los posibles casos con alto riesgo, permite enfocar las diferentes estrategias de adherencia terapéutica, los monitoreos y controles de la enfermedad más personalizados en casos focalizados.

Actualmente el uso combinado de técnicas y modelos predictivos de la estadística clásica con los nuevos modelos de inteligencia artificial, han permitido la detección oportuna de posibles pacientes clave con alto riesgo de fracaso terapéutico, así como en diversas investigaciones, han facilitado la detección oportuna de reacciones adversas a los medicamentos antituberculosos, permitiendo la intervención rápida.[6]

El análisis multivariante consiste en métodos o técnicas con el objetivo del análisis simultáneo de múltiples variables de un individuo u objeto de estudio, así como proporcionar técnicas que permitan el análisis conjunto de datos cuando el estudio unidimensional o bidimensional no es posible. Además, facilita la tarea del analista o investigador, permitiéndole tomar decisiones adecuadas, en función de la información brindada por el conjunto de datos analizados.[7]

Uno de los instrumentos de análisis multivariantes explicativo y predictivo es la regresión logística (RL). Este instrumento es ampliamente utilizado en salud para predecir eventos a partir de variables predictoras. Este modelo parte del conjunto de modelos lineales generalizados (GLM), mismo que utiliza la función logit como función de enlace. En la regresión logística binaria la variable dependiente sigue una distribución binomial (solo dos clases), por lo que el propósito es establecer causalidad en presencia de variables dicotómicas.[8]

Por otro lado, el Random Forest (RF) es un método de aprendizaje automático tipo ensamble, que consiste en la agregación de múltiples árboles de decisión, lo que proporciona una robustez y estabilidad en la predicción final en comparación con la predicción que pueda realizar un solo árbol. Cada árbol es entrenado con una muestra bootstrap del conjunto de datos, y en cada nodo de división, se selecciona al azar un subconjunto de características a evaluar para decidir la mejor partición, reduciendo la correlación entre árboles. En el caso de la clasificación, la predicción final se obtiene a través de una votación mayoritaria de los árboles, mientras que en regresión, se realiza el promedio de las predicciones. Este método incrementa la precisión de las predicciones, minimiza el sobreajuste y responde de forma natural a la heterogeneidad en las características, datos faltantes, relaciones no lineales y demás. También, es de fácil implementación y de fácil interpretación en cuanto la importancia de las características, aunque el ensemble de modelos en comparación con un árbol, es más complicado de interpretar en nivel de las predicciones.[9]

Las redes neuronales artificiales (RNA) son modelos de la Inteligencia Artificial (IA) que trata de imitar el funcionamiento de un cerebro biológico. Estas redes están construidas con capas de unidad simple que se denominan ‘neuronas artificiales’; distribuidas en neuronas de entrada, ocultas y salida. Cada unión neuronal posee un peso que se puede ajustar y modula la cantidad de la señal que se emite, además, utilizan ‘activación’ que constituye el proceso donde la señal de entrada se convierte en salida a partir de una función que activa dicha señal. Las redes pueden aprender representaciones complejas de datos ajustando la red y minimizando el error en la salida prevista con respecto a la salida deseada, este proceso puede ser guiado o no. Las RNA brindan mejores resultados en clasificación y regresión, así como el reconocimiento de imagen y el procesamiento de lenguaje natural.[10 11 12]

La combinación de técnicas de la estadística clásica, así como de la IA en el ámbito de la salud, y especialmente en enfermedades como la TB, surge como una necesidad urgente para controlar esta pandemia. Diversos estudios han analizado la eficiencia de modelos de IA para el manejo de esta enfermedad, incluyendo la máquina de vectores de soporte y la red neuronal convolucional, con resultados prometedores para la predicción de la eficacia del tratamiento.[13]

Mediante técnicas de análisis multivariante, esta investigación busca desarrollar un modelo predictivo para la detección de posibles casos de fracaso terapéutico antituberculoso, partiendo del modelo clásico como la regresión logística y comparando sus métricas de evaluación con modelos más complejos como la RF y RNA.

MATERIALES Y METODOS

La investigación se realiza bajo el enfoque cuantitativo, explicativo y predictivo, con datos obtenidos del Centro de Salud Tipo C Augusto Egas en Santo Domingo de los Tsáchilas, Ecuador, perteneciente al

Ministerio de Salud Pública (MSP), en donde se estudia una población de 1154 casos confirmados de tuberculosis tratados entre el periodo 2002 al 2024.
El objetivo principal es desarrollar un modelo predictivo basado en técnicas de clasificación multivariante para predecir el resultado del tratamiento antituberculoso. Para ello, se analizan las siguientes variables como punto de partida y posteriormente se someterán a la selección de aquellas con mayor interés clínico predictivo: Tabla 1

Variables independientes	
Sociodemográficas	Edad, sexo, nacionalidad, ocupación.
Epidemiológicas	Investigación de contactos censados, tamizaje respiratorio y contactos sintomáticos.
Clínicas	Comorbilidades. Tipo y localización de tuberculosis. Clasificación del caso al inicio del tratamiento: abandono recuperado, fracaso, nuevo, perdida en el seguimiento, recaída. Carga bacilar en esputo al diagnóstico. Resultado del tamizaje de VIH. Peso del paciente al inicio del tratamiento y al final del tratamiento. Ganancia de peso al primero, segundo y tercer mes de tratamiento, así como la ganancia total final. Días totales en tratamiento hasta el alta del paciente.
Variable Dependiente	
Resultado del tratamiento	Éxito: Todos los casos declarados curados o tratamientos completos dentro del unidad de salud. Fracaso: Los declarados como abandono, cambio o derivación del caso a otra unidad de salud y que no se conoce su condición final de egreso, fracaso terapéutico o fallecimiento.

Tabla 1. Variables dependientes e independientes para el análisis inicial.

El estudio se desarrolla en fases que incluyen la recolección de datos de historias clínicas, análisis exploratorio para identificar correlaciones, análisis de técnicas multivariantes, desarrollo y entrenamiento de modelos predictivos basados en modelos de IA y finalmente, el análisis de la importancia relativa de cada variable en la predicción del resultado del tratamiento.

El proceso de análisis para lograr cumplir los objetivos propuestos en esta investigación se realiza bajo 4 fases según el siguiente detalle: Figura 1

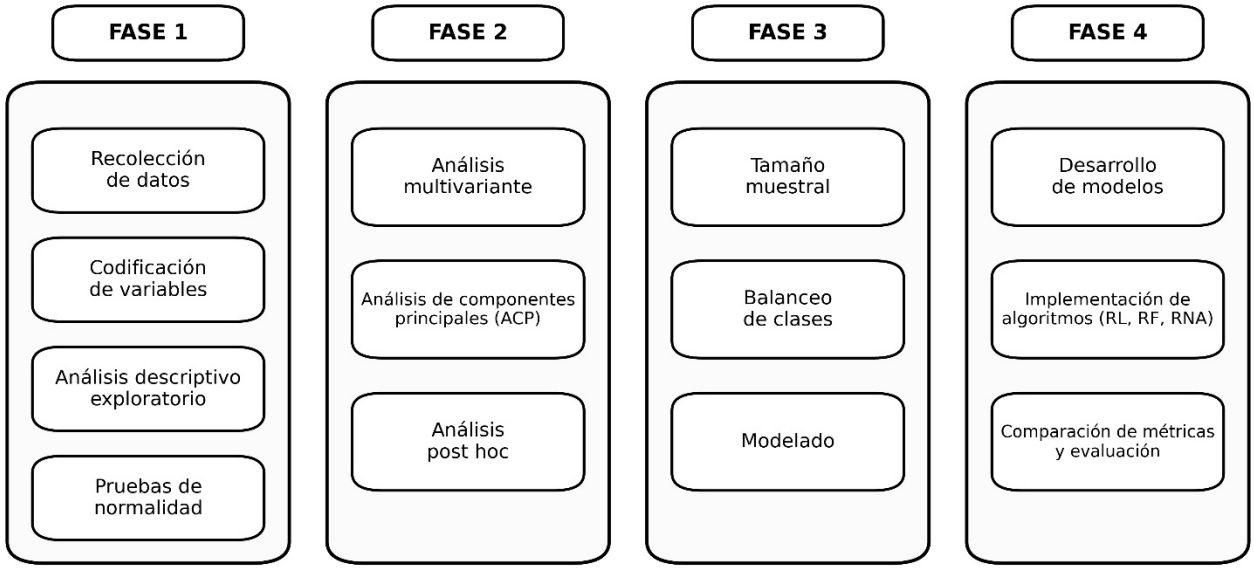


Figura 1. Fases de realizadas para cumplir los objetivos de la investigación.

Fase 1: Recolección de los datos, preparación y análisis exploratorio.

Se procede a la recolección y digitalización de los datos registrados en las tarjetas de control y seguimiento de los tratamientos de los pacientes con tuberculosis tratados entre 2002 al 2024 en el Centro de Salud Augusto Egas. La recolección se realiza de forman manual, revisando y registrando las variables de interés en un documento Excel versión 2016. Se realiza posteriormente la limpieza de los datos y se verifica valores faltantes. Para completar los datos faltantes se revisa las historias clínicas físicas y digitales en la plataforma de registro de atención en salud (PRAS) del ministerio de salud del Ecuador.
Se corrigen los errores de entrada y digitalización.

Codificación de variables

Se realiza el proceso de codificación de variables de interés, así como la asignación de categoría. Tabla 2

Clasificación, categorización y codificación de variables	
Variables	Codificación
Tipo de tuberculosis	Los casos se agrupan en tres categorías: Tuberculosis pulmonar (TB_PULMONAR) Tuberculosis extrapulmonar (TB_EXTRAPULMONAR) Tuberculosis clínico epidemiológico (TB_CRI_CLINICO_EPIDEMIO)
Carga bacilar diagnóstica	Ante la existencia de dos baciloscopias en esputo diagnosticas se selecciona la de mayor carga bacilar y se categoriza: A: (+++) Mayor a 10 BAAR por campo en 20 campos microscópicos. B: (++) 1 a 10 BAAR por campo en 50 campos microscópicos. C: (+) 10-99 BAAR en 100 campos microscópicos. D: (menor de +) 1 a 9 BAAR en 100 campos microscópicos. E: (-) En 100 campos microscópicos examinados no se encuentran BAAR (criterios clínicos epidemiológicos).

	F: No se realizó tomo de muestra* o corresponde a Tb extrapulmonar.
Resultado del tratamiento	Éxito: (A) que agrupa las categorías curado y tratamiento terminado. Fracaso: (B) agrupa las categorías abandono, fracaso, fallecido, perdida en el seguimiento y suspendido. Otros (C) está conformada por la subcategoría transferido, pero no se conoce su estado final del resultado**
Comorbilidades	A: Todos los registros con comorbilidad registradas siendo estas diabetes mellitus tipo 2, hipertensión arterial, cáncer de varios tipos e hipotiroidismo. B: Se agrupan los que no registran comorbilidades.
Ocupación	Según la clasificación Internacional Uniforme de Ocupaciones 2008.[14] Se agrupa en categorías como profesionales de la salud, profesionales del derecho, profesional de apoyo administrativo, profesional docente, vendedor ambulante, vendedor de tienda, oficiales artesanos y operarios, agricultura y ganadería, desempleado, jubilado, otros servicios y ninguna.
Contactos sintomáticos	SI: Considera los contactos del caso índice que presentan síntomas respiratorios NO: para los que no lo presentan.
Contacto positivo	SI: Contactos censados que resultaron positivos para tuberculosis No: Contactos que resultaron negativos.
Tamizaje de VIH	A: Casos con tamizaje Reactivos. B: No reactivos. C: No se realizan tamizaje o sin resultado.
Peso	Inicial: Peso en kilogramos (Kg) al inicio de la terapia. Final: Peso en Kg al final de la terapia. Ganancia total de peso: Diferencia entre el peso final y el inicial en Kg. Ganancia de peso del primero, segundo y tercer control ***.
Total de contactos	Número total de contactos censado en cada caso.
Diferencia de días entre el ingreso y el egreso	Días transcurridos entre la fecha de ingreso y la fecha de egreso del programa.

*Pacientes con diagnóstico clínico y epidemiológico o Tb extrapulmonar. ** Se identificaron dos casos (n:2) que ingresan en la categoría C, para el análisis predictivo estos casos ingresan a la categoría B.*** Corresponde a la diferencia entre el peso del primer control (generalmente al primer mes de terapia) y el peso inicial en Kg. En el caso la ganancia de peso del segundo control corresponde a la diferencia entre el peso registrado el segundo mes de tratamiento y el peso inicial, y para el tercer control es la diferencia entre el peso registrado en el tercer control (al tercer mes de tratamiento) y el peso inicial.

Tabla 2. Clasificación, categorización y codificación de variables para el análisis inicial.

Variable dependiente (resultado del tratamiento): Se procede a la creación de la variable dependiente éxito (A) que agrupa las categorías curado y tratamiento terminado. Para la variable fracaso (B) agrupa las categorías abandono, fracaso, fallecido, perdida en el seguimiento y suspendido. La categoría otros (C) está conformada por la subcategoría transferido, pero que no se conoce su estado final del resultado (n:2). Esta agrupación de 3 categorías se utiliza para el análisis descriptivo exploratorio inicial. Para el análisis predictivo la categoría C se integra a la categoría B.

