Aplicación de Modelos *Machine Learning* para predecir el riesgo de pérdida de seguimiento en tuberculosis

Diana Azucena Guerrero Barreto código 8985852
Rubén Darío Rodríguez Camargo código 8986356

Proyecto Aplicado para optar al título de Magister en Ciencia de Datos

Directora Delia Ortega Lenis

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
FEBRERO 2025

TABLA DE CONTENIDO

Pág.

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE ANEXOS

INTRODUCCIÓN

La tuberculosis (TB) es una enfermedad infectocontagiosa, prevenible y curable, que ha reemergido como una amenaza para la salud pública en el mundo, aunque cuenta con un tratamiento efectivo, la no continuidad del tratamiento se convierte en un factor de alto riesgo tanto para los pacientes y para la comunidad en general, convirtiéndose en una problemática de salud pública; ya que no se logra cortar la cadena de transmisión, generando nuevos casos y puede dar lugar a aparición de cepas resistentes a los medicamentos de primera elección. La cura y el control de la enfermedad dependen en forma directa de la adherencia al tratamiento por parte del paciente, el cual tiene una duración aproximada de entre seis a nueve meses según la gravedad de la enfermedad. La pérdida de seguimiento se ha asociado a diferentes determinantes en salud, como la falta de acceso a los servicios de salud, la falta de conocimiento y percepción del riesgo, la existencia de comorbilidades, los frecuentes efectos secundarios de los medicamentos o las barreras socioeconómicas de los pacientes.

Cada vez que un paciente no logra el éxito terapéutico debe reiniciar el tratamiento; elevando los costos en el sistema de salud, sociales y familiares. De igual manera, se incrementa el riesgo de padecer formas resistentes de la enfermedad, las cuales tienen un mayor tiempo de tratamiento, presentan aún más efectos secundarios y las tasas de curación son significativamente, más bajas.

Para dar respuesta a esta problemática, se desarrolló el presente proyecto con el fin de obtener un modelo de *Machine Learning* para predecir el riesgo de pérdida de seguimiento en pacientes pertenecientes al programa de TB en el distrito capital, a partir de determinantes sociales en salud contenidos en el sistema de información. El cual se obtuvo posterior a realizar diferentes análisis estadísticos de los datos, aplicando diferentes técnicas de selección de variables, aplicación de 4 algoritmos diferentes de aprendizaje automático supervisado con diferentes técnicas de remuestreo; para finalmente obtener 3 modelos que fueron validados comparando sus métricas de desempeño (Sensibilidad, Exactitud y área bajo la curva).

Se obtuvo un modelo de aprendizaje supervisado de tipo predictivo (XGBoost con submuestreo) que mostró buen desempeño en la métrica de sensibilidad (70 %), generado a partir de la identificación por métodos estadísticos de 3 variables con 8 atributos relacionadas con el riesgo de pérdida de seguimiento en pacientes del distrito capital; el modelo fue entrenado a partir de bases históricas del sistema de información del programa de TB (2016 a 2022) y validado en una segunda fase con base preliminar de 2023; este modelo será entregado a la entidad con el fin de permitir a los tomadores de decisiones, enfocar sus esfuerzos y recursos en la priorización de los pacientes identificados por el modelo como futuras pérdidas en el seguimiento, actividades desarrolladas por los equipos locales en salud del distrito.

1. DEFINICIÓN DEL PROBLEMA

La tuberculosis (TB) al ser una enfermedad de origen bacteriano, es prevenible y curable. Se estima que entre 2000 y 2021 se salvaron 74 millones de vidas gracias al diagnóstico y el tratamiento. Según la Organización Mundial de la Salud (OMS), esta enfermedad es la decimotercera causa de muerte y la más mortífera por detrás de la COVID-19, entre las enfermedades infecciosas. Cerca de una cuarta parte de la población mundial se encuentra infectada, y entre el 5 y el 10 % de éstos, desarrollarán la enfermedad. Siendo las de mayor riesgo las personas inmunodeprimidas (VIH, cáncer, desnutrición, diabetes) o personas con estilos de vida poco saludables, consumo de tabaco o alcohol [1]. Por tanto, acabar con la epidemia de TB para 2030, corresponde a una de las metas de los Objetivos de Desarrollo Sostenible (ODS).

En este sentido, se aprueba por parte de la OMS en 2014, la estrategia: *Fin a la tuberculosis*, con la cual se busca reducir la mortalidad en un 90 % y la incidencia para 2030 en un 80 %, comparado con 2015. Para lograrlo, se necesitan $13.000 millones US anuales [1]. En Colombia, la TB también es una prioridad en salud pública, por lo cual desde el Ministerio de Salud y Protección Social (MSPS) se cuenta con el Programa Nacional de Prevención y Control de la Tuberculosis (PNPCT), que garantiza acciones para el diagnóstico, tratamiento y seguimiento de casos y sus contactos, y está articulado con el Plan Decenal de Salud Pública (PDSP) [2], el Plan estratégico programático y los lineamientos nacionales.

Con la expedición de la Resolución 227 de 2020, el MSPS actualizó los algoritmos diagnósticos y esquemas de tratamiento para esta enfermedad, el cual consta de cuatro antibióticos suministrados en un solo comprimido (tetraconjugado). La eficacia del tratamiento depende de la adherencia del paciente y tiene una duración de aproximadamente 6 meses.

Una suspensión del tratamiento sin autorización médica por un tiempo mayor a 30 días, hace que el paciente deba reiniciar su esquema, aumentando además del riesgo de sufrir complicaciones de la enfermedad, la aparición de formas resistentes a estos antibióticos [1] (TB farmacorresistente); convirtiéndose una amenaza a nivel de salud pública, ya que el paciente continúa transmitiendo la enfermedad a personas no infectadas [3], incluso de las cepas resistentes.

1.1 PLANTEAMIENTO DEL PROBLEMA

De acuerdo con el informe de evento tuberculosis año 2022 del MSPS, el éxito terapéutico para Colombia desde el año 2009 al 2020 ha oscilado entre 60.9 % para el 2016 y para el año 2014 de 76.5 %, como se observa en la Figura 1; no alcanzando el cumplimiento a los estándares mínimos establecidos por la OMS (tratamiento exitoso debe ser al menos del 90 %), en los últimos 11 años.

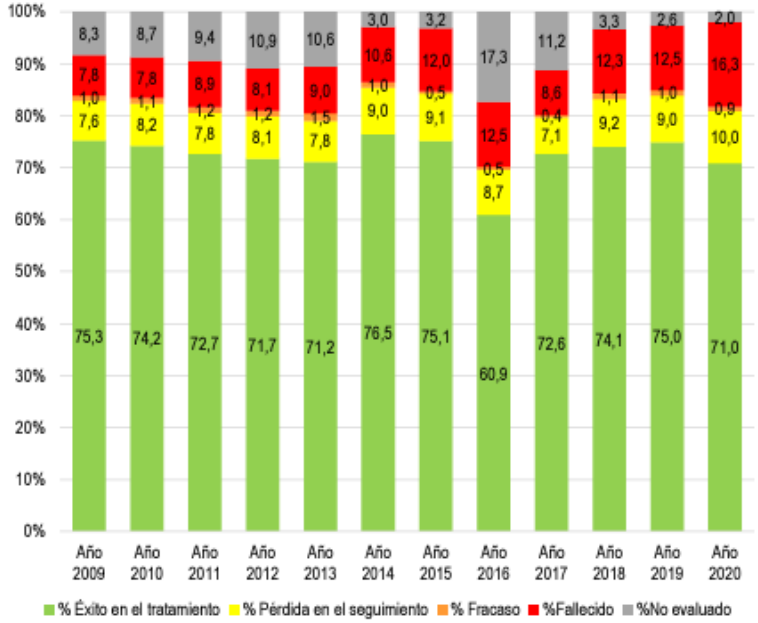

Figura 1. Evaluación de resultados al tratamiento casos de TB pulmonar y extrapulmonar (nuevos y recaídas) en Colombia años 2009 al 2020 *[2]*.

En este mismo informe, se indica además que el índice de pacientes con pérdida en el seguimiento (abandono) para 2020 (año pandémico) en el país es del 10 %, el de fallecidos 16.3 % y los casos resistentes fueron 350; situación que se asemeja al distrito capital según cifras del mismo informe: 68.6 % de éxito en el tratamiento, 6.4 % pérdida en el seguimiento, 23.3 % fallecidos y aportando el 13 % de la totalidad de casos resistentes del país para ese año [4].

La no continuidad en el tratamiento de tuberculosis se ha relacionado en la literatura, con factores socioeconómicos y el estado de salud de las poblaciones; la enfermedad se presenta en personas de bajos recursos con bajo nivel educativo, circunstancias laborales no favorables y falta de alimento [5]. Se han descrito otros factores individuales como: antecedente de tratamiento previo para TB, bajo Índice de Masa Corporal (IMC) asociado a estados de malnutrición [3], hábitos como alcoholismo, tabaquismo, consumo de sustancias psicoactivas, antecedente familiar de tuberculosis, presencia de comorbilidades como VIH, predominancia en el sexo masculino [6], ausencia de reporte de supervisión del tratamiento, pertenecer a un grupo social vulnerable: indígena, habitante de calle o privados de la libertad, ser inmigrante, incluso estar desempleado [7]. Otra causa común de no continuidad del tratamiento, son los síntomas generados como efectos secundarios de los medicamentos [8]. Por otro lado, también se ha reportado que no contar con afiliación al sistema de salud es un factor relacionado en el abandono al tratamiento de TB [9]. Las enfermedades mentales y la escasa red de apoyo familiar, también se han descrito como factores que predisponen la no adherencia al tratamiento [10].

A través del PNPCT se realiza la captura, consolidación y análisis de la información de manera rutinaria y sistemática, con la cual es posible realizar monitoreo y evaluación de indicadores programáticos; mucha de esta información se relaciona con los Determinantes Sociales en Salud (DDS) de los pacientes, la cual se encuentra disponible en el sistema de información del programa distrital de tuberculosis.

A nivel mundial se ha descrito la construcción de diferentes modelos de *Machine Learning* para optimizar el diagnóstico, predecir incidencias, desarrollo de resistencias de ésta y otras enfermedades, también se describen estudios predictivos para identificar pacientes con pérdida de seguimiento. No obstante, en Colombia estos estudios son escasos, si bien se encuentran caracterizados factores asociados a la falla terapéutica de TB en otros territorios; existen limitantes en el sector salud, sobre generación de modelos predictivos apoyados en el uso de herramientas computacionales avanzadas.

En este sentido, el problema se plantea en términos de utilizar metodologías modernas como *Machine Learning*, para identificar de manera temprana (predictiva) a los pacientes que presenten un alto riesgo de pérdida de seguimiento para el distrito basado en sus DSS.

1.2 FORMULACIÓN DEL PROBLEMA

¿Cómo clasificar de manera predictiva y con un buen desempeño, a los pacientes con riesgo de pérdida de seguimiento para el programa de tuberculosis del distrito capital, según sus determinantes sociales en salud a partir del sistema de información del programa?

Preguntas de sistematización:

¿Cuáles son los determinantes sociales en salud relacionados con la pérdida de seguimiento de los pacientes del programa de tuberculosis de Bogotá?

¿Cómo predecir el riesgo de pérdida en el seguimiento de las actuales cohortes del programa de tuberculosis de Bogotá, a partir de registros históricos de pacientes que no tuvieron éxito en el tratamiento?

¿Cómo determinar que el modelo propuesto presenta un buen desempeño en la predicción de pacientes con alto riesgo de pérdida de seguimiento en el programa de tuberculosis en el distrito capital?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Desarrollar un modelo de *Machine Learning* para predecir el riesgo de pérdida de seguimiento en pacientes pertenecientes al programa de tuberculosis en el distrito capital, a partir de determinantes sociales en salud contenidos en la base de tuberculosis sensible, el cual hace parte del sistema de información del programa.

2.2 OBJETIVOS ESPECÍFICOS

- Determinar las variables relacionadas con el riesgo de pérdida de seguimiento de los pacientes del programa de tuberculosis en el distrito capital, a partir del sistema de información.

- Entrenar un modelo de *Machine Learning* para la predicción del riesgo de pérdida en el seguimiento de los pacientes del programa de tuberculosis en el distrito capital.
- Evaluar el rendimiento del modelo de *Machine Learning* utilizando las métricas exhaustividad (Recall), exactitud (Accuracy), precisión y F1-score, para la predicción del riesgo de pérdida en el seguimiento de los pacientes del programa de tuberculosis en el distrito capital.

3. MARCO DE REFERENCIA

A continuación, se presentarán los temas que se relacionan con el desarrollo del proyecto, relacionados con la enfermedad, el aprendizaje automatizado, métodos de selección de variables, técnicas de remuestreo para balanceo de clases y las métricas empleadas en la evaluación de estos modelos.

3.1 Marco teórico tuberculosis

3.1.1. Tuberculosis

La TB es una enfermedad causada por la bacteria *Mycobacterium tuberculosis*, también llamado bacilo tuberculoso o bacilo de Koch. La bacteria se transmite vía aérea cuando una persona enferma habla, tose, ríe, canta o estornuda expulsando al aire pequeñas partículas de secreciones respiratorias que contienen bacilos. Estos bacilos ingresan a las vías respiratorias de personas sanas y se alojan en los pulmones. Por lo general, el bacilo es contrarrestado por la respuesta inmune del huésped (infección latente); no obstante, cuando el sistema inmune se debilita por diferentes factores, la bacteria se multiplica originando una infección activa la cual puede diseminarse a otros órganos. Cerca del 90 % de las personas infectadas con el bacilo nunca llegan a desarrollar la enfermedad, sin embargo, el 5 % generará manifestaciones tempranas en los cinco años posteriores a la exposición, mientras que el otro 5 % puede presentar manifestaciones tardías (varias décadas) [11].

Dentro de los factores que favorecen el desarrollo de la enfermedad se han descrito, el número de bacilos y su virulencia (determinada genéticamente), el hacinamiento que favorece la transmisión, las edades extremas (primeros dos años y posterior a los 65 años), el sexo (más frecuente en hombres adultos), la raza (personas afro presentan 2 veces más riesgo ante la misma intensidad de exposición), la presencia de comorbilidades que debilitan el sistema inmune o factores que favorecen el daño pulmonar (tabaquismo, silicosis, desnutrición, diabetes mellitus, neoplasias, hemodiálisis, trasplantados y VIH) [11].

En 1993, la OMS declara la TB como “emergencia mundial” con el fin de llamar la atención de los países miembros y aunar esfuerzos sobre una enfermedad que es curable y prevenible [11]. Entre 2014 y 2015, los Estados Miembros de la OMS y la ONU se comprometieron a poner fin a la epidemia de TB, a través de la adopción de la estrategia “Poner Fin a la Tuberculosis” y de los Objetivos de Desarrollo Sostenible (ODS) [12]. Con esta estrategia se busca que ninguna persona con TB y su familia, tengan que hacer frente a gastos catastróficos para tratar esta enfermedad (gastos superiores al 20 % de la renta doméstica). Sin embargo, cerca de la mitad de los pacientes y sus familias actualmente, se siguen viendo afectadas económicamente por ese motivo [1]. Entre 2017 y 2019 se intensificaron esfuerzos para lograr el compromiso político, y se reafirmaron metas mundiales para la movilización de fondos para prevenir y atender la enfermedad. No obstante, con la pandemia COVID-19, se generaron retrocesos debido a la no disponibilidad

de los servicios de salud para los pacientes. La consecuencia más inmediata fue la gran caída en el número de personas recién diagnosticadas con TB y un aumento en el número de fallecidos por TB en 2020 [12].

La OMS estima que alrededor de un cuarto de la población mundial está infectada por el bacilo [12]. La mayor parte de los casos se presentan en países pobres o de ingreso medio, donde los menores de 50 años (población económicamente activa) son los principales afectados [11]. Para el año 2020, en todo el mundo 9.9 millones de personas enfermaron de TB de los cuales más de 1 millón fueron niños. Geográficamente, Asia aportó 43 % de los casos, África 25 %, el Pacifico Occidental 18 %, Mediterráneo Oriental 8.3 %, América 3 % y Europa 2.3 %. Los 30 países con alta carga de TB representaron el 86 % de casos incidentes en todo el mundo, siendo los mayores aportantes: India (26 %), China (8,5 %), Indonesia (8,4 %), Filipinas (6 %), Pakistán (5,8 %), Nigeria (4,6 %), Bangladesh (3,6 %) y Sudáfrica (3,3 %), países que representaron dos tercios del total mundial de casos. A nivel de las Américas, para 2019 el 88.1% del total de casos diagnosticados se concentraron en 12 países, siendo los mayores aportantes en su orden: Brasil (33,1 %), Perú (13,4 %), México (10,3 %) y Colombia (6.6 %) [12].

Para 2021, en nuestro país se notificaron al sistema de vigilancia en salud pública (SIVIGILA) 14.060 casos de tuberculosis de todas las formas. Las tasas de incidencia más altas se presentaron en orden descendente en: Amazonas, Risaralda, Meta, Barranquilla, Cali, Guaviare y Arauca. En cuanto al contexto sociodemográfico el 66.2 % de los casos ocurrieron en hombres, el grupo de edad más afectado fue el de 25 a 34 años (23.6 %), seguido de mayores de 65 años (18.5 %). La pertenencia étnica más afectada corresponde a afrocolombianos 3.6 % seguido de indígenas 3.5 %. A nivel clínico, la forma pulmonar aporta el mayor porcentaje de los casos (84,3 %). Los grupos poblacionales afectados más frecuentes son: Personas Privadas de la libertad (PPL), indígenas, habitantes de calle, trabajadores de la salud y migrantes. Las comorbilidades más frecuentes asociadas con TB fueron: desnutrición (15,2 %), seguida de la coinfección TB-VIH (12,1 %) y diabetes (9,7 %) [12].