El análisis estadístico inicial se lo realiza en el programa R studio versión R.4.4.3. Las librerías empleadas son descriptr, ggplot2, readr, patchwork, psych, dplyr, car, qqplotr, nortest, GPArotation, tidyverse, broom, recipes, recipes, keras, tidymodels, RColorBrewer, rsample, tidyr, rlang, GGally, skimr, y otras que se describen a continuación según su uso. Estas permiten realizar el análisis descriptivo con el cálculo de frecuencias y graficas para el análisis exploratorio.

Análisis descriptivo exploratorio y pruebas de normalidad.

Se realiza el análisis descriptivo de los datos de la base original total en donde se obtiene las frecuencias de mayor interés para esta investigación. Tabla 6, 7 Por otro lado, se aplican las pruebas de Shapiro-Wilk y Lilliefors para determinar normalidad de los datos. Tabla 8.

Fase 2: Realización del Análisis Multivariante y Selección de Variables de interés.

En esta fase se procede a aplicar técnicas de análisis multivariante con el fin de reducir la dimensionalidad previas a la construcción de los modelos predictivos.

Se emplea la base de datos original (adultos de 18-64 años) de la cual se procede a seleccionar las variables cuantitativas de interés siendo estas la edad, diferencia de días transcurridos entre el ingreso y el egreso del paciente, peso inicial, peso final, ganancia total del peso, ganancia de peso al primer mes de control, ganancia al segundo mes de control, ganancia al tercer mes de control y el total de contactos censados.

Se procede a la detección de multicolinealidad, para ello se calculó la matriz de correlaciones de Pearson utilizando la función pairwise.complete.obs. Se detectan dos variables con muy fuerte correlación ($|r| > 0{,}90$) siendo el peso inicial y peso final. Se decide eliminar la variable peso final y se repite el análisis, identificándose que la mayor parte de variables mantiene una débil correlación ($|r| \leq 0{,}30$), y 2 variables con moderada ($0{,}30<|r|\leq 0{,}70$) como son la ganancia de peso total y la ganancia del peso en el primer control, y ninguna con alta correlación ($|r| > 0{,}70$), valores que son clínicamente esperables.

Mediante el cálculo de las distancias de Mahalanobis se trata a los datos atípicos, Para ello se usó la media muestral y la matriz de covarianza empírica (use = "pairwise.complete.obs"), Como punto de corte se seleccionó el percentil 99,9 de la distribución x^2 con k grados de libertad (siendo k el número de variables, equivalente a $qchisq(0{,}999, k) \approx 22{,}46$. Se observan 19 casos (2,06%) considerados como atípicos debido a que están fuera del lumbral, y que son excluidos del análisis final. Se procede a generar un nuevo conjunto de datos eliminando los valores atípicos. Figura 2

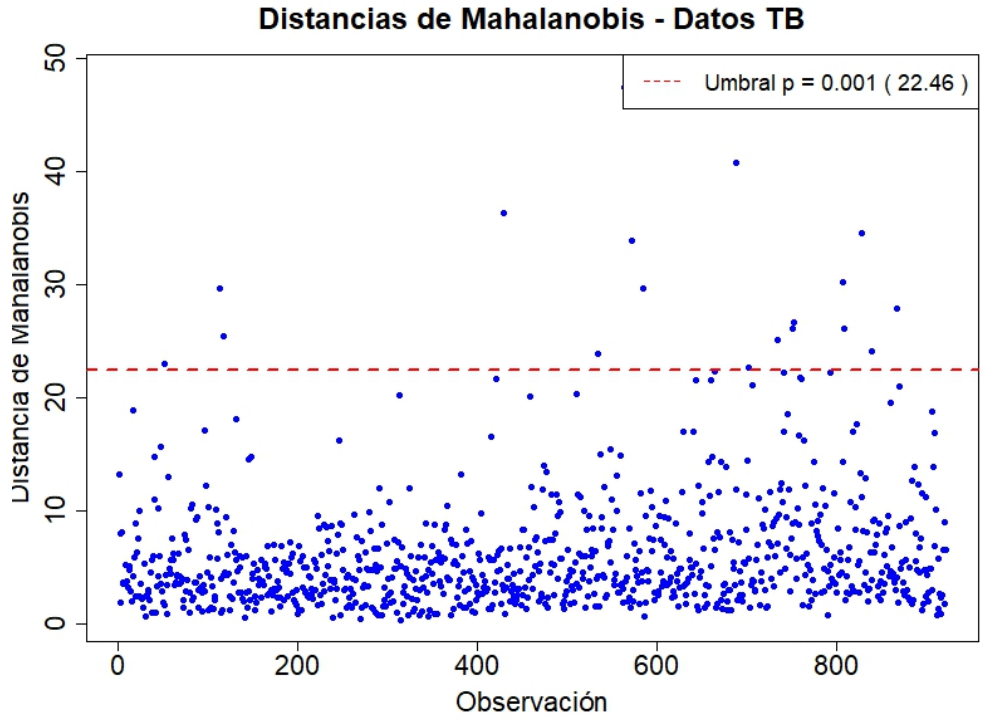


Figura 2.- Calculo de distancias de Mahalanobis de base de datos original de adultos de 18-64 años.

Se procede a la evaluación y comprobación de los supuestos previos al análisis factorial. Para evaluar si la matriz de correlaciones difiere significativamente de la identidad se aplica el test de esfericidad de Bartlett, mismo que resultó muy significativo ($p < 0,001$), de modo que las variables se correlacionan entre sí en grados suficientes. Para medir la proporción de varianza compartida frente a la varianza parcial no compartida se aplica el índice de Kaiser–Meyer–Olkin (KMO), esto nos permite observar la existencia de un exceso de redundancia en las variables correlacionadas. Para ello se establece valores de KMO < 0,50 como inapropiado para factorizar, 0,50 ≤ KMO < 0,70: mediocre, pero aceptable y KMO ≥ 0,70: bueno o excelente, indicando correlaciones parciales pequeñas favoreciendo la agrupación en factores definidos y estables. Los resultados obtenidos del índice KMO general es de 0,56 lo que señala una idoneidad muestral mediocre pero aceptable. El histograma de residuos estandarizados muestra una distribución casi cercana a la normalidad, sin embardo se evidencia una leve asimetría derecha, el Q-Q indica la mayor parte de puntos cercanos a la línea teórica, con ligeras desviaciones en los extremos, sugiriendo una normalidad multivariada aceptable. En lo relacionado a la homogeneidad de varianzas, se evidencia un patrón aleatorio en torno al cero, no heterocedasticidad. Al satisfacer los principales criterios, se continua con en análisis de factores. Figura 3

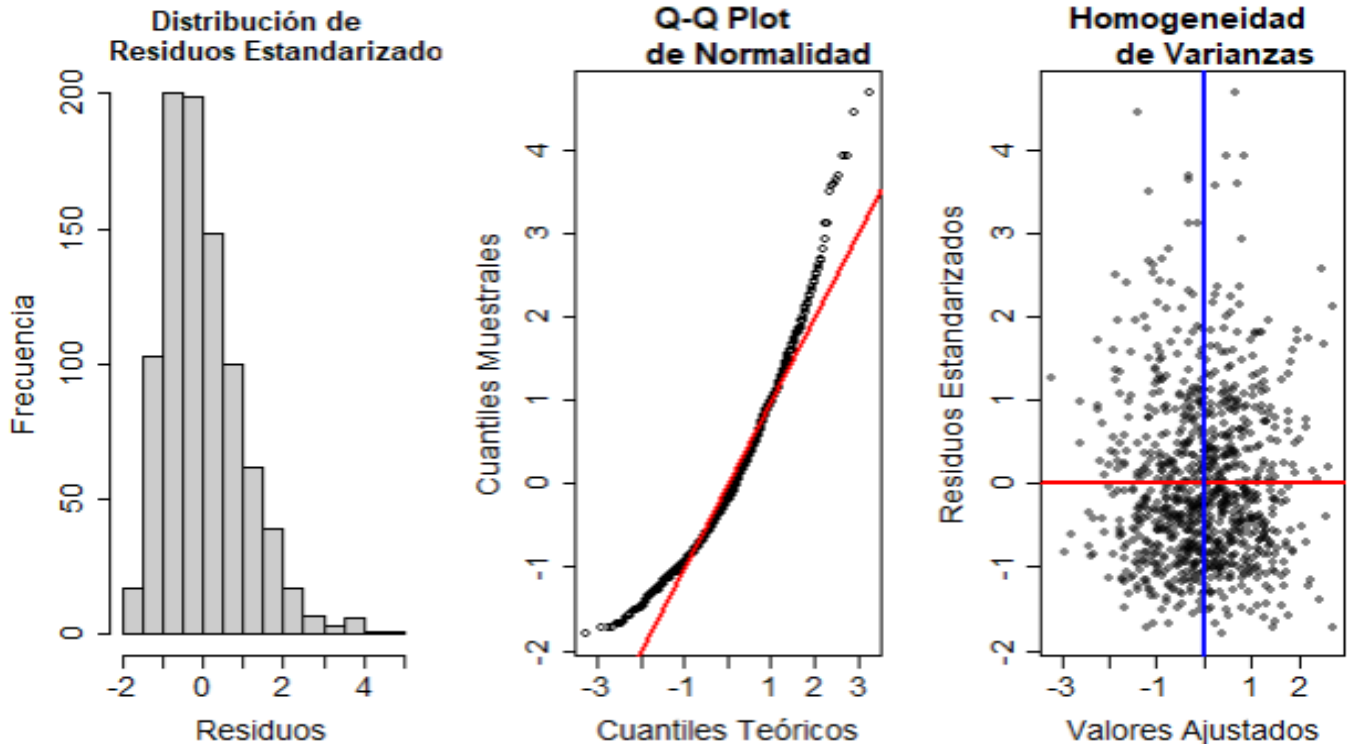


Figura 3.- Supuestos de normalidad del análisis factorial.

El Análisis de Componentes Principales (ACP) se efectúa con el fin de reducir la dimensionalidad y explorar los patrones. El análisis paralelo, comparando los autovalores obtenidos con los esperados bajo simulación aleatoria, permite determinar el número óptimo de componentes a retener. El scree plot mostró que el primer componente se ubica sobre las líneas asociadas a los datos simulados y remuestreados, y los demás componentes caen rápidamente del umbral establecido.

El análisis factorial realizado reveló consistentemente que no se justifica la extracción de más de un factor latente significativo. El criterio de Kaiser recomienda la selección de un solo factor, el mismo que supera el valor referencial de uno, y que es corroborado en la gráfica en donde no se observa retención de otros factores. Las métricas obtenidas del análisis del único factor seleccionado muestran un ajuste razonablemente bueno, con un RMSEA (Root Mean Square Error of Approximation) de 0,04, Raíz cuadrada media residual (RMSR) de 0,04, Tucker-Lewis Index (TLI) de 0,893, y con una varianza global del 15,6%. Las cargas factoriales revelan que únicamente dos ítems vinculados a la ganancia ponderal durante el tratamiento, ganancia total de

peso (0,78) y la ganancia de peso en el primer control (0,48), aportan saturaciones relevantes; por el contrario, las restantes variables edad, día de duración de tratamiento, el peso inicial en (KG) y el total de contactos muestran cargas muy bajas y comunalidades inferiores a 0,08, indicando que proporcionan información esencialmente independiente y no comparten varianza significativa con el único factor obtenido. Estos hallazgos permiten determinar variables no redundantes y que aporten robustes al modelo predictivo desde una base clínica. Figura 4

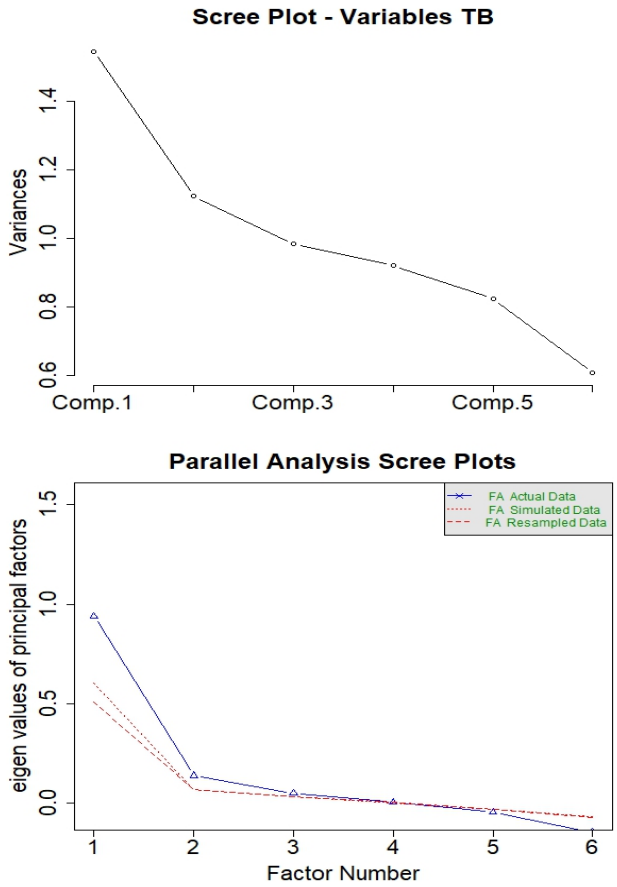


Figura 4.- Determinación del número de factores.