El distrito capital no es ajeno a esta situación, acorde a la información de SaluData en los últimos 10 años la incidencia de TB ha oscilado entre 11,5 y 16,0 casos por cada 100.000 habitantes. Siendo la tasa de incidencia (casos nuevos y recaídas) para 2022 la más alta de los últimos cinco años (15.7 por cada 100 mil habitantes). Al analizar comparativamente, con el año inmediatamente anterior, se observa un incremento del 13,3 % en la detección de casos, atribuido a la implementación de los algoritmos diagnósticos de la resolución 227 del 2020 del Ministerio de la Protección Social, donde se incluyó el cultivo líquido y la prueba molecular para el diagnóstico de TB. La tasa de mortalidad para 2022 fue de 1 por cada 100 mil habitantes, manteniéndose estable en los últimos diez años. Las localidades con mayores tasas de incidencia para 2022 fueron: Los Mártires, Antonio Nariño, Santafé, La Candelaria, Sumapaz y Rafael Uribe Uribe. Con relación a variables sociodemográficas se evidencia que la enfermedad para Bogotá predomina en hombres 65.9 %, los grupos de edad más afectados son mayores de 65 años (32 %) seguido del grupo de 25 a 39 años (25.5 %). La forma pulmonar aportó el 75 % de los casos y las comorbilidades más

frecuentes fueron: VIH 19 %, desnutrición 14 %, EPOC y diabetes cada una 11 %. Los grupos poblacionales más afectados fueron: migrantes 8.7 %, habitantes de calle 6.3 %, PPL 3.7 % y trabajadores de la salud 2.1 % [13].

3.1.2. Tratamiento

Dentro de los factores que han impedido el control de la enfermedad se encuentran: falta de financiación, apoyo gubernamental o mala planificación de los programas de control; deterioro de las condiciones de vida de la población (migraciones, guerras y hambrunas), subdiagnóstico y aparición de formas resistentes a los fármacos. Por lo anterior, los objetivos generales de control de la TB apuntan a 3 objetivos: disminuir morbimortalidad, transmisión de la enfermedad; y prevenir el desarrollo de resistencias. En este marco, la OMS estableció en 1999 la estrategia DOTS (Tratamientos Directamente Observados), es un método que asegura la adhesión al tratamiento por parte del paciente, y requiere que un trabajador de la salud u otra persona designada para ello, presencie la ingesta de la medicación por parte del paciente [11] hasta la finalización del tratamiento; estrategia que se encuentra implementada actualmente en nuestro país, en el marco del plan estratégico programático.

Para lograr la curación, los medicamentos deben tomarse durante 6 meses en promedio, la suspensión del tratamiento antes de tiempo, sin supervisión médica o una mala formulación puede provocar que los bacilos tuberculosos vivos presentes en el huésped desarrollen mecanismos de resistencia a estos medicamentos (tuberculosis farmacorresistente), requiriendo un tratamiento farmacológico más agresivo (segunda línea) y prolongado. En algunos casos incluso, se puede desarrollar resistencias a medicamentos de segunda línea, dejando pocas opciones de tratamiento disponibles para el paciente (formas extremadamente resistentes). Por lo anterior, la tuberculosis multirresistente se considera una amenaza para la salud pública, según la OMS solo dos de cada cinco personas con tuberculosis farmacorresistente (TB FR) tuvieron acceso al tratamiento en 2022 [1].

En Colombia, para el año 2021 se reportaron desde el PNPCT del MSPS un total de 14.091 casos de TB; según la estimación de la OMS el país debió detectar para ese año cerca de 21.000 casos, representando una detección de solo el 67 % del total de casos estimados. Con corte al tercer trimestre de 2022, se reportó de manera preliminar un total acumulado de 13.487 casos de TB todas las formas, con un incremento del 31 % comparativo con el mismo periodo del año 2021, debido al aumento del diagnóstico molecular, captación de sintomáticos respiratorios, acciones de búsqueda activa de casos y contactos a nivel institucional y comunitario (acciones programáticas). Para 2021, el 64.2 % de la carga nacional de casos de TB del país se concentró en 8 departamentos y distritos, siendo el distrito capital el tercero aportando el 8.2 % del total de casos. En cuanto al indicador programático de éxito terapéutico entre casos nuevos y recaídas en el 2020, se reportó un 71 % de casos curados y terminados, 51 % en personas con coinfección TB y VIH, 53.7 % en personas previamente tratadas y 60 % en los casos TB FR (indicador procedente del año 2019). Comparado con 2019, se evidencia una disminución del éxito terapéutico del 5.3 %, dado porque persiste una alta proporción de fallecidos con el 16.3 % (n=1.891), la

cual se incrementó frente a un 12.3 % del año anterior; también se reportaron 10 % (n=1.168) pérdidas en el seguimiento al tratamiento, 2 % (n=237) casos sin evaluar y un 0.9 % (n=110) de fracasos [4].

El informe destaca los esfuerzos que ha realizado el país en el abastecimiento de medicamentos para tratar el 100 % de los casos, así como adquisición de nuevos fármacos (2021-2022) para tratar TB FR y profilaxis acortadas en personas con VIH, con el fin de fortalecer la adherencia terapéutica por parte de los pacientes. Adicionalmente, enfatiza que se requiere fortalecer un sistema de información interoperable que facilite las salidas de tableros de control automatizados para consolidar y optimizar información que sirva para la toma de decisiones y modelos en TB y VIH, TB infantil, entre otros. Así como incrementar el abordaje intersectorial de personas en contexto de vulnerabilidad y efectuar su monitoreo a partir del sistema de información [4].

3.1.3. Pérdida de seguimiento

De acuerdo con la Resolución 227 de 2020, la pérdida de seguimiento en personas afectadas por tuberculosis sensible se define como el no inició de tratamiento o la interrumpió durante un mes o más sin autorización médica [14]. En la literatura se han descrito como principales determinantes sociales relacionados con la pérdida de seguimiento los factores socioeconómicos [3], que afectan de forma directa el estado de salud tales como: bajos recursos económicos, falta de acceso a una adecuada alimentación [5], barreras de acceso al sistema de salud [9], presencia de comorbilidades [6], o pertenencia a grupos sociales en condiciones de vulnerabilidad [7]. Los pacientes con pérdida de seguimiento se consideran un riesgo a nivel de salud pública ya que perpetúan la transmisión de la enfermedad en la comunidad, generan riesgo de aparición de cepas resistentes a los medicamentos y generan una mayor carga de morbimortalidad. Todo esto constituye una barrera para el cumplimiento del ODS de controlar la TB para 2030 [15] y poner fin a la epidemia.

En el distrito capital el seguimiento de los pacientes en tratamiento para tuberculosis se encuentra a cargo de los equipos locales de las Subredes Integradas de Servicios de Salud (SISS), quienes realizan el respectivo seguimiento de cada paciente acorde con la localidad de residencia de éste; dentro de las acciones descritas se encuentra realización de visitas domiciliarias, seguimiento a los contactos, apoyo en la gestión de ayudas sociales en caso de requerirlas, disminución de barreras de acceso en salud en articulación con las IPS y EAPB, así como oferta de apoyo psicosocial en caso de requerirlo. No obstante, pese a estas acciones en muchas ocasiones no es posible asegurar la adherencia terapéutica por parte de los pacientes.

3.2 Marco teórico Aprendizaje Automático

3.2.1 Aprendizaje automático

Acorde con IBM el *Machine Learning* (ML) o aprendizaje automatizado se define como una rama de la inteligencia artificial (IA) que se centra en el uso de datos y algoritmos con el fin de imitar por parte de computadoras la forma en que los humanos aprenden [16], esto sucede al alimentarlas con datos e información en forma de observaciones de interacciones de la vida real [17] lo que mejora de manera gradual su precisión de manera autónoma [16]. El ML permite analizar de manera efectiva grandes volúmenes de datos, utilizando métodos estadísticos para entrenar algoritmos que posteriormente permiten clasificar o predecir, así como descubrir patrones de comportamientos, que posteriormente son usados en la toma de decisiones dentro de aplicaciones o empresas [16]; requiere de tiempo y expertos para su configuración [17]. Se divide en tres tipos: supervisado, no supervisado y semisupervisado.

3.2.1.1. Aprendizaje Supervisado

En el aprendizaje automático supervisado, se crea un modelo que realiza predicciones basadas en pruebas en presencia de incertidumbre. El algoritmo se entrena con un conjunto de datos (entrada) y respuesta (salida) conocidos y entrena un modelo para generar predicciones razonables como respuesta a datos nuevos. Se utiliza cuando se dispone de datos conocidos para la salida que se desea predecir. Este tipo de aprendizaje emplea técnicas de clasificación y regresión para el desarrollo de los modelos [18].

Técnicas de clasificación: estos modelos clasifican los datos de entrada en categorías y predicen respuestas de variables discretas. Esta técnica se utiliza cuando los datos se pueden etiquetar, categorizar o dividir en clases o grupos específicos [18]. Luego de ser expuesto a suficientes ejemplos, el sistema de aprendizaje supervisado comenzará a reconocer las imágenes y podrá distinguirlas de manera automática [17].

Técnicas de regresión: predicen respuestas de variables continuas, se utiliza cuando se trabaja con un intervalo de datos o la respuesta es un número real. Se aplica en la sensorización virtual, predicción de carga eléctrica y trading algorítmico [18].

3.2.1.2. Aprendizaje No Supervisado

Este tipo de aprendizaje identifica patrones ocultos o estructuras intrínsecas en los datos, sin necesidad de la intervención humana, permitiendo descubrir similitudes y diferencias en los datos [16]. Se utiliza para determinar conclusiones sobre conjunto de datos de entrada sin respuestas etiquetadas. También se utiliza para reducir la cantidad de funciones en un modelo a través del proceso de reducción de dimensionalidad [16]. La técnica más común es la agrupación en clústeres, que permite realizar análisis exploratorios para identificar patrones ocultos en los datos, como en las ciencias ómicas, segmentación de mercado y reconocimiento de imágenes y patrones [18].

3.2.1.3. Aprendizaje Semisupervisado

El aprendizaje semisupervisado brinda un punto intermedio entre el aprendizaje supervisado y el no supervisado. En este tipo de aprendizaje, durante el entrenamiento se utiliza un conjunto de datos etiquetado más pequeño para guiar la clasificación y la extracción de características de un conjunto de datos más grande y sin etiquetar. A través de este, se puede resolver el problema de no tener suficientes datos etiquetados para aplicar un algoritmo de aprendizaje supervisado o cuando el costo de etiquetarlos es muy alto [16].

3.2.1.4. Algoritmos más comunes en ML

Redes neuronales: las redes neuronales simulan la forma en que funciona el cerebro humano, con una gran cantidad de nodos de procesamiento vinculados. Las redes neuronales son buenas para reconocer patrones y juegan un papel importante en las aplicaciones, como la traducción de lenguaje natural, el reconocimiento de imágenes, el reconocimiento de voz y la creación de imágenes [16]

Regresión lineal: este algoritmo se utiliza para predecir valores numéricos, basándose en una relación lineal entre diferentes valores. Por ejemplo, la técnica podría usarse para prever los precios de la vivienda en función de los datos históricos del área [16].

Regresión logística: hace predicciones para variables de respuesta categórica (como las respuestas "sí/no" a las preguntas). Se puede utilizar para aplicaciones como la clasificación de spam y el control de calidad de una línea de producción [16].

Agrupación en clústeres: pueden identificar patrones en los datos para que puedan ser agrupados. También pueden utilizarse para identificar las diferencias entre los elementos de datos que los humanos han pasado por alto [16].

Árboles de decisión: los árboles de decisión se pueden usar tanto para predecir valores numéricos (regresión) como para clasificar datos en categorías. Los árboles de decisión utilizan una secuencia de ramificación de decisiones vinculadas que se pueden representar con un diagrama de árbol. Una de las ventajas de los árboles de decisión es que son fáciles de validar y auditar, a diferencia de la caja negra de la red neuronal [16].

Bosques aleatorios: en un bosque aleatorio, el algoritmo predice un valor o categoría combinando los resultados a partir de una serie de árboles de decisión [16]. Una variante de este algoritmo es Random Forest balanceado (BRF) el cual presenta restricciones con las clases desequilibradas porque utiliza una muestra de arranque del conjunto de entrenamiento para formar cada árbol. Para superar esta limitación, se debe realizar previamente un balanceo de clases, ya sea mediante un muestreo descendente o un muestreo excesivo. El algoritmo BRF lo hace extrayendo iterativamente una muestra de arranque con proporciones iguales de puntos de datos tanto de la clase minoritaria como de la mayoritaria [19].

***Naive Bayes*:** Es un algoritmo de aprendizaje automático generativo de tipo supervisado que utiliza principios de probabilidad, supone que los predictores son condicionalmente independientes, es decir, que no están relacionados con ninguna de las otras características del modelo y también supone que todas las características contribuyen por igual al resultado. Se utiliza para tareas de clasificación, como la clasificación de texto. A diferencia de los clasificadores discriminativos, como la regresión logística, no aprende qué características son más importantes para diferenciar entre clases [20].

XGBOOST (*Extreme Gradient Boosting*): este método permite generar un modelo de clasificación o regresión sobre una muestra descrita por variables cualitativas y/o cuantitativas. El método maneja eficazmente grandes conjuntos de datos con un gran número de variables. Puede usarse en tareas de **clasificación** ya que permite predecir la clase a la que pertenece cada observación, basándose en variables explicativas que pueden ser cuantitativas y/o cualitativas. Un ensemble es una combinación de modelos individuales simples que al interactuar de manera conjunta generan un modelo más potente. El boosting de aprendizaje automático es un método que crea este tipo de conjuntos. Comienza ajustando un modelo inicial (en nuestro caso un árbol de regresión o clasificación) a los datos. A continuación, se construye un segundo modelo que se centra en predecir con exactitud las observaciones que el primer modelo predijo mal. Se espera que la combinación de estos dos modelos sea mejor que cada uno de ellos. Este proceso de refuerzo se repite varias veces, y cada modelo sucesivo intenta corregir las deficiencias del conjunto refuerzo combinado que contiene todos los modelos anteriores. El refuerzo de gradiente se basa en la intuición de que el mejor modelo siguiente posible, cuando se combina con los modelos anteriores, minimiza el error de predicción global. La idea clave es establecer el peso de cada observación para este próximo modelo con el fin de minimizar el error. En cada paso de boosting y para cada observación, se calcula una puntuación basada en el error de predicción del modelo. El nombre de boosting de gradiente surge del hecho de que cada peso se establece en función del gradiente del error con respecto a la predicción. Cada nuevo modelo da un paso en la dirección que minimiza el error de predicción, en el espacio de predicciones posibles para cada observación [21].

3.2.2 Selección de variables

Una selección de variables adecuada tiene grandes ventajas: reducción del sobreajuste (*overfitting*), mejora de la precisión de las predicciones, favorece la eliminación de atributos redundantes (multicolinealidad), simplificación de los modelos y reducción del tiempo de proceso. El objetivo que persiguen las técnicas de selección de variables es obtener la lista de atributos más relevantes de un conjunto de datos, es decir seleccionar aquellas variables que contienen la mayor información con respecto a la variable objetivo, de resultado o dependiente. Si dicho proceso se realiza de forma exitosa, un subconjunto de atributos aportará la misma información que el conjunto original sin tener en cuenta los atributos irrelevantes o redundantes que pudieran existir en los datos originales [22].

Existen diferentes métodos descritos para ello a continuación, se describen los utilizados en la selección de variables categóricas del proyecto.

3.2.2.1 LASSO

Este modelo de regularización determina la magnitud de los coeficientes del modelo, para ello deben estar escalados. La regularización Lasso penaliza la suma del valor absolutos de los coeficientes de regresión y tiene el efecto de forzar a que los coeficientes de los predictores tiendan a cero [23]. Por lo que, consigue excluir los predictores menos relevantes. Cuando las variables conjunto de datos corresponden a variables categóricas, aplicar Lasso no es apropiado como técnica de selección de características debido a que trata a cada atributo de forma independiente, ignorando cualquier relación entre ellas (variables dummy) es decir, que se ignora el hecho de que las variables ficticias generadas, tomadas en su conjunto, representan la misma variable categórica.

Para ello, se requiere una penalización que incorpore la “estructura de grupo” subyacente entre las variables ficticias cuando se trabaja con variables categóricas. Existe entonces una variante de este modelo llamado Group LASSO, en el cual se modifica ligeramente su penalización para que pueda manejar grupos de variables, reorganizando los coeficientes de regresión que pertenecen al mismo grupo en vectores. De esta manera, permite realizar la selección de características, de una mezcla de variables continuas/categóricas.

3.2.2.2 Regresión Logística

El objetivo de este modelo es predecir la probabilidad de que una observación determinada pertenezca a una clase particular y permite identificar las características más relevantes de un conjunto de datos. Para ello se pueden incorporar una a una las variables ya sea con selección hacia adelante o *Fordward*, la cual consiste en que el modelo comienza sin predictores y, sucesivamente, ingresa predictores significativos hasta alcanzar un criterio estadístico de detención. En este enfoque, se agregan variables al modelo una a la vez. En cada paso, se prueba cada variable que aún no está en el modelo para su inclusión en el modelo. La más significativa de estas variables se agrega al modelo, siempre que su valor p esté por debajo de un nivel preestablecido. De igual manera, la incorporación de variables se puede hacer hacia atrás o *Backward*; en este enfoque, se comienza ajustando un modelo con todas las variables de interés; posteriormente, se descarta la variable menos significativa, siempre que no sea significativa en el nivel crítico elegido [24] y se itera esta operación hasta tener la totalidad de variables con p valor significativo para el modelo.

A su vez, la selección por pasos (*Stepwise*) es un método que permite movimientos en cualquier dirección, eliminando o agregando variables en los distintos pasos. El proceso consiste en alternar la selección de variables, iniciando por las menos significativas para eliminar y luego volver a considerar todas las variables eliminadas (excepto la eliminada más recientemente) para reintroducirlas en el modelo. Esto significa que se deben elegir dos niveles de significación separados para eliminar del modelo y para agregar al modelo. La significancia de la adición debe ser más estricta (valor p menor) que la significancia de la eliminación (valor p mayor). El método de regresión por pasos combina estos dos enfoques, agregando y eliminando predictores a medida que construye el modelo [24].