Análisis Post-Hoc - Validación de Factores

Dado que en el análisis factorial exploratorio solo se determinó que lo ideal es la extracción de un solo factor mientras que las demás variables mostraron cargas bajas, realizar el cálculo del alfa de Cronbach o de cualquier otra prueba de consistencia interna, no se justifica dado que aportaría poco valor. Al no contar con varios factores que agrupen múltiples variables, cada una con saturaciones robustas y que arrojen índices más reales, se decide no realizarla dado que desde punto de vista estadístico y practico carecería de una real utilidad.[15]

Fase 3.- Tamaño muestral y Balanceo de clases.

Posteriormente al análisis factorial y luego de haber tratado los datos atípicos, se procede a generar un nuevo conjunto de datos (dataset_sinatipicos). En esta se procede a asignar la variable independiente del "éxito en el tratamiento" en una variable con datos dummy (1=éxito,0=fracaso), además se integran a la data set las variables cualitativas restantes. En las variables cualitativas de mayor interés clínica y epidemiológica se transforma las categorías a valores dummy, esto permitirá tener datos que aporten un real interés clínico para el modelo predictivo sin variables que podrían confundir o que cuyos datos no aportan al modelo. Las variables seleccionadas y que mostraron mayor significancia en los análisis previos son la edad del paciente, el sexo masculino, el peso al inicio del tratamiento, casos con antecedentes de recaída, abandono, fracaso, y perdida del

seguimiento, además los pacientes con desempleo, los que presentar un resultado de VIH reactivo, así como los casos de tuberculosis pulmonar y extrapulmonar.

Para establecer la muestra total requerida que asegure la validez y estabilidad a los modelos predictivos se utiliza la Regla de Eventos por Variable (EPV), la cual establece que el número de ocurrencias del evento menos frecuente debe ser proporcional al número de variables predictoras que se van a incluir. Por ello en estudios previos como Peduzzi et al.[16 17 18], establecen un mínimo requerido de 10 a 20 eventos. En esta investigación el evento menos frecuente son los pacientes que fracasaron en su tratamiento y se determinan 11 variables predictoras. Considerando el mínimo requerido de fracasos de 10, se establece la necesidad mínima de 110 casos. Dado que el total muestral de nuestro data set final fue de 903 pacientes con tratamiento de tuberculosis, donde 127 (14,1%) presentaron fracaso en el tratamiento y 776 (85,9%) tuvieron éxito, los fracasos de la muestra obtenida superan al mínimo requerido de la muestra estimada, además se obtiene un EPV de 11,5:1 (127/11) y un total muestral estimado N= 785,6 (Fracasos mínimos proyectados: 110 y éxitos mínimos proyectados: 675,9), cumpliendo y asegurando con nuestra muestra la validez y estabilidad de los coeficientes de regresión obtenidos.
Se observa un desbalance con una proporción 6:1, a favor de casos de éxito, por lo que se justifica el uso de técnicas de balanceo para mejorar la capacidad predictiva de los modelos, dado que al no realizarlo nuestro modelo estaría sesgado a la clase mayoritaria (éxitos) y no a los casos que son los de real interés clínico para esta investigación (fracasos). Las librerías utilizadas son ROSE, themis (tidymodels) y Caret.

Se aplican varias técnicas de balanceo sobre los datos, una de estas es el undersampling aleatorio, mismo que se enfoca en equilibrar y recortar las clases desiguales, de esta manera se logra equiparar la proporción de la clase mayoritaria con la minoritaria. Este procedimiento equivale a un muestreo sin reemplazo que altera las probabilidades a priori de cada clase $P(Y)$, para aproximarla al 0,50 en problemas binarios. Con este procedimiento logra evitar la dominancia de las clases mayoritarias, pero corre el riesgo de aumentar la perdida de información y elevar la varianza del estimador. Por otra parte, al modificar el conjunto de datos de entrenamiento se asume que las instancias retenidas siguen una distribución representativa de la población.[19]

Por otro lado, el Oversampling aleatorio, otra técnica empleada, genera nuevas instancias de la clase minoritaria duplicando ejemplos existentes mediante muestreo con reemplazo, ajustando las probabilidades a priori $P(Y)$. Se basa en el principio Bootstrap, por lo que este método preserva la distribución de características de la clase minoritaria, lo que le permite mejorar la capacidad del clasificador en el aprendizaje de nuevos patrones. La fundamentación estadística se basa en la teoría del muestreo e inferencia por re muestreo para la estimación de parámetros de la población con muestras pequeñas. Una dificultad encontrada con esta técnica es que se puede incurrir en el sobreajuste dado que, al basarse esta técnica en la replicación, se corre el riesgo de repetir excesivamente los mismos datos.[19]

Se realiza la combinación de las dos técnicas de balanceo, el undersampling aleatorio (de la clase mayoritaria) + oversampling aleatorio (de la clase minoritaria). Esta técnica es conocida como balanceo híbrido y permite lograr un equilibrio de los conjuntos de datos mitigando a la vez el sesgo y la varianza. Busca la optimización de la función de pérdida ajustando el tamaño de la muestra de la clase mayoritaria n_{maj} como el de la minoritaria n_{min} de modo que $n_{maj} = n_{min}$, Esto permite que se amplifique la información de la clase minoritaria apoyada a la reducción de las instancias redundantes mejorando el rendimiento de las métricas.[20]

La técnica de balanceo ROSE (Random Over-Sampling Examples), se enfoca principalmente en la generación de nuevos datos sintéticos mediante técnicas de Bootstrap suavizado aplicados a la clase minoritaria, evitando así la excesiva replicación de instancias existentes. Mediante el uso de Kernels para cada clase, se logra la estimación de la densidad conjunta $f(X,Y)$, permitiendo así la simulación de nuevos pares $(X*,Y*)$ siguiendo la densidad condicionada $f(X|Y = \min)$. Esto proceso equivale a un remuestreo con perturbaciones aleatorias, permitiendo la reducción de la varianza del estimador al comparar con el oversampling puro, manteniendo la estructura local de la data. Al intercalar opcionalmente undersampling aleatorio de la clase

mayoritaria, ROSE realiza un balanceo híbrido. El undersampling aleatorio de la clase mayoritaria permite un balanceo híbrido además permite controlar de forma explícita la relación entre las clases.[21]

Por otro lado, la técnica SMOTE (Synthetic Minority Oversampling Technique), se enfoca en la generación de nuevos datos partiendo de la clase minoritaria interpolando entre las instancias originales x_i y uno de sus k vecinos mas cercanos x_{zi}, creando para cada par un punto sintético:
$x_{\text{new}} = x_i + \delta \times (x_{zi} - x_i), \delta \sim U(0,1)$. Este método permite la reducción de la varianza sin aumentar la distorsión de la frontera de decisión dado que se apoya en el espacio métrico de características y evita la duplicación exacta.[22]

Modelamiento

Para la realización de los experimentos se inicia dividiendo la base de datos en el 80% para el entrenamiento y 20% para la prueba. Para la evaluación de las métricas de balanceo se emplea la regresión logística binaria, este método permite predecir nuestro resultado dicotómico en función de varias variables predictoras. Su ecuación matemática es determinada por: [23] [24]

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

En donde
p= Representa la probabilidad de que ocurra los casos “Fracaso”.
β_0=Intercepto.
β_i= Miden el impacto del cada predictor x_i sobre el $log - odds$.

Siendo su función inversa logística para la estimación de p:[2324]

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}}$$

Posteriormente se calcula las probabilidades sobre los datos de prueba y se asigna umbrales, si la probabilidad es >0,5 se asigna el valor de 1 (Predicción de éxito) y si es <0,5 se agina el valor de 0 (Predicción de fracaso).

Se realiza las diferentes técnicas de balanceo tanto para datos no balanceados, así como para datos con balanceo y se compara las siguientes métricas: Tabla 3

Técnica	Accuracy	Sensitivity	Specificity	Precision	F1 Score	Balanced Accuracy	AUC ROC	Kappa
Sin Balanceo	0,85	0,00	0,99	0,00	NA	0,49	0,65	-0,02
Oversampling	0,32	0,48	0,30	0,10	0,16	0,39	0,31	-0,09
Undersampling	0,33	0,52	0,30	0,11	0,18	0,41	0,39	-0,07
Hibrido	0,33	0,56	0,30	0,11	0,19	0,43	0,37	-0,05
ROSE	0,48	0,44	0,49	0,12	0,19	0,47	0,44	-0,03
SMOTE	**0,71**	**0,44**	**0,75**	**0,22**	**0,29**	**0,59**	**0,66**	**0,13**

Los valores en **negrita** representan a la técnica que mostró el mejor desempeño comparativo.
Tabla 3.- Análisis comparativo de las métricas obtenidas con las diferentes técnicas de balanceo.

En la tabla se muestra el impacto de las diferentes técnicas de balanceo empleadas, es evidente las diferencias en el desempeño predictivo de los modelos de regresión logística binaria, para la detección de resultados terapéuticos para tuberculosis. Todas las técnicas de balanceo aportan en la identificación de la clase minoritaria (fracasos), a diferencia de la técnica sin balanceo, en donde se evidencia la falta de detección dado el sesgo a la clase mayoritaria de éxito terapéutico (Sensitivity = 0, Balanced Accuracy = 0,49). Es evidente el descenso de Accuracy global luego del balanceo yendo de 0,85 para pasar de un 0,32 al 0,71, produciendo un

fenómeno de tradeoff, relacionado al maximizar la detección de casos fracasos (sensibilidad) así como aumentar la detección de falsos positivos, lo que desde el contexto clínico resulta relevante.
La técnica que mejores métricas muestra es el método SMOTE, con un buen rendimiento promedio en Balanced Accuracy (0,59), AUC-ROC (0,66) y un Kappa positivo (0,13). Esto significa que SMOTE logra un equilibrio superior detectando más casos de fracasos (Sensitivity = 0,44), reduciendo el número de falsos positivos (Precision = 0,22) y preservando la discriminación general del modelo. Técnicas como Oversampling y Undersampling, aunque incrementan la sensibilidad (0,48 y 0,52, respectivamente), presentan valores muy bajos de precisión (0,10 y 0,11) y un acuerdo (Kappa) negativo, lo que indica una alta proporción de alarmas falsas y una menor consistencia predictiva. El método híbrido y ROSE alcanzan desempeños intermedios, aumentando la sensibilidad, pero sin corregir completamente el descenso en especificidad que acompaña al aumento en sensibilidad.

Fase 4.- Desarrollo y evaluación de modelos.

Modelización

Luego de identificar a la técnica de balanceo SMOTE como la más óptima, se procede al desarrollo de tres algoritmos de aprendizaje automático sobre la base de datos originales, enfocados en la predicción de los casos fracaso en el tratamiento antituberculoso. Los modelos entrenados son la Regresión Logística Binaria (RLB), RF y RNA. Se plantea para el entrenamiento el uso de 80% de datos, mientras que para la evaluación el 20%. Se experimenta bajo dos escenarios, el (**a)** con el uso de datos originales sin balanceo (conjunto de entrenamiento n=723, y evaluación n= 180 sin aplicar técnicas de balanceo), y **(b)** con datos balanceados mediante SMOTE aplicados sobre el 80% de los datos de entrenamiento previo a modelización (conjunto de entrenamiento n=1117) mediante la función step_smote() de los paquetes recipes y themis de R, con parámetro over_ratio = 0.8 para equilibrio parcial de clases, reduciendo el riesgo de sobre ajuste.[25] Es importante recordar que al aplicar la técnica SMOTE sobre el entrenamiento en la clase fracaso, aumentaron las observaciones mediante la interpolación sintética, pasando de 102 a 496 observaciones y con un incremento del conjunto de datos totales en el entrenamiento de 723 a 1117.

Preparación de datos y validación

Las observaciones que conforman el conjunto de datos original sin atípicos de adultos de 18 a 64 años, son alrededor de 903, con 11 variables predictoras y que fueron seleccionadas en pasos anteriores (edad, peso inicial, sexo (masculino), antecedentes de recaída, abandono previo, fracaso anterior, pérdida de seguimiento, ocupación (desempleado), tipo de TB: pulmonar/extrapulmonar y estado VIH). Es notoria un desbalance de las clases hacia el éxito (776 siendo el 85,9%) con una relación 6:1 frente al fracaso (127 casos siendo un 14,1%).
El conjunto de datos fue dividido en proporciones, el 80% para el entrenamiento y 20% para la evaluación, gracias a la función “caret::createDataPartition()”, que permite una partición estratificada manteniendo la distribución de clases. Con el “set.seed(1234)” se garantiza la reproductividad. Tabla 4

Escenario	Fase	(N) Total	(N) Fracaso (%)	(N) Éxito (%)	Razón
Original	Entrenamiento	723	102 (14,1%)	621 (85,9%)	1:6,1
	Prueba	180	25 (13,9%)	155 (86,1%)	1:6,2
SMOTE	**Entrenamiento**	**1117**	**496 (44,4%) ***	**621 (55,6%)***	**1:1,25**
	Prueba	180	25 (13,9%)	155 (86,1%)	1:6,2

*Mediante la técnica SMOTE sobre los datos de entrenamiento se generó 394 observaciones sintéticas en la case fracaso, lo que permite el balanceo. Los valores en **negrita** representan la técnica con mejor balanceo para las clases analizadas.