3.2.2.3 Eliminación Recursiva de Características (RFE)

Es un método iterativo de selección de características que funciona eliminando recursivamente características del conjunto de datos y evaluando el rendimiento de un modelo de aprendizaje automático en cada paso. Comienza con todas las características y las clasifica en función de su importancia o relevancia para la variable de destino, y luego elimina las características menos importantes y repite el proceso hasta que se alcanza el rendimiento deseado del modelo [25]

3.2.2.4 Estadísticos para comparación de modelos

El mejor modelo por elegir corresponde a aquel que proporciona los mejores valores predichos por lo que para comparar los resultados obtenidos entre los diferentes modelos generados, se debe utilizar un estadístico. La elección de modelos candidatos de crecimiento individual no deben basarse en criterios como valores de coeficientes de determinación (R2) o coeficientes de variación (CV), y en su lugar se sugieren criterios robustos como la teoría de la información [22]. Dentro de estos se encuentran:

El criterio de información de Akaike (AIC) proporciona un método simple y objetivo que selecciona el modelo más adecuado para caracterizar los datos experimentales. Se define como: $AIC = -2 \log(L(\theta b)) + 2K$ (1) 22, donde $\log(L(\theta b))$ es el logaritmo de la máxima verosimilitud, que permite determinar los valores de los parámetros libres de un modelo estadístico y K es el número de parámetros libres del modelo. Esta expresión proporciona una estimación de la distancia entre el modelo y el mecanismo que realmente genera los datos observados, que es desconocido y en algunos casos imposible de caracterizar. Como la estimación se hace en función de los datos experimentales, esta distancia es siempre relativa y dependiente del conjunto de datos experimentales. Por tanto, un valor individual de AIC no es interpretable por sí solo, y los valores AIC sólo tienen sentido cuando se realizan comparaciones utilizando los mismos datos experimentales [22].

Otro criterio de selección de modelos se basa en el denominado Criterio de Información Bayesiano (BIC). La expresión matemática que estima el BIC se define como: $BIC = (-\ln L \times 2) + [\theta \times \ln (n)]$ Donde θ es el número de parámetros estimados y n es el número de observaciones. Ya con el valor estimado de BIC, el modelo candidato que resulta ganador se define con la menor estimación de BIC [22].

3.2.3 Balanceo de clases

El desbalanceo de clases se presenta cuando en un conjunto de datos existe desequilibrio entre grupos en la variable objetivo uno dominante y otro minoritario. Estadísticamente el grupo con mayor número de observaciones acumula el resultado final de la clasificación opacando la clase minoritaria [26]. Para superar esta dificultad se ha empleado diferentes técnicas de remuestreo o *resampling,* las cuales se describen a continuación:

Sobremuestreo (*Oversampling*): Genera nuevas muestras para la clase que está

subrepresentada.

Submuestreo (*Undersampling*): Elimina muestras de la clase que están sobrerrepresentadas.

En la mayoría de los casos, se prefiere el sobremuestreo en lugar del submuestreo, ya que no es ideal eliminar información importante para el modelado. De igual manera, ambas técnicas pueden introducir sesgos, dado que donde se toman más muestras de una clase que de la otra para neutralizar el efecto del desequilibrio presente en los datos, por lo cual se han desarrollado otro tipo de técnicas [27].

Técnica de sobremuestreo sintético de minorías (SMOTE): genera nuevas ilustraciones a través de la interpolación con los k-vecinos más cercanos de la clase minoritaria [27].

Enlaces Tomek (Tomek links): se utiliza para reducir el desequilibrio en conjuntos de datos desequilibrados eliminando instancias de la clase mayoritaria que están cerca de instancias de la clase minoritaria. Las ventajas de utilizarla incluyen: simplicidad, eficacia reducción de ruido y mejora la precisión de la clasificación, y la capacidad de identificar instancias importantes de la clase minoritaria. Una desventaja de utilizar esta técnica consiste en que eliminar demasiadas instancias de la clase mayoritaria, puede provocar un ajuste insuficiente y una disminución del rendimiento general de la clasificación [28].

SMOTE-Tomek Links: este método combina la capacidad de SMOTE para generar datos sintéticos para la clase minoritaria y la capacidad de Tomek Links para eliminar los datos que se identifican como enlaces Tomek de la clase mayoritaria (es decir, muestras de datos de la clase mayoritaria que son más cercanas a los datos de la clase minoritaria). Este método es eficaz porque los datos sintéticos que se generan son relativamente cercanos al espacio de características de la clase minoritaria, lo que agrega nueva “información” a los datos, a diferencia del método de sobremuestreo original. [29].

3.2.4 Medidas de evaluación del modelo

Posterior a la construcción de un modelo y elegir los parámetros, es de gran importancia evaluar su desempeño con el fin de escoger el más adecuado, esto depende de la forma en que se está llevando a cabo la clasificación de los datos. La matriz de confusión se basa en el teorema de Bayes (probabilidades condicionales), es útil para evaluar el rendimiento de un modelo de clasificación binaria, ya que presenta la relación entre las predicciones del modelo y las clases verdaderas [30]: Verdaderos Positivos (TP) y Verdaderos Negativos (TN), así como Falsos Positivos (FP) y Falsos Negativos (FN). A partir de esta matriz se pueden calcular las siguientes métricas:

Exactitud o *Accuracy*: Es una métrica que mide la proporción de predicciones correctas del modelo (TP+TN) sobre el total de predicciones (TP+TN+FP+FN). Tiene buen rendimiento cuando las clases están equilibradas en el conjunto de datos [30].

Precisión: Es el número de clases identificados correctamente como positivos (TP) de un total de datos identificados como positivos en los datos de desempeño (TP +FP), es decir que solo da cuenta de los casos clasificados como positivos [31].

Sensibilidad, Exhaustividad o *Recall*: Conocida como tasa de verdaderos positivos. Calcula la capacidad del modelo para identificar correctamente las instancias positivas de la clase objetivo, identificando instancias positivas correctamente clasificadas por el modelo. Se define como la proporción de verdaderos positivos (TP) sobre la suma de verdaderos positivos y falsos negativos (TP+FN) [30].

Especificidad: También llamada tasa de verdaderos negativos. Se refiere a la capacidad del modelo para identificar correctamente las instancias negativas de la clase objetivo, identificando instancias negativas que fueron correctamente clasificadas por el modelo. Se define como la proporción de verdaderos negativos (TN) sobre la suma de verdaderos negativos y falsos positivos (TN+FP) [30].

Puntaje F1: Es la medida armónica entre la precisión del modelo y su exhaustividad (*Recal*l), y la relación se basa que a mayor porcentaje de F1 el modelo es mejor [31].

Curva ROC y AUC: La curva ROC es una representación gráfica de la sensibilidad, contra el inverso de la especificidad (1- especificidad) para un clasificador binario y de acuerdo con cierto umbral. Esto se mide con el área bajo la curva (AUC) que muestra el funcionamiento del modelo al diferenciar y predecir entre las dos clases. Si el resultado del AUC es 1 demuestra que el modelo distingue perfectamente todas las muestras o características de cada una de las clases asignándoles la clase correcta, denota que el modelo está sobreentrenado y solo es útil para los datos que se usaron para su entrenamiento, pero no para nuevos datos; si el AUC es de 0,5 o menos, el modelo es incapaz de diferenciar entre ambas clases. Por lo anterior, un valor entre 0,8 y 0,9 es el ideal [31].

3.3 Antecedentes

A continuación, se relacionan algunos trabajos en los que se ha hecho uso de herramientas de ciencia de datos como aporte en el tratamiento y pronóstico de la TB. Se hace especial énfasis en aquellos estudios enfocados en detectar riesgo de falla terapéutica utilizando metodologías de *Machine Learning* (ML), los cuales se han realizado principalmente en países con alta carga de la enfermedad.

En el trabajo realizado por Naidu A y su equipo en 2023, mencionan que para lograr el objetivo de la estrategia de la OMS “Acabar con la tuberculosis” para 2035, se requiere de un enfoque multisectorial que se vería beneficiado por los últimos avances computacionales. Presenta un resumen de estudios en los que han utilizado herramientas y algoritmos computacionales avanzados para el diagnóstico precoz de la tuberculosis,

descubrimiento de fármacos antimicobacterianos, el diseño y selección de fármacos y en el diseño de la próxima generación de vacunas contra la TB. Menciona el uso de algoritmos de aprendizaje automático supervisado, no supervisado, el análisis de componentes principales y las redes neuronales, para comprender los patrones subyacentes asociados a la presencia de componentes genéticos específicos del bacilo. Propone el uso de enfoques de aprendizaje automático en el proceso de descubrimiento de fármacos, ya que su aplicación implica desarrollo de modelos de clasificación basados en propiedades antibacterianas y la extracción de características que definen estas propiedades [32] [33] [34]. En este estudio se resalta las bondades del uso de *Machine Learning* en la generación de clasificadores para el desarrollo de nuevos fármacos que permitan el control de la enfermedad brindando evidencia al presente proyecto en cuanto a la utilidad de estas herramientas.

Por su parte Peetluk y colaboradores en 2021, realizaron una revisión sistemática de modelos de predicción de resultados de tratamiento en adultos con TB pulmonar, a partir de publicaciones de estudios en bases de datos entre el 01-01-1995 y el 09-01-2020. La calidad de los estudios se evaluó mediante la herramienta de evaluación del riesgo de sesgo del modelo de predicción. Identificaron 14 739 artículos relacionados, de los cuales revisaron 536 textos completos y se incluyeron dentro de la revisión sistemática 33 estudios que presentaban 37 modelos de predicción. Los resultados del tratamiento incluyeron muerte (n=16, 43 %), fracaso del tratamiento (n=6, 16 %), incumplimiento (n=6, 16 %) o un resultado compuesto (n=9, 25 %). La mayoría de los modelos (n=30, 81%) midieron la discriminación (mediana del estadístico c=0,75; IQR: 0,68-0,84), y 17 (46 %) informaron de la calibración, a menudo aplicando la prueba de Hosmer-Lemeshow (n=13). Los factores predictivos más frecuentes fueron la edad, el sexo, la TB extrapulmonar, el índice de masa corporal, los resultados de la radiografía de tórax, la TB previa y el VIH. El riesgo de sesgo varió de unos estudios a otros, pero todos presentaron un alto riesgo de sesgo en su análisis; concluyendo que los modelos de predicción de resultados de la tuberculosis son heterogéneos con definiciones de resultados, predictores y metodología dispares [35] [36]. Los trabajos de revisiones sistemáticas permiten sintetizar factores predictivos a partir de los resultados obtenidos del análisis de varios modelos, en este caso 37 que pueden analizarse acorde con los resultados obtenidos en el presente proyecto.

En la publicación de Ma y colaboradores en 2023, se desarrolló y validó un modelo clínico sencillo para predecir resultados de fallos terapéuticos en pacientes con tuberculosis pulmonar multirresistente; a través de un estudio de cohortes retrospectivo entre enero de 2017 y diciembre de 2019 en un hospital en China. Se utilizaron la regresión del operador de selección y contracción mínima absoluta (LASSO) y la regresión logística multivariante, para seleccionar los factores pronósticos de los resultados del tratamiento no exitoso. Se construyó un nomograma basado en cuatro factores pronósticos. Para evaluar el modelo se utilizó la validación interna y la validación cruzada *leave-one-out.* Se identificaron como factores pronósticos la falta de educación en salud, la edad avanzada, el sexo masculino y mayor extensión de la afectación pulmonar. El área bajo la curva del modelo fue de 0,757 (IC del 95 %: 0,711 a 0,804) y el índice de concordancia (índice C) fue de 0,75. La pendiente de la curva de calibración fue de 0,968, esto indica que el modelo era exacto en la mayoría

de los casos y preciso para predecir los resultados del tratamiento sin éxito [37]. Este estudio fue relevante para el proyecto, ya que proporcionó metodologías utilizadas en el proyecto para la selección de las variables.

En 2017 Perfura-Yone y su equipo, desarrollaron y validaron una sencilla puntuación para predecir la mortalidad durante el tratamiento de la tuberculosis en zonas de alta endemicidad de Camerún entre 2012 y 2013. Las características basales asociadas con la mortalidad se investigaron mediante regresiones logísticas. Se construyó una puntuación de pronóstico simple (CABI) con coeficientes de regresión para los predictores en el modelo final. La validación interna utilizó procedimientos de remuestreo bootstrap. La discriminación de los modelos se evaluó mediante el estadístico c y la calibración mediante gráficos de calibración y el estadístico de Hosmer y Lemeshwow (H-L). La puntuación óptima se basó en el índice de Youden. La forma clínica de la tuberculosis (C), la edad (A, años), el índice de masa corporal ajustado (B) y el estado de infección por el VIH (I) fueron predictores significativos en el modelo final ($p < 0,0001$) [38], apoyando los predictores descritos en el trabajo de Moreno *et al* [39]. El umbral de riesgo absoluto óptimo fue del 4,8 %, lo que corresponde a una sensibilidad del 81 % y una especificidad del 67 % [38]. Este trabajo aporta al proyecto las técnicas de remuestreo empleadas, el uso de coeficientes de regresión logística para selección de los predictores y resultados de métricas obtenidas que son discutidas acorde con resultados obtenidos.

En el estudio realizado por Sauer y otros en 2018, se identificaron las características asociadas al fracaso terapéutico, logrando discriminar pacientes que predictivamente presentan mayor riesgo de fracaso terapéutico, utilizando un conjunto de datos de varios países a los que se aplicaron diferentes técnicas de ML para identificar factores estadísticamente significativos a partir de variables demográficas y clínicas: antibiograma, hallazgos imagenológicos, resultados de baciloscopia, nivel educativo y situación laboral, tal como se describe en otro trabajo de Chen y colaboradores [40]. El modelo más predictivo fue la selección escalonada hacia delante (AUC: 0,74), aunque la mayoría de los modelos funcionaron con un AUC igual o superior a 0,7. El aprendizaje automático ayuda a identificar a los pacientes con mayor riesgo de fracaso terapéutico. Un seguimiento más estrecho de estos pacientes puede disminuir las tasas de fracaso del tratamiento y prevenir la aparición de resistencias a antibióticos [6]. Este trabajo resalta el uso de variables demográficas y clínicas (insumos bajo costo) como útiles predictores de la falla terapéutica utilizando modelos de ML, enfoque llamativo para ser implementado en países de renta baja y media.

Por su parte en el trabajo de Hussain y Junejo de 2019, se buscó predecir el resultado del tratamiento de un paciente concreto al inicio de éste, de modo que el personal de salud pueda utilizar esta información de forma selectiva y rentable. Se realizó con información de 4 213 pacientes de Pakistán (país alta carga de la enfermedad). Se modeló como un problema de clasificación, y el resultado del tratamiento se predijo utilizando 3 algoritmos de aprendizaje automático (Neural Networks, Random Forest y Support Vector Machines). Los resultados se evaluaron utilizando cuatro medidas de rendimiento: exactitud, precisión, sensibilidad y especificidad. Los modelos ofrecieron una mejora de más del 12 % con respecto a la predicción de referencia [41] [42]. Este estudio es muy importante ya que, en

él se utiliza uno de los algoritmos empleados en el presente proyecto permitiendo realizar comparaciones.

En 2021 Peetluk y colaboradores, generaron un modelo de predicción del fracaso terapéutico de la tuberculosis a partir de datos clínicos de referencia en Brasil, evaluando el valor incremental con el VIH y el uso de isoniazida en la gravedad de la enfermedad. El estudio incluyó 944 participantes con TB pulmonar confirmada por cultivo, susceptible a fármacos, quienes comenzaron la terapia antituberculosa de primera línea y estuvieron ≥12 meses de seguimiento. El criterio de valoración de fracaso terapéutico TB fue: muerte, fracaso del tratamiento, cambio de régimen, tratamiento incompleto o no evaluado. Los predictores se eligieron mediante selección hacia atrás con bootstrap. La discriminación y la calibración se evaluaron con c y gráficos de calibración, respectivamente. El 20 % de los participantes no tuvo éxito terapéutico. El modelo final incluía 7 predictores basales: hemoglobina, infección por VIH, consumo de drogas, diabetes, edad, educación y consumo de tabaco, similares a los descritos en otros trabajos [38]. El modelo demostró una buena discriminación (estadístico c = 0,77; intervalo de confianza del 95 %, 0,73- 0,80) y estaba bien calibrado. Utilizando información fácilmente disponible al inicio del tratamiento, el modelo de predicción funcionó bien en esta población; los factores relacionados con el VIH y el estado uso de isoniazida no mejoraron la predicción del modelo final [43]. Este estudio aportó metodologías para selección de predictores, así como el tipo de predictores finales y comparativo del porcentaje de clases, para ser discutidos acorde con los resultados obtenidos en el presente proyecto.

De igual manera en 2022 Kulkarni y otros, implementaron el uso de un indicador medida de la falta de adherencia extrema a partir de un conjunto de datos de casi 700 000 pacientes de cuatro estados de la India. Se formuló y resolvió un problema de ML de predicción temprana en la falta de adherencia terapéutica basada en una métrica personalizada. Entrenaron modelos de ML (árboles de decisión, k-NN y clasificadores bayesianos), y los evaluaron frente a líneas de base, logrando una mejora de ∼ 100 % con respecto a las líneas de base basadas en reglas y ∼ 214 % con respecto a un clasificador aleatorio. Los resultados obtenidos indican que la estratificación del riesgo de los pacientes no adherentes es una solución viable y desplegable a gran escala [44]. Esta publicación demuestra la utilidad de otro tipo de algoritmos de ML en la predicción de fallas terapéuticas, los cuales se discuten en los resultados obtenidos en el documento.

A su vez Kheirandish y colaboradores en 2022, establecieron un modelo de predicción de los resultados de tratamiento utilizando el seguimiento a pacientes nuevos con TB en Moldavia, con el fin de detectar a tiempo a aquellos en los cuales el plan de tratamiento puede no ser eficaz. Se diseñó un marco de predicción dinámica que integra modelos de referencia y algoritmos de aprendizaje automático, para predecir los resultados del tratamiento de los pacientes durante el seguimiento. Se calcularon la sensibilidad y el valor predictivo positivo (VPP) para evaluar el rendimiento del modelo en puntos críticos. Se definieron nuevas medidas para determinar cuándo debían realizarse pruebas de laboratorio de seguimiento (cultivo y baciloscopia). El algoritmo de bosque aleatorio funcionó mejor que los modelos de máquina de soporte vectorial y regresión logística

multinomial penalizada para predecir los resultados del tratamiento de la tuberculosis. Para las 3 clases de resultados (curado, no curado y fallecido a los 24 meses del inicio del tratamiento); la sensibilidad y el VPP de los modelos de predicción mejoraron a medida que se recopilaba más información de seguimiento. En concreto, la sensibilidad y el VPP aumentaron de 0,55 a 0,84 y de 0,32 a 0,88, respectivamente, para la clase no curada [45]. Este estudio aportó el tener en cuenta dentro del modelo variables de laboratorio se seguimiento (resultados de cultivos y BK), así como el uso del métricas como sensibilidad y VPP como indicador de desempeño, en la evaluación de los modelos desarrollados en el proyecto.