Tabla 4.- Composición tanto para entrenamiento y prueba de datos no balanceados (conjunto de datos original sin atípicos) así como balanceados (SMOTE)

Implementación de Algoritmos

Regresión Logística Binaria

Sobre las 11 variables predictoras se implementa la función "glm()" de la familia binomial y con enlace logit. sin aplicación de técnicas de regularización adicionales (penalidades L1 Lasso o L2 Ridge), permitiendo el desempeño del modelo clásico de regresión logística. Además, este modelo actúa como un enfoque "baseline" dado que permite evaluar la capacidad predictiva de un modelo básico y no muy sofisticado, además de otorgar una interpretabilidad directa a los coeficientes de las variables clínicas de tuberculosis.[26] Los supuestos del modelo como la linealidad en el logit, independencia, ausencia de colinealidad severa se evaluaron mediante VIF y pruebas de residuos.

Random Forest

Mediante el paquete Random Forest de R y bajo la configuración de "ntree=500" y "mtry=5 (500 árboles de decisión y como máximo 5 variables por nodo) parámetros seleccionados tras la realización de la validación cruzada interna, además de ser valores recomendados para la convergencia de métricas de rendimiento y evitando el incremento computacional prohibitivo.[27] [28] Se utilizó criterio de Gini para división óptima de nodos, permitiendo maximizar la homogeneidad de grupos resultantes en cada bifurcación.[29] [30]

Redes Neuronales Artificiales

La estructura que conforma la red neuronal está dada por 2 capas ocultas: la primera capa con 26 neuronas (aproximadamente 1,1 × número de variables de entrada = 11 × 1,1 ≈ 12, redondeado a 26 para mayor capacidad expresiva), la segunda conformada por 8 neuronas que actúan como un "embudo" que comprime la información para llevarla de mayor a menor dimensión para la toma de decisión final. Para evitar el problema de gradiente que desvanece se emplea la función de activación ReLU (Rectified Linear Unit), de esta manera favorece el aprendizaje de relaciones no lineales mediante $(f(x) = \max(0, x))$.[31]

Mediante la función layer_normalization() de Keras, se aplicó la normalización de los datos adaptadas principalmente al conjunto de entrenamiento y prueba. Esto permite el escalamiento de cada una de las variables a una distribución estandarizada con media=0 y varianza=1. Además, facilita la aceleración de la convergencia del optimizador, evita el dominio de las variables con altas magnitudes y en los datos de entrenamiento mejora la estabilidad numérica de los datos.[32]

Configuración del Entrenamiento:

Compilación: Optimizador Adam con learning_rate=0,01 (tasa de aprendizaje adaptativa que ajusta dinámicamente la magnitud de actualizaciones de pesos), [33] pérdida binaria cruzada (appropriate para clasificación binaria), métrica de evaluación = accuracy.
Entrenamiento: Mediante 60 épocas (sobre conjunto de entrenamiento), batch_size=32 (64 observaciones procesadas simultáneamente), validation_split=0,3 (30% de datos de entrenamiento reservados internamente para monitoreo de sobreajuste sin contaminar prueba final).
Entorno: Keras/TensorFlow v2.13.0+, Python 3.10.9 en entorno Conda 3 específico, en entorno conda específico (C:/ProgramData/miniconda3/envs/tf), accessed via reticulate package de R.[34] Se entrena de forma independiente los escenarios (a) con n=723 para entrenamiento (desbalanceado) y (b) con n=1117 para entrenamiento (balanceado con SMOTE). Esto permite evaluar el impacto del balanceo sobre a la arquitectura profunda neuronal, así como el desempeño.[35]

Métricas de evaluación.

Mediante las funciones de caret, yardstick, y pROC, se realizó el cálculo de 9 métricas según la siguiente descripción: Tabla 5

Métrica	Fórmula	Interpretación Clínica
Sensitivity (Recall)	TP/(TP+FN)	Proporción de fracasos reales correctamente identificados. Crítica en TB: detectar pacientes en riesgo [36]
Specificity	TN/(TN+FP)	Proporción de éxitos reales correctamente identificados. Evita falsas alarmas innecesarias[36]
Precision (PPV)	TP/(TP+FP)	Proporción de predicciones de fracaso que fueron correctas. Confiabilidad de avisos clínicos[37]
Accuracy	(TP+TN)/(TP+TN+FP+FN)	Proporción total de predicciones correctas. Sesgada en desbalance[37]
F1-Score	2×(Precision×Sensitivity)/(Precision+Sensitivity)	Media armónica. Equilibra precision y recall. Preferible a accuracy en desbalance[37]
Balanced Accuracy	(Sensitivity+Specificity)/2	Promedio de sensibilidad y especificidad. Elimina sesgo de desbalance[37]
AUC-ROC	Área bajo curva sensibilidad vs (1-especificidad)	Discriminación general, independiente del umbral. Rango 0,5-1,0 (0,5=aleatorio, 1,0=perfecto)[37]
Kappa de Cohen	(po-pe)/(1-pe)	Acuerdo más allá del azar. Rango -1 a 1. >0.6=sustancial[3637]
Error de Clasificación	(FP+FN)/(Total)	Proporción de predicciones incorrectas. Complemento de accuracy [3637]

TP=verdaderos positivos (fracasos detectados), TN=verdaderos negativos (éxitos detectados), FP=falsos positivos (éxitos predichos como fracasos), FN=falsos negativos (fracasos no detectados).

Tabla 5.- Métricas empleadas para la evaluación de los modelos.

Por último, es importante mencionar que para la generación de código de programación estadística para R studio, tanto para la generación de gráficas, pruebas y búsqueda de artículos, nos apoyamos de la IA, misma a la que se le solicitó la explicación detallada de las sugerencias entregadas, así como se corroboro en las fuentes de donde se extrajeron. La toma de decisiones finales, luego de un análisis profundo, fue realizado por el equipo que participó en esta investigación.[38, 39]

RESULTADOS

Los resultados del análisis descriptivo exploratorio inicial se detallan a continuación: Tabla 6

VARIBLES	CATEGORIAS	FR	FA	´P (%)	PA (%)
sexo	Femenino	488	488	42,29	42,29
	Masculino	**666**	**1154**	**57,71**	**100**
clasificación del caso por sus antecedentes	Abandono recuperado	18	18	1,56	1,56
	Fracaso	3	21	0,26	1,82
	Nuevo	**1051**	**1072**	**91,07**	**92,89**

	Perdida en el seguimiento	7	1079	0,61	93,5
	Recaída	75	1154	6,5	100
ocupación	Agricultura ganadería	44	44	3,81	3,81
	Estudiante	133	177	11,53	15,34
	Jubilado	12	189	1,04	16,38
	Militar policía	3	192	0,26	16,64
	Ninguno	**304**	**496**	**26,34**	**42,98**
	Oficiales artesanos, operarios	173	669	14,99	57,97
	Otro servicio	11	680	0,95	58,93
	Apoyo administrativo y derecho	60	740	5,2	64,12
	Profesional docente	17	757	1,47	65,6
	Profesionales salud	11	768	0,95	66,55
	Quehaceres Domésticos	225	993	19,5	86,05
	vendedor ambulante	155	1148	13,43	99,48
	vendedor tienda	6	1154	0,52	100
tipo de tuberculosis	Tb criterio clínico epidemiológico	98	98	8,49	8,49
	Tb extrapulmonar	178	276	15,42	23,92
	Tb pulmonar	**878**	**1154**	**76,08**	**100**
resultado del tratamiento	**a: éxito**	**1005**	**1005**	**87,09**	**87,09**
	b: fracaso	141	1146	12,22	99,31
	c: otros: transferido o descartado	8	1154	0,69	100
tamizaje de Vih	a: reactivo a la prueba	45	45	3,9	3.9
	b: no reactivo a la prueba	**560**	**605**	**48,53**	**52,43**
	c: no realizado	549	1154	47,57	100

FR: frecuencia relativa, FA: Frecuencia acumulada, P: Porcentaje PA: Porcentaje acumulado. Los valores con **negrita** representan a los más frecuentes dentro de cada variable analizada.

Tabla 6.- Tabla de frecuencia de la base original total de datos de tuberculosis.

A continuación de muestra las distribuciones por edad de la base original de los casos de tuberculosis. Tabla 7

Frecuencias de afectación de la tuberculosis por grupo de edad.								
GRUPO_EDAD	n	P (%)	$\tilde{x}$	DE	Md	As	KU	EEM
Niño (0-14)	31	2,68%	11	3,61	13	-1,217023	0,4589922	0,64923445
Adolescente (15-17)	78	6,75%	16,1	0,85	16	-0,191523	-1,5963038	0,09585089
Adulto (18-64)	**922**	**79,8%**	**34,4**	**12,8**	**31**	**0,6649787**	**-0,7300555**	**0,42057761**
Adulto Mayor (65+)	123	10,7%	73,8	7,11	73	0,8720057	0,2264921	0,64140192

n: total de observaciones, P (%): Porcentaje,$\tilde{x}$: media muestral, DE: Desviación estándar, Md: Mediana, Asimetría, KU: Curtosis, EEM: Error estándar de la media. Los valores con **negrita** representan al grupo etario más frecuente.

Tabla 7.- Distribución de frecuencias, principales medidas de tendencia central, dispersión y de forma de la base original total de datos de tuberculosis.

Al observarse en la base original total de datos la presencia de varios estratos poblacionales, se decide tomar para el análisis a las observaciones de adultos de 18 a 64 años que conforman el 79.8% del total.

Pruebas de Normalidad.

Para el análisis de las pruebas de normalidad se emplea la variable cuantitativa del peso de los pacientes adultos de 18 a 64 años, y se procede al análisis de las pruebas de normalidad con datos sin y con transformación logarítmica. Las pruebas utilizadas son Shapiro-Wilk y Lilliefors, mismas que presentan una alta potencia y al emplearse en conjunto se realiza una validación cruzada entre estas. Los valores obtenidos en los p-valores, son bajos por lo que se confirma la no normalidad de los datos, reflejándose en los histogramas y en las gráficas de Q-Q-plot. Si bien la transformación logarítmica corrige de cierta manera la desviación con respecto a la normalidad, no fue lo suficiente como se confirma con las pruebas. Tabla 8

	Variable	Prueba	estadístico	P-Value	Conclusión
W	Peso inicial (Kg)	Shapiro-Wilk	0,95573093	4,682885e-16	NO NORMALIDAD
D	Peso inicial (Kg)	Lilliefors	0,08146111	7,385552e-16	NO NORMALIDAD
W2	**Log. Peso inicial (Kg)**	**Shapiro-Wilk**	**0,99389963**	**8,357538e-04**	**NO NORMALIDAD**
D2	**Log. Peso inicial (Kg)**	**Lilliefors**	**0,04448371**	**1,889752e-04**	**NO NORMALIDAD**

W: corresponde al cálculo de la prueba Shapiro-Wilk con la variable Peso Inicial. D: corresponde al cálculo de la prueba Lilliefors con la variable Peso Inicial. En la W2 y D2 se utiliza la variable Peso inicial con trasformación logarítmica para el cálculo de las pruebas respectivas. Los valores en **negrita** destacan las mejoras mostradas con la transformación logarítmica en los datos.

Tabla 8.- Resultados de pruebas de Normalidad

Estos resultados recalcan la necesidad de métodos robustos y/o métodos no paramétricos.

Las métricas comparativas para los modelos de RLB, RF, RNA con 2 capas para la predicción del fracaso terapéutico en pacientes con tuberculosis se evaluaron bajo dos escenarios, tanto para datos originales desbalanceados con el 86% para el éxito y 14% para el fracaso, además de datos balanceados con SMOTE aplicados a la partición de entrenamiento. Es notorio como el desbalance hacia los datos mayoritarios “éxito” sesga los modelos hacia esta, produciendo valores de Sensitivity nula o baja (0-12%) para los modelos RLB y RNA, además con Precision y F1-Score indefinidos o mínimos debido a la ausencia de verdaderos positivos, mientras que Specificity y Accuracy se inflan (hasta 0,9871 y 0,85, respectivamente). Por otro lado, las métricas de los modelos con SMOTE no mostraron mejoras significativas esperadas con la generación de datos sintéticos hacia la clase minoritaria “fracaso”, a excepción de la RLB, elevando drásticamente la Sensitivity (de 0 hasta 44%), Precision (22,45%) y F1-Score (29,73%), con AUC-ROC mejorado, pasando de un rendimiento pobre (0,64) a un rendimiento moderado (0,659) y Balanced Accuracy superando 0,597 en todos los casos, aunque con una ligera reducción en Accuracy global (0,711). RLB emerge como el modelo superior en SMOTE, con el mayor equilibrio entre todos los modelos comparados. Tabla 9

Modelo	Esc.	ACC	K	Sn.	Sp.	Pre.	F1	BAcc	AUCROC	IC 95%	Err.
Reg. Logística	Original	0,85	-0,021	0	0,9871	0	NA*	0,4935	0,6485	(0,5376-0,7595)	0,15
	SMOTE	**0,7111**	**0,1389**	**0,44**	**0,7548**	**0,2245**	**0,2973**	**0,5974**	**0,6591**	**(0,5511-0,767)**	**0,2889**
Random Forest	Original	0,8389	0,0114	0,04	0,9677	0,1667	0,0645	0,5039	0,6319	(0,5256-0,7381)	0,1611

	SMOTE	0,8167	-0,0241	0,04	0,9419	0,1	0,0571	0,491	0,6693	(0,5707-0,7679)	0,1833
RNA (2 Capas)	Original	0,8389	0,1	0,12	0,9548	0,3	0,1714	0,5374	0,619	(0,5047-0,734)	0,1611
	SMOTE	0,8333	0,0476	0,08	0,9548	0,222	0,1176	0,5174	0,601	(0,4743-0,7293)	0,1667

Esc: Escenario, ACC: Accuracy, K:Kappa, Sn: Sensitivity, Sp: Specificity, Pre: Precision, F1: F1-score, BAcc: Balanced Accuracy, AUCROC: Area Under the ROC Curve, IC 95%: Confidence Interval, Err: Classification Error Rate. *Las métricas NA son no calculables dado que presentaron ausencia de verdaderos positivos "fracasos" (sensitivity=0). Los valores en negrita representan a la técnica que mostró el mejor desempeño comparativo.