En nuestro país, se encuentra el reporte de un estudio realizado por Díaz en el año 2012 en Cartagena, en el cual utilizó las redes neuronales artificiales para clasificar el riesgo de contraer tuberculosis en población vulnerable, a partir de factores predictivos y propone una clasificación matemática para el nivel de riesgo. Se seleccionó una muestra probabilística conformada por 370 individuos, de una población de 10 363 personas (IC 95%, error 5%). Se realizó un análisis de tres fases: análisis descriptivo de factores; definición de propuesta de clasificación para el riesgo de tuberculosis por medio de un lenguaje de modelado matemático de factores predictivos; y construcción del modelo de RNA Perceptron Multicapa de Red Supervisada Unidireccional (MLP: Multilayered Perceptron). Dentro de los resultados se describen como factores de riesgo predictores: género (63.7 %), fumador (100 %), ingesta de alcohol (65,5 %), ingreso familiar (57,6 %), conocimiento sobre identificación de los signos y síntomas de la enfermedad (54,5 %) y en menor grado los propios síntomas (35,2 %) para la predisposición de presentar tuberculosis [46]. En este estudio se describen predictores acordes con el contexto del país y la población objeto del proyecto, los cuales se discuten más adelante acorde con los resultados obtenidos.

Por otro lado, se encuentra el estudio de Bedoya y equipo en 2023, en el cual se proponen modelos basados en cuatro técnicas de aprendizaje supervisado (redes neuronales, árboles de decisión, y dos métodos de ensamble) que permiten realizar un diagnóstico, positivo o negativo, de tuberculosis pulmonar a partir de unas variables de entrada y de diagnósticos anteriormente registrados de pacientes sanos y otros con tuberculosis pulmonar, de la ciudad de Cali. De acuerdo con los resultados obtenidos, el método de ensamble Extra Trees resulta ser el más exacto comparado con las otras técnicas utilizadas para la predicción de tuberculosis pulmonar alcanzando un área bajo la curva ROC de 95.63% [47]. Este estudio demuestra que, en nuestro país ya se utilizan herramientas avanzadas para la generación de modelos propios, acordes con nuestra realidad, para solucionar temas de diagnóstico temprano de tuberculosis. Dado que el enfoque es distinto, no se comparan los resultados de este estudio con el presente proyecto.

En el trabajo de grado de Sánchez de 2023, se diseñaron e implementaron algoritmos basados en redes neuronales artificiales, el filtro de Kalman y modelos autorregresivos, para realizar la predicción de los posibles nuevos casos de Tuberculosis en los departamentos de Colombia, basado en notificaciones de casos realizados en SIVIGILA. Los algoritmos implementados, arrojaron resultados aceptables en la predicción de casos de TB. Para medir el porcentaje de eficiencia se implementaron los índices de error MAE,

RMSE y TRS. Los índices MAE y RMSE permitieron conocer que tan dispersos estaban los datos obtenidos con los algoritmos en comparación a los datos de la serie real. Por su parte el TRS indicó que tan distantes se encontraban los datos obtenidos con relación a los picos de la señal original. El estudio hace parte del proyecto de investigación “Generación de modelos alternativos basados en inteligencia computacional para tamización y diagnóstico de Tuberculosis pulmonar” financiado por Minciencias y ejecutado por la Universidad Antonio Nariño [48]. El estudio demuestra la voluntad política para dar respuesta a esta problemática, a través del uso de herramientas de ciencias de la computación en la construcción de modelos de tipo predictivo para determinar incidencias por esta enfermedad, acorde con datos históricos del país. Dado que el enfoque de este estudio es distinto, no se comparan los resultados con el presente proyecto.

En el distrito capital se encuentra un trabajo de grado relacionado con el tema realizado por Orjuela y colaboradores en 2022, quienes utilizaron técnicas de inteligencia computacional para realizar tamización y diagnóstico de TB pulmonar activa en conjunto con profesionales dela Unidad de Servicios de Salud Santa Clara (Subred Centro Oriente), empleando algoritmos basados en redes neuronales y lógica difusa aplicados en datos que permitieron tener una caracterización de la TB en entornos con situaciones precarias [49]. Se utilizaron factores de riesgo médicos y sociodemográficos de pacientes con sospecha de TB (2017-2019) como variables predictoras para los siguientes modelos: Maquinas de soporte vectorial (MSV), Bosques aleatorios (RF), arboles de decisión (AD) y regresión logística (RL), obteniendo las siguientes métricas: exactitud con los siguientes resultados: 86 % exactitud para RL, 95 % sensibilidad MSV y 68 % especificidad MLP. Estos resultados permiten establecer que para el distrito capital ya se utilizan herramientas computacionales desde la academia y aportan otro tipo de algoritmos a ser tenidos en cuenta para la fase de modelado.

4. METODOLOGÍA

El presente proyecto se realizó siguiendo la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*), la cual consta de seis fases: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación e implementación. Es importante tener en cuenta que la fase de implementación se encuentra fuera del alcance del proyecto dado que es potestad de la entidad [50]. En la Figura 2, se describe las fases con las cuales se llevó a cabo el proyecto.

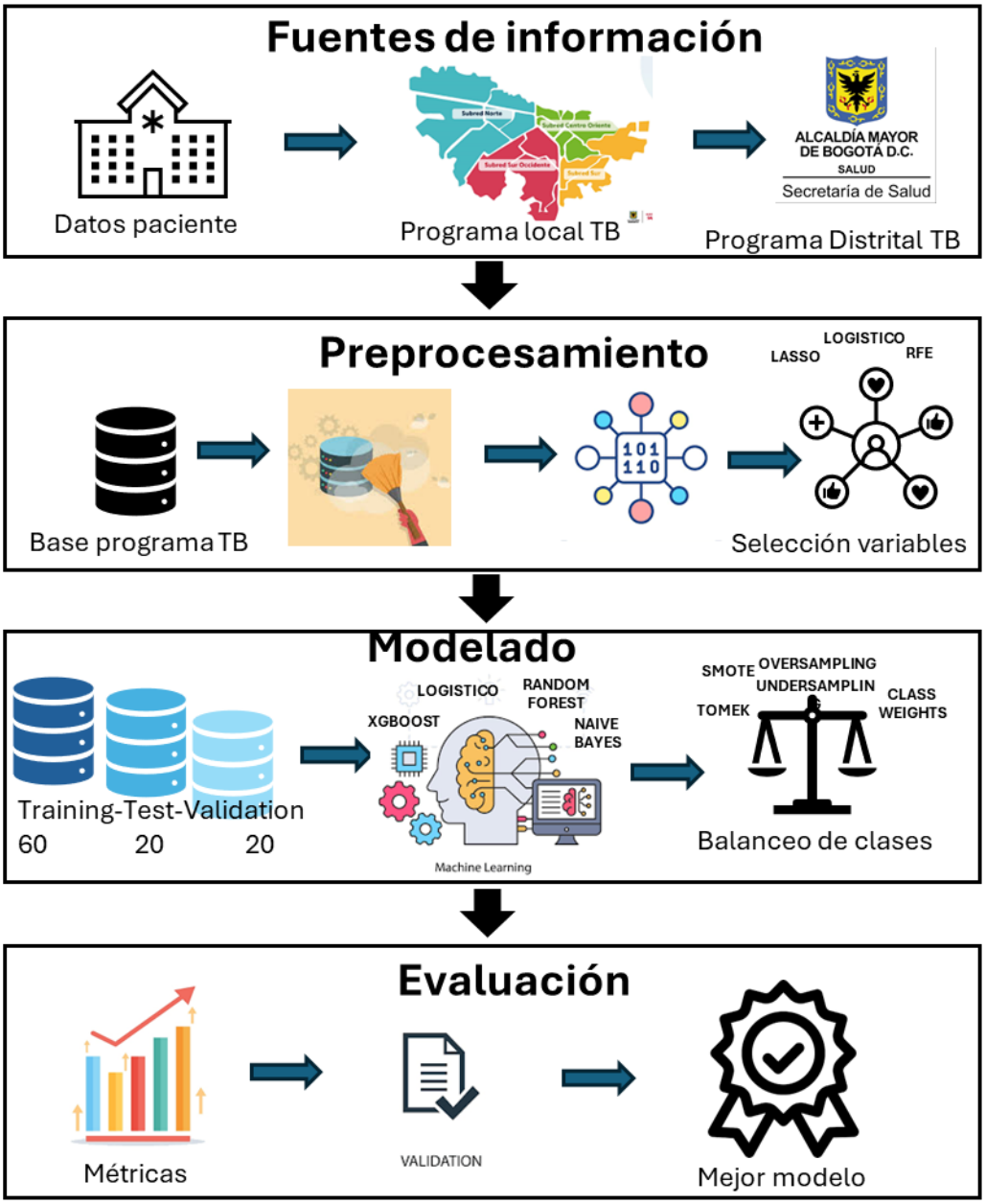

Figura 2. Diagrama de flujo del proyecto (Fuente: construcción propia).

4.1 Comprensión del negocio

El objetivo de esta fase consistió en alinear los objetivos del proyecto con los objetivos de la entidad y específicamente del programa distrital de tuberculosis. Durante esta fase, se evaluó la situación de la pérdida de seguimiento en el distrito, se fijaron los objetivos para analizar el conjunto de datos con el fin de descubrir información útil, y se trazó el plan para alcanzar los objetivos planteados en el proyecto.