Tabla 9.- Comparación de resultados obtenidos por cada modelo evaluado para la predicción de casos de fracaso terapéutico en pacientes con tuberculosis.

Por otro lado, al cuantificar las diferencias absolutas delta (Δ) en puntos porcentuales (pp) de los resultados con SMOTE en ambos escenarios de estudio, no muestran mejoras drásticas en la detección de la clase minoritaria. En RLB, SMOTE genera un salto inicial de +44 pp en Sensitivity (de 0% a 44%) y +1,06 pp en AUC-ROC (hasta 0,6591), con una pérdida moderada de -23,2 pp en Specificity ; RF, no mostró diferencias absolutas obteniendo +0 pp en Sensitivity (hasta 4%), -0.74 pp en F1-Score (0,491) y +3,74 pp en AUC-ROC (0,6693), con solo -2,58 pp en Specificity, mientras que RNA mostró -4 pp en Sensitivity (12%), -5,38 pp en F1-Score y -1,76 pp en AUC-ROC, sin diferencias en Specificity y -2 pp en Balanced Accuracy, reafirmando que para el RF y RNA muestran resistencia al balanceo, lo que sugiere la presencia de limitaciones fundamentales en la capacidad de aprendizaje, posiblemente con los datos y variables actuales. A pesar de ello el modelo RLB muestra mejores diferencias absolutas en relación a los otros modelos. Tabla 10

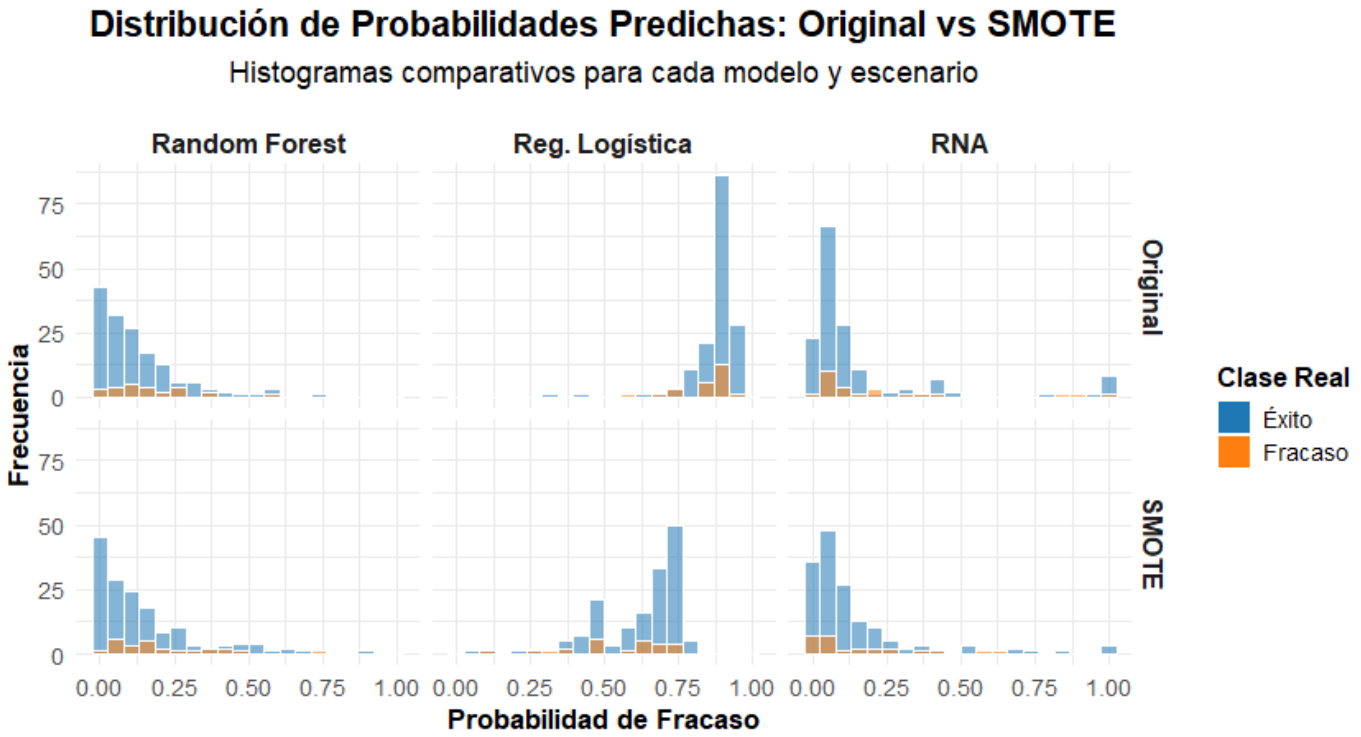
Modelo	Δ Sensitivity	ΔSpecificity	ΔF1-Score	ΔAUC-ROC	ΔBalanced Accuracy
Regresión Logística	**+44 pp**	**-23.2 pp**	**N/A**	**+1,06 pp**	**+10,39 pp**
Random Forest	0 pp	-2,58 pp	-0.74 pp	+3,74pp	-1,29 pp
RNA (2 Capas)	-4 pp	0 pp	-5,38 pp	-1,76 pp	-2 pp

Delta(Δ): Corresponde al cálculo del porcentaje de la diferencia absoluta del valor con SMOTE menos el valor original. (V. SMOTE-Original) x100. pp: Puntos porcentuales. N/A: valor no calculable. (** y con **negrita)**: representan a la técnica que mostró el mejor desempeño comparativo.

Tabla 10.- Análisis de impacto de balanceo SMOTE frente a valores Originales para cada modelo.

Probabilidades predichas

A continuación, se muestra la Figura 5 en donde es evidente las diferencias de las probabilidades para cada modelo estudiado.



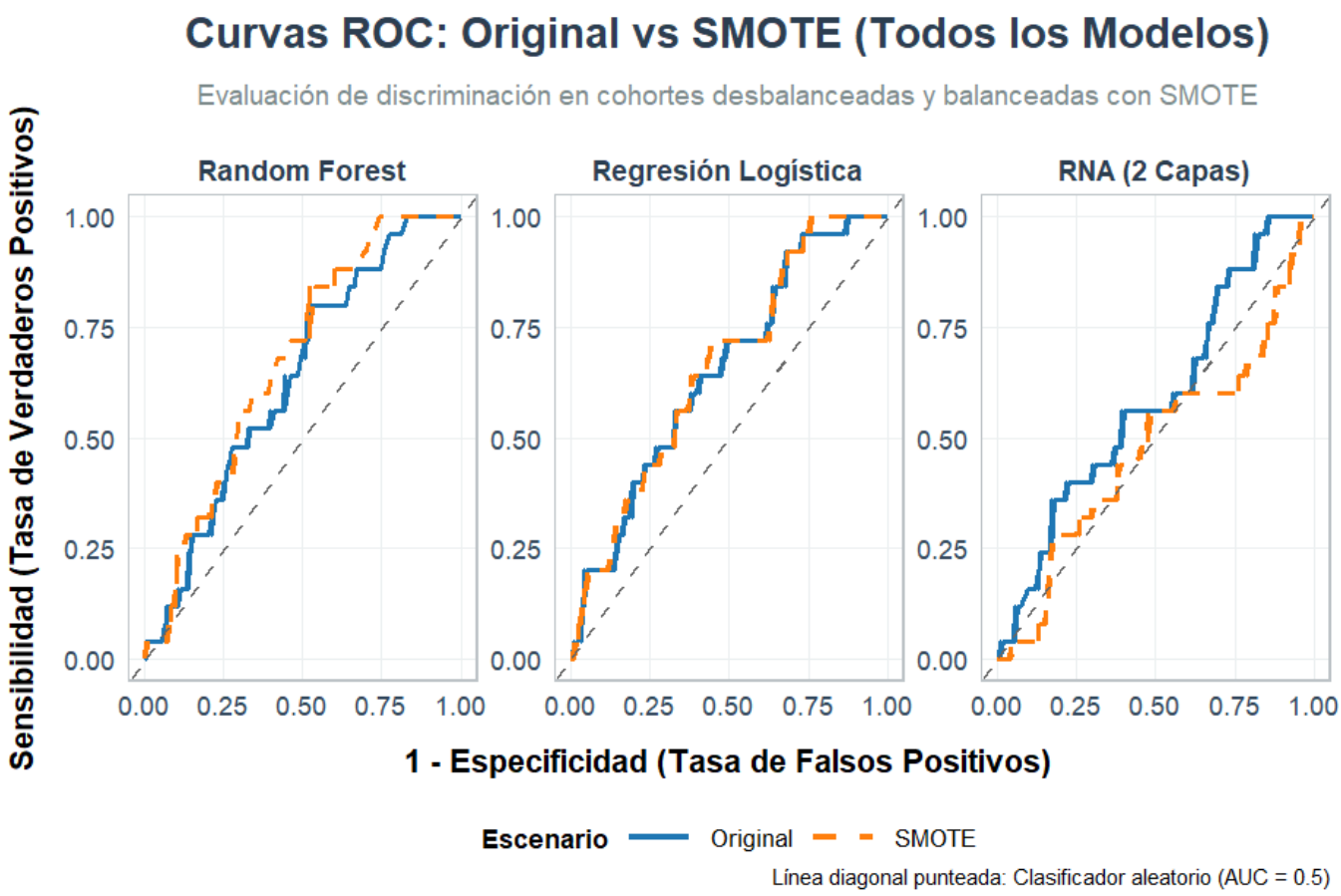
Las Barras representan el Éxito en azul y Fracaso en naranja. Las figuras superiores para los escenarios (a) sin balanceo (b) con balanceo SMOTE.

Figura 5.- Histogramas de probabilidades predichas comparativa

El histograma muestra el predominio de la concentración predictiva hacia la clase mayoritaria (éxito, P<0.2) en los datos originales sin balanceo, en donde se observa concentraciones extremas >90% (0,0-0,2). En cuanto a la aplicación de SMOTE no surgen diferencias significativas en la probabilidad de fracaso, salvo en el modelo RLB SMOTE en donde se observa dispersión significativa (RLB: 23-30% casos P>0,4), evidenciando su relativa superioridad. Por otro lado, RF y RNA mantienen distribuciones unimodales, incluso luego del balanceo lo que nos sugiere la incompatibilidad fundamental entre los algoritmos y los datos sintéticos con SMOTE. Además, es evidente la incapacidad de estos modelos para capturar incertidumbre probabilística por la ausencia de probabilidades intermedias (0,3-0,7).

Capacidad de discriminación.

A continuación, se muestran las curvas ROC y la capacidad de discriminación comparativa con los modelos. Figura 6



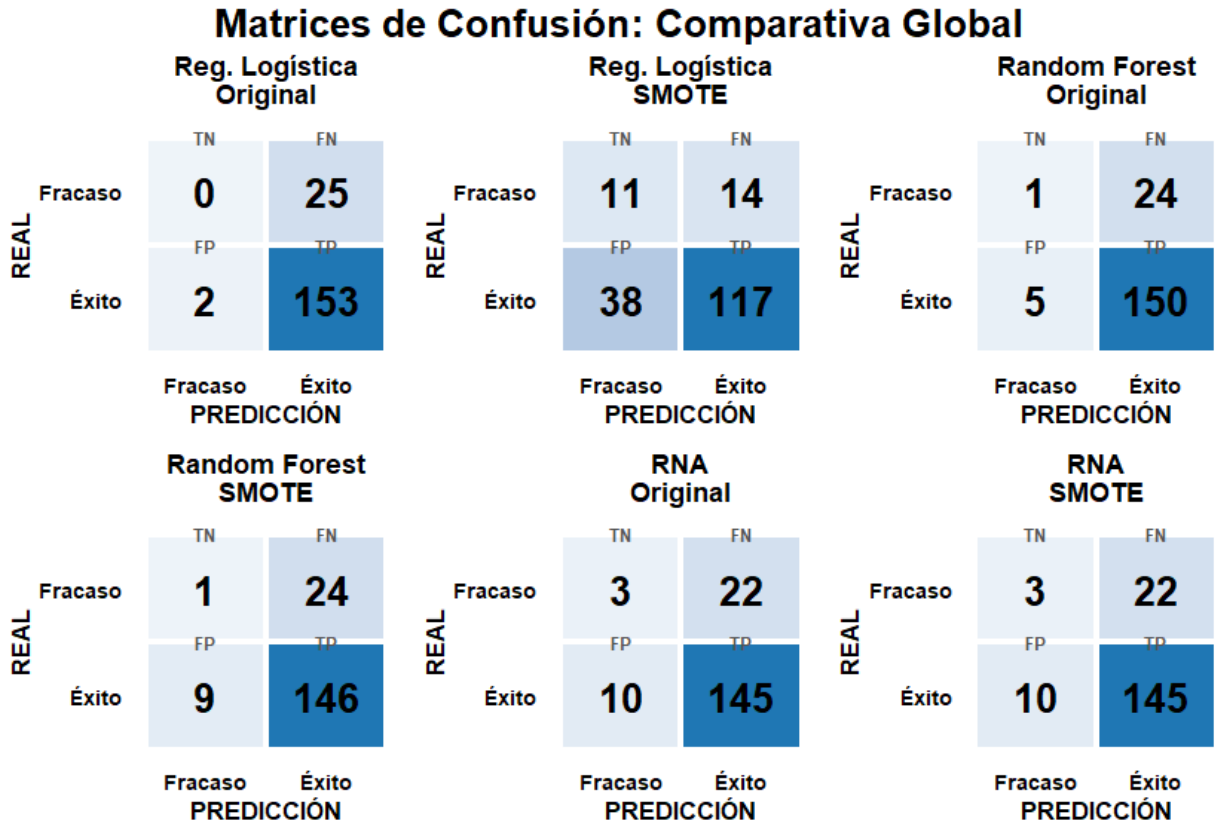
Panel Izquierdo: Random Forest, Panel Central: Regresión logística Binaria, Panel derecho: Redes neuronales artificiales dos capas. Línea azul solida: escenario con datos originales. Línea naranja discontinua: escenario con SMOTE.

Figura 6.- Curvas ROC comparativas para los modelos empleados.

La figura 6 muestra la evaluación de la discriminación en los escenarios sin balanceo y con balanceo SMOTE. El modelo RF con escenario original presenta una capacidad discriminativa moderada, en donde es evidente el desvió de la curva del clasificador aleatorio, pero sin alcanzar una discriminación fuerte (AUC=0,6319). Con la aplicación de SMOTE hay mejoras leves en la capacidad clasificatoria (AUC=0,6693), evidenciándose en la curva naranja un desplazamiento ligero a la esquina izquierda. Si bien el balanceo de clases mejora la diferenciación entre fracaso y éxito, aun el modelo presenta limitaciones significativas para lograr discriminar entre ambas clases. Por otra parte, la RLB muestra métricas muy cercanas al RF mostrando un AUC sin balanceo= 0,6485 y AUC SMOTE= 0,6591. Es evidente que estas curvas presentan un comportamiento más lineal en ambos escenarios, lo que nos podría sugerir que el modelo captura patrones más generales, pero presenta limitaciones en la capacidad de ajustarse ante casos minoritarios, así como balanceados. En cuanto a las RNA el desempeño es muy menor y más problemático en relación a los otros modelos. En el escenario original el AUC = 0,5768, es levemente superior al clasificador aleatorio, mientas que en el escenario con SMOTE se presenta un fenómeno de degradación del rendimiento con un AUC= 0,5112. Esto nos sugiere que el modelo RNA con 2 capas, resulta insuficiente para capturar la complejidad del problema al incorporar datos sintéticos de SMOTE, generando ruido y degradando aún más la capacidad de generalización del modelo.

Distribución de predicciones

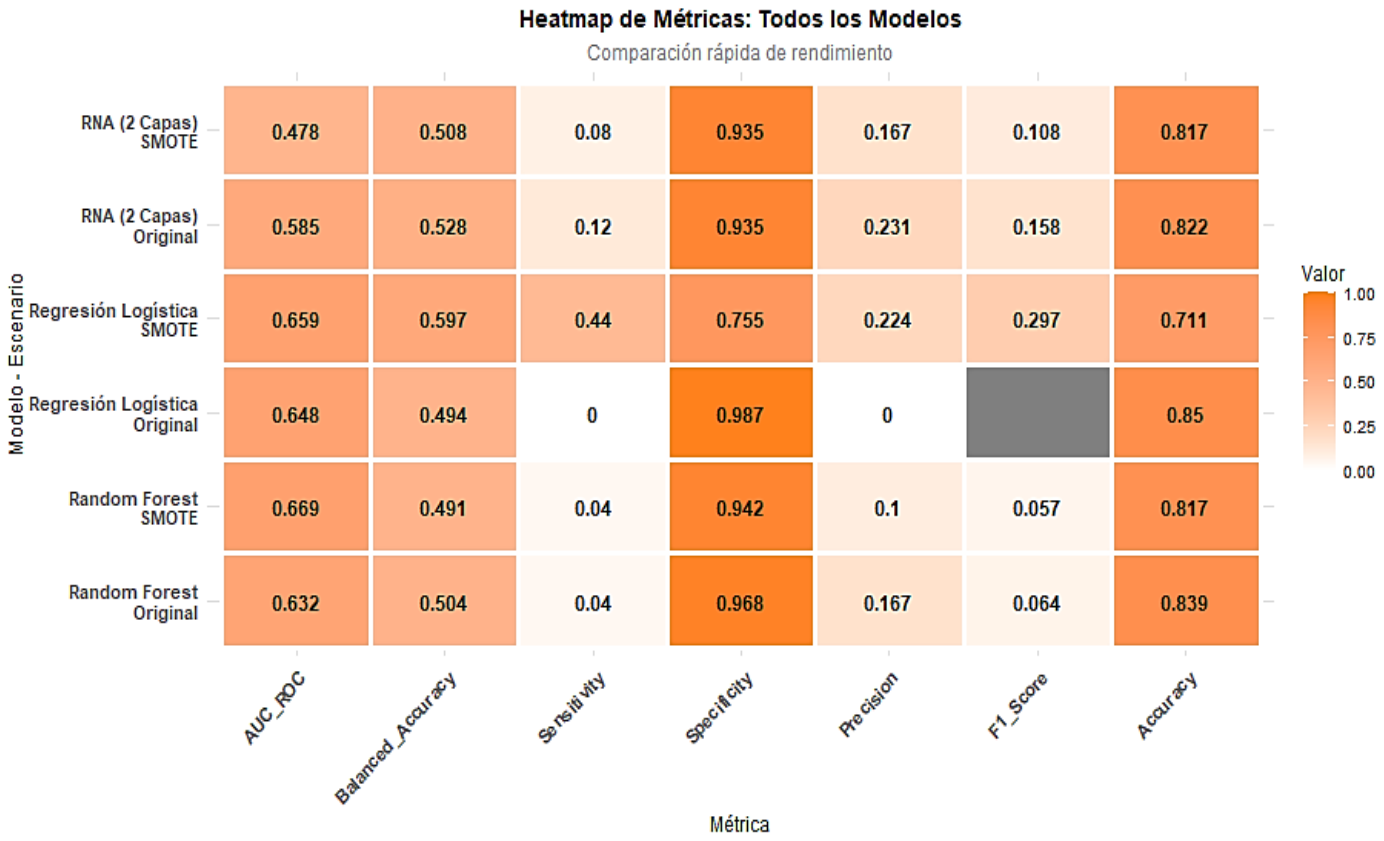
Las matrices de confusión comparativas en los dos escenarios de estudio revelaron el sesgo a la clase mayoritaria en todos los modelos, pese a la aplicación del balanceo SMOTE. La RLB original clasificó correctamente 153 éxitos (TP) pero no logró detectar 25 fracasos reales (FN=25, TN=0), esto produjo una sensibilidad de cero, confirmando la incapacidad de identificar casos críticos clínicamente. Al aplicar SMOTE de evidencia una mejora en la sensibilidad (44%) dado que se produce una elevación en la detección de fracasos reales pasando a lograr detectar 11 fracasos (TN=11), aunque se incrementan los falsos positivos a 38, mostrando los beneficios del balanceo. En cuanto al RF mantiene una sensibilidad baja TN=1 tanto para modelos balanceados y no balaceados. Para la RNA logra una detección moderada (TN=3) sin mejorar pese al balanceo. Esto demuestra que el modelo tiene limitaciones serias en la predicción de casos de tuberculosis en nuestro estudio, al presentarse restricciones serias debido al desbalanceo y la presencia de datos sintéticos. Esto muestra que ningún modelo supera una sensibilidad del 44% con falsos negativos persistentes de 14-25 por matriz. Figura 7



Eje X: valores reales (fracaso-éxito), Eje Y: valores predichos (fracaso-éxito), tablas para datos no balanceados(original) y balanceados (SMOTE).

Figura 7.- Matriz de confusión del desempeño en la predicción.

En la Figura 8 se muestra el desempeño comparativo de las métricas de todos los modelos para los dos escenarios estudiados. Se evidencia que ningún modelo alcanza un rendimiento optimo para poder ser aplicado clínicamente. RF con SMOTE obtiene el AUC-ROC más alto (0,669), mientras que la RLB con SMOTE logra la mejor sensibilidad (0,44) y el mayor F1-Score (0,297), siendo este el único modelo que mejora la detección de fracasos tras el balanceo, aunque sacrificando especificidad y precisión. Por su parte la RLB original (sin balanceo) muestra una especificidad más alta (0,987) y un accuracy mayor (0,85), pero con nula sensibilidad y F1-Score, lo que nos estaría indicando un fuerte sesgo hacia la clase mayoritaria de éxito con incapacidad total para detectar fracasos en el tratamiento. Las RNA por su parte muestran en ambos escenarios métricas que la sitúan en puntos intermedios con respecto a los otros modelos, si bien los valores de exactitud son aceptables (=0,82) y especificidades altas (>0,93), las sensibilidades son bajas (0,12 y 0,08) y F1-Score modestos. Esto confirma que su capacidad discriminativa es limitada pero no supera a la RLB con balanceo.

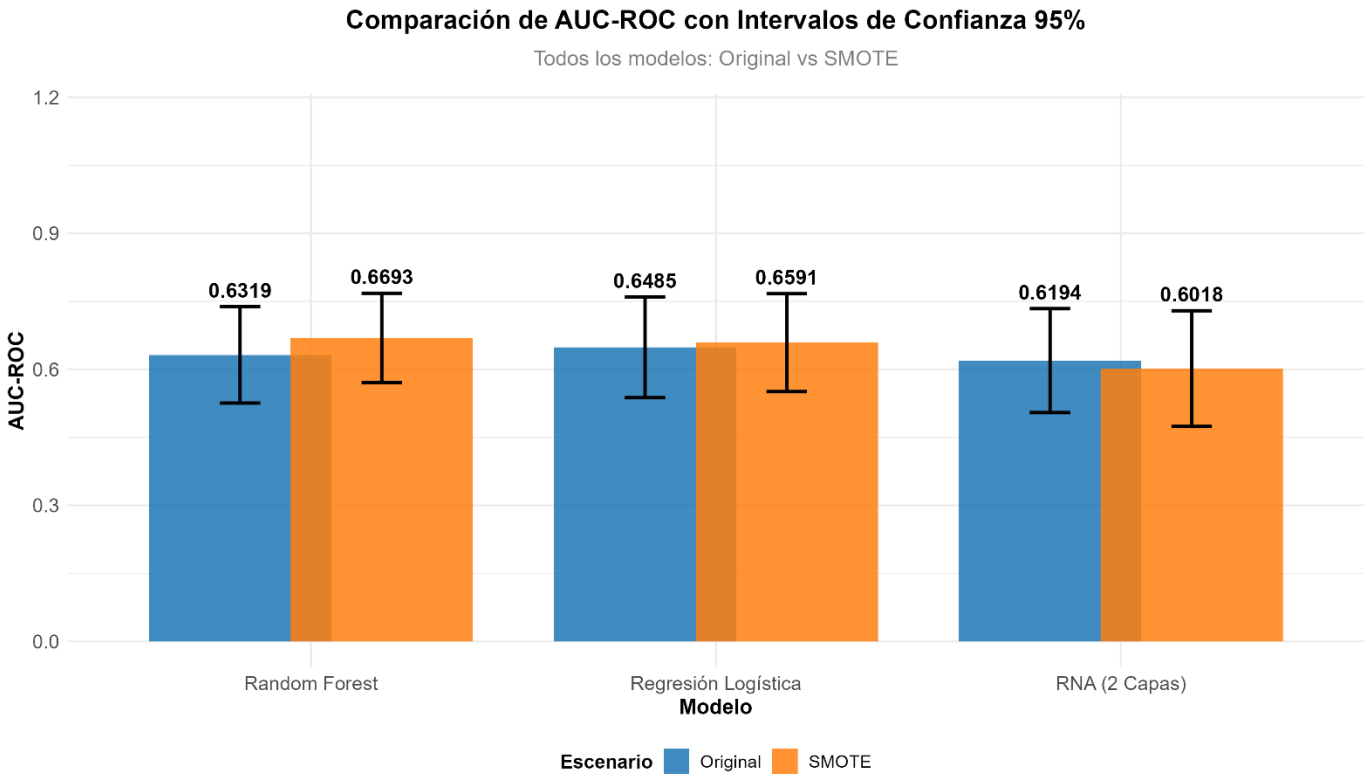


Tonalidad naranja oscuro: valores de 0.9+, tonalidades blancas/grises: valores entre 0 a 0.3

Figura 8.- Desempeño comparativo de todos los modelos basados en sus métricas de evaluación.

Validación con intervalos de confianza.

Se realiza un análisis de los intervalos de confianza basados en al AUC-ROC, mismo que es obtenido con la función `ci.auc()` de pROC. Figura 9



Barras azules: Corresponde a datos originales, Barras naranja: corresponde a datos con SMOTE.

Figura 9.- Comparación del AUC-ROC e intervalos de confianza para cada escenario y modelo empleado.

En la figura 9 es evidente que la implementación de SMOTE en los modelos genera efectos divergentes, para el modelo RF se observa la mejora mas notable en donde su capacidad de discriminación pasa de 0,63 a 0,67 (IC 95%: 0,57–0,77), en la RNA con 2 capas sucede todo lo contrario, en donde baja su capacidad de discriminación de 0,62 a 0,60 (IC 95%: 0,47–0,73). El modelo que mantiene mayor estabilidad es la RLB con un incremento marginal de 1,06 pp, situándose en 0,66 (IC 95%: 0,55–0,77). Ante estos resultados obtenidos es evidente que las diferencias en el rendimiento obtenido en el AUC-ROC no son concluyentes para ambos escenarios, por lo que mantiene a los modelos en un rango de desempeño clasificatorio moderado.

DISCUSIÓN

La tuberculosis es una enfermedad multifactorial que requiere de varias intervenciones de salud. Estas van desde una adecuada detección de signos y síntomas, posteriormente la realización de pruebas diagnósticas adecuadas, tamizaje de los contactos del caso índice e iniciar de forma rápida y oportuna el tratamiento.[1] Lograr la adherencia al tratamiento antituberculoso evitando el abandono o fracaso terapéutico, surge como una de las principales estrategias para erradicar esta enfermedad.[40] Diversos estudios a nivel mundial han tratado de crear modelos predictivos que logren de forma precisa identificar casos con alto riesgo de fracaso terapéutico. Este estudio se enfocó en el desarrollo de un modelo predictivo, que adaptado al contexto del sub trópico ecuatoriano y con datos provenientes de tarjetas de control terapéutico de un centro de salud, permite la detección de casos con alto riesgo de fracasar en su tratamiento desde el primer contacto. Esto favorece a que los trabajadores sanitarios pongan más énfasis en el seguimiento y control de casos con alto riesgo.

Situación epidemiológica de la tuberculosis.

Esta investigación permite observar la presencia clara de tuberculosis a lo largo de los años evaluados. Se determina una tasa de prevalencia aproximada de 44 casos/100,000 habitantes, en donde el sexo masculino con el 57,1% es el más afectado. Las edades más frecuentes fueron de 18 a 64 años con el 79,8%, siendo la población económicamente activa con mayor porcentaje. Es importante mencionar que el 26,34% de los afectados refiere encontrarse en el desempleo o en situación de calle. Por otro lado, la tuberculosis pulmonar con el 76,08% fue la que más presencia tuvo. Las comorbilidades reportadas (Diabetes Mellitus, Hipertensión, cáncer, etc) correspondieron al 3,21%, así como los pacientes con coinfección VIH en un 3,9%, es importante mencionar que la falta de registro de estos datos en las tarjetas físicas, así como la no realización de estos tamizajes años atrás a toda la población, podría influir en un subregistro y baja detección de estos casos. La mayor parte de pacientes que ingresaron al programa tuvieron una tasa de éxito terapéutico del 87,09% mientras el 12,22% fracasaron. Es importante recalcar que los que reportaron antecedentes previos de fracaso terapéutico, el 100% volvió a fracasar en un nuevo ciclo, contrastándose con varios estudios y reportes.[41] [5] Para evaluar la eficacia terapéutica en el contexto de programas locales de tuberculosis la OMS recomienda superar el 85% de curaciones para considerarlo como un programa exitoso.[1]

La regresión logística como punto de partida.

Un modelo predictivo base que se empleó fue la RLB, dado que permite generar resultados binarios con alto potencial para producir razonablemente ajustes que versan sobre las razones de verosimilitud. La característica de RLB es que cada coeficiente cuantifica el log-odds del evento de interés, ajustados por los demás factores incluidos.[42] Los supuestos del modelo como la linealidad en el logit, independencia, ausencia de colinealidad severa se evaluaron mediante VIF y pruebas de residuos. Estos permitieron comprobar los supuestos de forma aceptable, además se logró la detección de potenciales interacciones no lineales sobre predictores clínicos como Peso-Edad, entre otros.[43]Si bien la regresión logística no puede, en su forma estándar, capturar interacciones complejas, su claridad a permitido la consolidación de su uso en la medicina.[44]

Regresión logística frente a los modelos Random Forest y Redes Neuronales Artificiales.

Un problema real y común que se encontró en este estudio es la presencia de datos no balanceados, es decir datos que presentan un mayor peso hacia una sola clase y que pueden sesgar el aprendizaje y predicción de los modelos. Con la aplicación de la técnica de balanceo SMOTE permitió mejoras leves sobre la clase minoritaria, pero pese a ello fueron insuficientes para la detección efectiva de fracasos, lo que conlleva a serias dificultades para la aplicación clínica. La RLB tras la aplicación de SMOTE aumentó la sensibilidad y AUC-ROC, pero sacrificando especificidad y exactitud. Pese a ello la RLB mostro ser el modelo con mayor equilibrio con respecto a los otros modelos más complejos, mientras que la RF y RNA, mostraron una baja sensibilidad con alta especificidad, pese a la aplicación del balanceo que corrigió levemente y en algunos casos como en la RNA, la introducción de datos sintéticos, generó ruido, saturó y degradó la capacidad discriminativa y de generalización del modelo. Esto se debe fundamentalmente a que estos modelos tienen una alta capacidad para modelar relaciones no lineales y patrones multivariantes complejos, y en este estudio una de las principales dificultades fue la obtención de variables clínicas predictivas que logren mostrar la mayor cantidad de patrones detectables.[45] Por otro lado la obtención de observaciones limitadas de la clase fracaso hace que el modelo pese al balanceo capte datos sintéticos, lo que dificulta el entrenamiento y más aún cuando el desbalanceo de las clases es 6:1.

El desempeño limitado de los modelos tras la aplicación de SMOTE se podría deber a factores estructurales y de generalización específicos de cada arquitectura. Para el RF, su robustez intrínseca ante el desbalanceo y el uso del criterio de Gini pudieron generar que el modelo deje de aprender de la clase fracaso dado que logró

memorizar las pocas observaciones entregadas, donde la creación de 394 ejemplos sintéticos terminó por diluir la información real, causando que posiblemente el modelo aprenda y memorice patrones de interpolación en lugar de relaciones causales. Por otro lado, las métricas de la RNA fueron más críticas, evidenciado por colapso del AUC; esto sugiere un overfitting severo sobre el ruido sintético, potenciado por una capacidad expresiva limitada que podría deberse a su arquitectura de dos capas, llevándole a entrega de métricas cercanos al azar. [9] [19] [46] [42]

Diversas investigaciones respaldan el uso del balanceo de clases dado que mejora la sensibilidad en la detección de la clase minoritaria como los fracasos, informando que se puede evidenciar mejoras hasta en 40 puntos porcentuales, pero otros métodos como el de ensambles, pueden dar mejoras superiores.[47] Por otro lado en escenarios en donde existe la prioridad en la explicación del riesgo clínico, la RL es preferida por su alta capacidad de interpretabilidad y calibración probabilística.[48]

Variables clínicas predictoras y su importancia.

Es posible confirmar el respaldo clínico y epidemiológico de las 11 variables predictoras empleadas. Se identifica a la edad en que se presenta la tuberculosis, el sexo masculino y otros determinantes de exposición social como factores que elevan la probabilidad del fracaso terapéutico.[49] El peso inicial y la ganancia de peso durante el tratamiento, además de ser marcadores robustos de respuesta, están relacionados con disminución de los riesgos y mejores tasas de curación, lo que se alinea con el cumplimiento de las recomendaciones en los estudios longitudinales de TB y las guías de soporte nutricional que se encuentran en uso a nivel internacional.[50]Los antecedentes de recaída, abandono y pérdida de seguimiento, además de ser precursores de eventos adversos, son reconocidos como factores de riesgo a considerar para la focalización de las intervenciones de adherencia. El desempleo y el vivir con VIH se han consolidado como determinantes estructurales de vulnerabilidad y mal desenlace, mientras que la TB extrapulmonar sigue asociándose con retraso diagnóstico y mayor complejidad terapéutica en el tratamiento.[51] La consideración conjunta de estas variables mejora la personalización de los esquemas de tratamiento y seguimiento, así como la toma de mejores decisiones clínicas.
Como es evidente las variables predictoras implementadas están altamente asociadas al fracaso terapéutico, sin embargo, en este estudio se puede evidenciar que aun son insuficientes para lograr mostrar todos los factores sociodemográficos, conductuales y clínicos que están inmersos en el abandono terapéutico en los pacientes con tuberculosis.

Comparación de resultados con otros estudios.

Estos resultados se alinean con estudios recientes que indican que el RL actúa como un estándar para la interpretación clínica pudiendo ser la base de partida ideal para el desarrollo de varios modelos. En otros estudios se han descrito mejoras consistentes en F1-score y AUC-ROC tras el balanceo.[45] La mayor parte de la literatura internacional aboga de manera consistente por la inclusión de curvas de calibración y el puntaje de Brier junto a métricas estándar, especialmente cuando los resultados están destinados a uso clínico o sistemas de alerta automatizados.[52]En general, las recomendaciones para estudios relacionados con la salud pública y la TB, la RL puede ser más adecuado debido a su facilidad de uso y contexto local, mientras que RF y RNA deberían ser incorporados cuando la sensibilidad operacional es la prioridad y existe apoyo institucional para la interpretación y gestión de una carga mayor de falsos positivos.[53]
Por otro lado la implementación práctica de modelos de RL que se encuentren bien calibrados, ofrecen un alto valor para la personalización del monitoreo y seguimiento de pacientes en programas de TB, facilitando la priorización de la asignación de recursos, la adherencia, y la planificación estratégica de seguimiento en visitas

para pacientes de alto riesgo.[44 47 48]Sin embargo, el aumento de casos detectados mediante modelos de RF y RNA debe equilibrarse con el costo de los falsos positivos que se presentarán. Como siempre, la carga de la enfermedad en el sistema de salud y el contexto operativo en el que se aplique, determinan la capacidad de la detección de nuevas alertas y el incremento de intervenciones más enfocadas.[47 50]

Limitaciones encontradas.

Las principales limitaciones en este estudio se basaron en la obtención de los datos, debido a que la mayor parte de información se la encontró de forma física en las tarjetas de control terapéutico, lo que obligó a la digitalización de forma manual. Además, se observaron datos faltantes por lo que se realizó la búsqueda en las historias clínicas, pudiendo incurrir en sesgos de selección y generando incertidumbre en la generalización de estos resultados.[43 52] Para la adaptación de diferentes escenarios epidemiológicos es indispensable reportar el análisis, la calibración y sobreajuste que permitan la validación externa. Para investigaciones posteriores se debe evaluar el impacto institucional de los resultados obtenidos.

Futuras líneas investigativas.

Las futuras investigaciones podrían centrarse en la validación externa de estos resultados que permita la generalización en el contexto latino americano y más específicamente en el subtrópico ecuatoriano. Es importante lograr integrar más variables predictoras que permitan robustecer los modelos, permitiendo mejor explicabilidad con modelos más avanzados, así como la generación de modelos piloto que permitan medir el impacto de los resultados de salud en el contexto comunitario.

CONCLUSIONES

En esta investigación identificó a varias variables que podrían tener un alto impacto en la predicción del fracaso terapéutico, siendo estas la edad del paciente, el peso al inicio de la terapia, el ser de sexo masculino y sobre todo si existen antecedentes previos de recaída, abandono, fracaso o pérdida del seguimiento del caso, entre otros. Además, se recalca la importancia del análisis estadístico multivariante y la RLB como técnicas y modelos de baja complejidad, de fácil aplicación en contextos de salud y de alto poder estadísticos para ser usados como línea base previos al desarrollo de modelos más complejos de IA. Esta investigación además permitió identificar a la RLB como el modelo que mayor estabilidad predictiva mostró (Accuracy: 0,7111, Sensitivity: 0,44, Specificity: 0,7548 Precision: 0,2245 F1-Score: 0,2973 AUC-ROC:0,6591 Kappa: 0,1389), frente a modelos de alta complejidad como RNA con 2 capas y la RLB, mostrando una mayor sensibilidad, baja especificidad, pero más equilibrado para la detección de casos con alto riesgo de fracaso terapéutico. Pese a estas métricas el uso en la práctica clínica se ve muy limitado, dado que aún son bajas, pero sirven como punto de partida para realizar las mejoras en los modelos, integrando más datos en la clase minoritaria fracaso, introducción de más variables predictivas y con la validación prospectiva antes de la aplicación clínica.

Funding Statement This research received no external funding.

Conflict of Interest Statement The authors declare no conflict of interest.

Author Contributions: For research articles with multiple authors, a short paragraph outlining each author's individual contributions must be provided. The following statements should be used: Conceptualization, Armijos A. and Sánchez N.; methodology, Armijos A. and Sánchez N.; software, Armijos A.; validation, Armijos A. and Sánchez N.; formal analysis, Armijos A.; investigation, Armijos A. and Sánchez N.; resources, Armijos A.; data curation, Armijos A.;

writing—original draft preparation, Armijos A.; writing—review and editing, Armijos A. and Sánchez N.; visualization, Armijos A.; supervision, Sánchez N.; project administration, Armijos A.

Funding: This research received no external funding.

Acknowledgments: We would like to express our sincere gratitude to the Nurse Dayse Zambrano Parraga (retired), who for several years monitored the administration of anti-tuberculosis therapies. Her exceptional organization of therapeutic management cards and digital matrices allowed for a faster and more accurate data collection process.

Conflicts of Interest: The authors declare no conflicts of interest.

Data Availability Statement The data used in this study belong to the Augusto Egas Type C Health Center of the Ministry of Public Health, Santo Domingo de los Tsáchilas, Ecuador. Their use was authorized under official document: MSP-CZ4S-DDS-Nº 23D01-2025-2213-O, dated June 5, 2025. The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request and subject to institutional regulations.

Institutional Review Board Statement The study was conducted in accordance with the Declaration of Helsinki and approved by the corresponding health authorities of the Ministry of Public Health of Ecuador (Document MSP-CZ4S-DDS-Nº 23D01-2025-2213-O).

Informed Consent Statement Patient consent was waived due to the retrospective nature of the study and the use of de-identified secondary data, as authorized by the institutional regulatory body.

AI-Assisted Tools Disclosure No artificial intelligence system was used to generate, manipulate, or analyze experimental data, images, or statistical output in this study. All quantitative assessments were performed directly by the authors using validated scientific methods. Generative AI tools were used exclusively for minor linguistic refinement and formatting standardization of the manuscript, under full human supervision. The authors independently verified all results, analyses, and conclusions in accordance with BioNatura Journal's policy: https://bionaturajournal.com/artificial-intelligence--ai-.html

REFERENCIAS

1. SALUD[OMS] OMDELA. Estrategia Fin a La Tuberculosis. Ginebra; 2023.
2. Data WHO. Tuberculosis Incidence, Ecuador. 2024.
3. P MS. Guía de Práctica Clínica (GPC) Tamizaje y Diagnóstico de La Tuberculosis. 2024.
4. Argentina O, Roca M, Moreno AJ. Abandono al Tratamiento Antifímico En Pacientes Con Tuberculosis Atendidos En Un Centro de Salud Público de Guayaquil. n.d.
5. Culqui DR, Munayco E. C V, Grijalva CG, et al. Factores asociados al abandono de tratamiento antituberculoso convencional en Perú. Arch Bronconeumol 2012;48(5):150–155; doi: 10.1016/j.arbres.2011.12.008.
6. S OM. Ethics and Governance of Artificial Intelligence for Health. World Health Organization; 2024.
7. Vertti J. Análisis Multivariado. Primera. 2019: Aguas Calientes; 2019.
8. Sánchez E., Tenesaca S. Modelo logístico y redes neuronales para pronóstico de anemia en menores de 5 años en el hospital pediátrico Alfonso Villagómez Román en el periodo 2020-2021. Riobamba; 2023.
9. Fernández A. Guía Completa Sobre Random Forest. 2024. Available from: https://anderfernandez.com/blog/guia-completa-random-forest/ [Last accessed: 10/27/2025].
10. Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks 2015;61:85–117; doi: https://doi.org/10.1016/j.neunet.2014.09.003.
11. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–444; doi: 10.1038/nature14539.

12. Montero LR, Bastián JA, Sanpablo AIP. Classification of Daily Living Activities in subjects with Parkinson's Disease using Artificial Neural Networks. Revista Mexicana de Ingenieria Biomedica 2023;44(4):128–139; doi: 10.17488/RMIB.44.4.9.
13. McClean M, Panciu TC, Lange C, et al. Artificial intelligence in tuberculosis: a new ally in disease control. Breathe 2024;20(3); doi: 10.1183/20734735.0056-2024.
14. INEC. Clasificación internacional uniforme de ocupaciones. 2010.
15. Briones H, Martínez V. Análisis factorial exploratorio del instrumento para medir el impacto del COVID-19 en estudiantes de educación superior Exploratory factor analysis of the instrument to measure the impact of COVID-19 on higher education students. Revista Espacios 2025;46:85; doi: 10.48082/espacios-a25v46n04p08.
16. Peduzzi P, Concato J, Kemper E, et al. A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. 1996.
17. Ortega M., Dominguez A. REGRESIÓN LOGÍSTICA NO CONDICIONADA Y TAMAÑO DE MUESTRA: UNA REVISIÓN BIBLIOGRÁFICA. RevEspSalud Publica 2002;76:1–9.
18. Ochoa C., Molina M., Ortega E. Regresión logística múltiple. Evidencias en Pediatría 2023;19:1–6.
19. Liebl S., Lemaître G., Nogueira F. 2.1.1. Naive Random over-Sampling; 2.1.2. From Random over-Sampling to SMOTE and ADASYN. In: Imbalanced-Learn 0.14.0. New York; 2025.
20. Wadhwa K, Kumari R, Gosain A. Enhancing Model Performance in Hybrid Class Imbalance Techniques. In: Procedia Computer Science Elsevier B.V.: Delhi; 2025; pp. 288–297; doi: 10.1016/j.procs.2025.04.266.
21. Menardi G., Torelli N. ROSE: Random Over-Sampling Examples. Viena, Australia.; 2021.
22. Lemaitre G, Nogueira F, Aridas C. Over-Sampling — Version 0.14.0. 2025.
23. Páez O, Sangrador C O, Arias M. Regresión logística binaria simple. Evidencias en Pediatria 2022;18:2–9.
24. Piury Pinzón J, Cayuela Rodríguez L, Cayuela Domínguez A, et al. Regresión logística binaria para clínicos poco amantes de las matemátiicas. NURE Investigación 2024;21:3–8; doi: 10.58722/nure.v21i131.2553.
25. Kuhn M, Wickham H. Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles. R Package Version 1.1.1. 2024. Available from: https://www.tidymodels.org/ [Last accessed: 10/23/2025].
26. Hosmer D, Lemeshow S, Sturdivant R. Applied Logistic Regression (3rd Ed.). 3rd ed. (John Wiley, & Sons. eds). 2013.
27. Probst P, Wright M, Boulesteix A. Hyperparameters and Tuning Strategies for Random Forest. Revista ISPRS de fotogrametría y teledetección 2019;9; doi: 10.1002/widm.1301.
28. Canty A, Wiener M. Classification and Regression by RandomForest. 2002.
29. Silfiana L., Asyifah Q., Rafika A., et al. Optimizing Random Forest Parameters with Hyperparameter Tuning for Classifying School-Age KIP Eligibility in West Java. Jambura Journal of Mathematics 2025;7; doi: https://doi.org/10.37905/jjom.v7i1.28736.
30. Liaw A, Wiener M. Classification and Regression by RandomForest. 2002.
31. Nair V, Hinton G. Rectified Linear Units Improve Restricted Boltzmann Machines. Israel; 2010.
32. Ba JL, Kiros JR, Hinton GE. Layer Normalization. ArXiv 2016.
33. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. ArXiv 2017.
34. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. 2016.

35. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (3rd Ed.). 2022.
36. Pongsuwun K, Puwarawuttipanit W, Nguantad S, et al. A Systematic Review of the Accuracy of Machine Learning Models for Diagnosing Pulmonary Tuberculosis: Implications for Nursing Practice and Implementation. Nurs Health Sci 2025;27(1); doi: 10.1111/nhs.70077.
37. Guerrero DA, Código B. Aplicación de Modelos Machine Learning para predecir el riesgo de pérdida de seguimiento en tuberculosis. Cali, Colombia; 2025.
38. Perplexity AI. Asistencia en búsqueda de artículos médico y resumen interpretativo. 2025. Available from: https://www.perplexity.ai [Last accessed: 10/28/2025].
39. Perplexity AI. Asistencia en generación de código para preprocesamiento de datos para R studio. 2025. Available from: https://www.perplexity.ai [Last accessed: 10/28/2025].
40. Jim A, González C. La Tuberculosis: una mirada desde la Atención Primaria de Salud. 2024;4–10.
41. de Lucena LA, Dantas GB da S, Carneiro TV, et al. Factors Associated with the Abandonment of Tuberculosis Treatment in Brazil: A Systematic Review. Rev Soc Bras Med Trop 2023;56; doi: 10.1590/0037-8682-0155-2022.
42. Melo Villalobos B., Weber S. Regresión logística en estudios epidemiológicos de casos y controles. Bogotá, Colombia; 1992.
43. Orwa J, Oduor P, Okelloh D, et al. Comparison of logistic regression with regularized machine learning methods for the prediction of tuberculosis disease in people living with HIV: cross-sectional hospital-based study in Kisumu County, Kenya. Res Sq 2023; doi: 10.21203/rs.3.rs-3354948/v1.
44. Tervi A., Junna N., Broberg M., et al. Large registry-based analysis of genetic predisposition to tuberculosis identifies genetic risk factors at HLA. Hum Mol Genet 2023;32(1); doi: doi:10.1093/hmg/ddac212.
45. Phat NK, Lee Y, Vu DH, et al. Risk factors for tuberculosis treatment outcomes: a statistical learning-based exploration using the SINAN database with incomplete observations. BMC Med Inform Decis Mak 2025;25(1); doi: 10.1186/s12911-025-03139-9.
46. Lino Ferreira da Silva Barros MH, Alves GO, Morais Florêncio Souza L, et al. Benchmarking machine learning models to assist in the prognosis of tuberculosis. Informatics 2021;8(2); doi: 10.3390/informatics8020027.
47. Asad M, Mahmood A, Usman M. A machine learning-based framework for Predicting Treatment Failure in tuberculosis: A case study of six countries. Tuberculosis 2020;123:101944; doi: https://doi.org/10.1016/j.tube.2020.101944.
48. Sekandi JN, Shi W, Zhu R, et al. Application of Artificial Intelligence to the Monitoring of Medication Adherence for Tuberculosis Treatment in Africa: Algorithm Development and Validation. JMIR AI 2023;2:e40167; doi: 10.2196/40167.
49. Umeta AK, Yermosa SF, Dufera AG. Bayesian parametric modeling of time to tuberculosis co-infection of HIV/AIDS patients at Jimma Medical Center, Ethiopia. Sci Rep 2022;12(1):16475; doi: 10.1038/s41598-022-20872-7.
50. Londoño Ruiz AM., Castaño Quintero AE. Regresión logística. 2025. Available from: https://es.scribd.com/document/485532978/practica-de-regresion-logistica-abandono-TB-2 [Last accessed: 10/27/2025].
51. Orjuela-Cañón AD, Jutinico AL, Awad C, et al. Machine learning in the loop for tuberculosis diagnosis support. Front Public Health 2022;Volume 10-2022.

52. Hossain MdS, Khandocar MdP, Riti FA, et al. A comprehensive machine learning for high throughput Tuberculosis sequence analysis, functional annotation, and visualization. Sci Rep 2025;15(1):25866; doi: 10.1038/s41598-025-98654-0.

Received: 12 Dec 2025 / **Accepted:** 21 Jan 2026 / **Published (online):** 15 Mar 2026 (Europe/Madrid)
Citation. Armijos A., Sánchez N. Modelos impulsados por inteligencia artificial para la predicción de la respuesta al tratamiento antituberculoso: estudio de cohorte retrospectivo en el subtrópico ecuatoriano. BioNatura Journal. 2026; 3(1): 2. https://doi.org/10.70099/BJ/2026.03.01.2
Additional Information Correspondence should be addressed to: alexmedicomauricio65@gmail.com
Peer Review Information BioNatura Journal thanks the anonymous reviewers for their valuable contribution to the peer-review process. Regional peer-review coordination was conducted under the BioNatura Institutional Publishing Consortium (BIPC), involving:

- Universidad Nacional Autónoma de Honduras (UNAH)
- Universidad de Panamá (UP)
- RELATIC (Panama)

Reviewer selection and assignment were supported via: https://reviewerlocator.webofscience.com/
Publisher Information Published by Clinical Biotec S.L. (Madrid, Spain) as the publisher of record under the BioNatura Institutional Publishing Consortium (BIPC). Institutional co-publishers:

- UNAH (Honduras)
- UP (Panama)
- RELATIC (Panama)

Places of publication: Madrid (Spain); Tegucigalpa (Honduras); Panama City (Panama) **Online ISSN:** 3020-7886