Debido a que la información de la base de TB sensible del programa corresponde a datos privados de la Secretaría Distrital de Salud, se tuvo la necesidad de contar con la aprobación por parte del comité de ética de la entidad, con el cual se realizó acercamiento en el mes de diciembre de 2023, se indagó por el proceso y requisitos para presentar el anteproyecto. En enero de 2024 se postuló el anteproyecto de manera oficial ante el comité, recibiendo respuesta en el mismo mes sobre la programación de la presentación del proyecto, el cual fue presentado en la sesión ordinaria del 21 de febrero de 2024, por parte de los investigadores y la directora del proyecto (Anexo 1). Se solicitaron ajustes por parte del comité a finales de febrero y en marzo se remitieron los ajustes anexando una consulta elevada a la oficina de asuntos jurídicos de la Secretaría Distrital de Salud, sobre cesión de derechos patrimoniales y de transformación (Anexo 2), recibiendo respuesta de este (Anexo 3). En el mes de abril se recibe concepto ético aprobatorio del proyecto a través de correo electrónico (Anexo 4), y también se recibe aval con autorización para entrega de bases de datos (Anexo 5), con el cual la referente distrital del programa del tuberculosis, procedió a realizar la entrega de las bases de datos solicitadas bajo las restricciones de seguridad informática indicadas por parte del comité de ética: anonimización de los datos de los pacientes, restricciones de acceso para la Universidad, restricciones de uso acorde con la política de seguridad de la información y política de protección de datos personales establecidas en la entidad.

4.2 Comprensión de los datos

Para el proceso de entrenamiento de los modelos de aprendizaje automático, se requiere la mayor cantidad de datos disponible por tal motivo, se realizó la solicitud formal a la entidad de la información a partir el año 2009. La población del estudio fueron los pacientes diagnosticados con tuberculosis sensible, que han sido reportados al sistema de información del programa distrital de tuberculosis (Libros de TB sensible 2009-2022). No se tuvo en cuenta año 2023, dado que existen cohortes que no han finalizado el tratamiento debido al tiempo de duración de éste.

Se recibieron dos bases de datos consolidadas (2009-2015 y 2016-2022) en el mes de mayo de 2024, las cuales se entregaron completamente anonimizadas. Por requerimiento del comité de ética, las bases de datos deben ser trabajadas dentro de la entidad y deben reposar en los Drive de las cuentas institucionales por seguridad de la información; también se gestionó la instalación de R Studio y Python en los equipos de cómputo en articulación con el área de TIC de la entidad, dado que el acceso se encuentra restringido por la entidad.

4.3 Preparación de los datos

Una vez verificada la información, se encontró diferencia sustancial de los dos conjuntos de datos entregados. Se realizó reunión con la directora del quien solicitó realizar un prediagnóstico por cada una de las variables para determinar viabilidad de uso de ambas bases de datos, este análisis se describe a detalle en la Tabla 1. De esta forma, se determinó que la calidad de la base de datos 2009-2015 no permite su uso para el presente proyecto, dado que no es posible obtener la información faltante o inferirla.

Tabla 1. Análisis comparativo de las variables de las dos bases de datos 2009-2015 y 2016 a 2022 (Fuente: construcción propia).

Variable	Base desde 2016 Total Registros 10.102	Bases antes de 2015 Total Registros 8.437
IPS de Diagnóstico	nominal-no faltantes-repetidos	no existe la variable
IPS de Seguimiento de Tratamiento	nominal-207SD-Repetidos	Nombre: IPS Donde recibió el TAES-50 % SD (4860)
Fecha de Inicio de Síntomas	formato Date- 54 SD	no existe la variable
Ingresa a Tratamiento	Dicotomica-1 SD	no existe la variable
Fecha de inicio de tratamiento (dd/mm/aaaa)	formato Date- 356 SD de los cuales 14 refieren si entraron a tratamiento y 341 no tratamiento - 74 No o no aplica no tratamiento - hay dos casos con inicio de tratamiento de 2023 (eliminar)	no existe la variable, esta fecha de ingreso al programa que es diferente al tratamiento
Pertenencia étnica	nominal 6 categorías, 2 SD	Mezclado en tres variables: Grupo Poblacional (columna D), Etnia/ situación social, Grupo Poblacional (columna AH). 2987 SD 12 categorías
Pueblo indígena	nominal, ajustar 3 no indígenas (No, NA, Ninguno), 42 indígenas sin pueblo diligenciado	no existe la variable-solo 3 registros con pueblo indígena inga
Grupo poblacional	nominal 16 categorías, 53 SD de los cuales 52 refieren otro en pertenencia étnica y 1 SD -dos categorías se deben unificar (otros y PPL)	nominal, 12 categorías que se encuentran en tres variables: Grupo Poblacional (columna D), Etnia/ situación social, Grupo Poblacional (columna AH). 2987 SD
Barrio de residencia	nominal, 114 SD, barrios de otras ET	nominal, 6148 SD
Tipo tuberculosis	dicotómica-completo	3 categorías- 8 SD
Localización de la TB extrapulmonar	24 registros inconcordantes-	no existe la variable
Prueba molecular	ordinal (-+) -SD 3957 (recodificar Pos o Neg),	no existe la variable
Se realizó APV	nominal, 4 categorías, 37 SD	Nombre: Consejería Pre y Post- 6 categorías – 30 % SD (2281)
Se realizó prueba	nominal, 4 categorias, 66 SD-	no existe la variable
Resultado prueba	nominal, 4 categorías, 531 SD-	no existe la variable
Prueba confirmatoria acorde a la norma	nominal, 4 categorías, 2 SD-4	no existe la variable
Recibe TAR	nominal, 3 categorías, 2 inconcordancias-53 SD	nominal, 5 categorías, 453 SD

Recibe trimetoprim (TMSX)	nominal, 3 categorías, 1 inconcordancias-56 SD	nominal, 4 categorias-729 SD
BK (Final 1ª Fase)	ordinal, 10 categorías (recodificar Pos o Neg), 4512 SD	ordinal, 4 categorías, (recodificar Pos o Neg), 50 % SD (4824)
BK (Mitad de la 2ª Fase)	ordinal, 8 categorías (recodificar Pos o Neg), 5201 SD	ordinal, 3 categorías, (recodificar Pos o Neg), 50 % SD (4975)
BK (Final del tto)	ordinal, 7 categorías (recodificar Pos o Neg), 3 5102 SD	ordinal, 4 categorías, (recodificar Pos o Neg), 50 % SD (4999)
Cultivo al final del tratamiento	ordinal, 4 categorías, SD 6329, con Fracaso y cultivo y BK negativo	no existe la variable
Tipo de farmacorresistencia	nominal, 13 categorías (recodificar)	no existe la variable
Condición de egreso	nominal, 8 categorías, 195 SD (27 migrantes, 10 habitantes de calle, carcelarios FDB, 2 desplazados-9	nominal, 11 categorías, 942 SD
FECHA DE EGRESO (dd/mm/aaaa)	2 fechas erradas, 1553 SD	no existe la variable
Comorbilidad 1	nominal, separar categorías, más de una comorbilidad asociada	nombre: Comorbilidades Asociadas, solo se diligenció una
Comorbilidad 2	nominal, unificar tipo de categorías	
Comorbilidad 3	nominal, unificar tipo de categorías	
Departamento de residencia	nominal, (recategorizar Bta-FDB)-5543 SD se pueden recuperar 5419 con barrio y comuna	no existe la variable, se puede inferior de localidad y barrio de residencia-763 SD
Municipio de residencia	nominal, (recategorizar Bta-FDB)-5545 SD se pueden recuperar 5419 con barrio y comuna	no existe la variable, se puede inferior de localidad y barrio de residencia-763 SD
Modalidad de tratamiento directamente observado	nominal, 5 categorias-50%SD (5572)	no existe la variable
Tipo de programas de protección social que recibe	nominal, 5 categorías-completo	no existe la variable
Reacciones adversas al tratamiento	ordinal, 4 categorías-completo	no existe la variable
Metodología de captación del caso	nominal, 4 categorías-completo	no existe la variable
Fecha de nacimiento	Formato Date-50% SD (5567)-	no existe la variable
Subred	nominal, 4 categorías, 8924 SD, no se especifica subred de residencia o de DX- Se puede generar a partir de residencia o DX o ambas	no existe la variable

Por lo anterior, el proyecto se desarrolló con la base de 2016 a 2022, ya que cuenta con calidad, completitud y suficiencia de registros. Esta base contiene un total de 104 variables y 10.102 observaciones. Las variables contienen información del paciente: sexo, edad, grupo poblacional, residencia, régimen de afiliación al SGSSS; contienen información de la enfermedad: fecha inicio de síntomas, IPS diagnóstico, resultados de pruebas de laboratorio, tipo de tuberculosis, tipos de resistencias y existencia de otras comorbilidades; contiene información del programa: fecha de ingreso, condición de ingreso y egreso al programa, IPS de seguimiento, modalidad de tratamiento, reacciones adversas a medicamentos, pertenencia a programas de protección social, metodología de captación del caso, localidad de seguimiento, persona que realiza el seguimiento al paciente, cierre

del caso, soporte del cierre del caso, observaciones.

Se realizó depuración de la información: se estandarizaron formatos de las variables, se unificaron datos, se diligenciaron campos vacíos como no aplica, se recategorizaron la mayoría de las variables (resultados de laboratorio), con el fin de facilitar los análisis estadísticos descriptivos. Producto de este proceso, se obtuvo una base de datos con un total de 60 variables y 10.102 observaciones.

Con el fin de realizar entendimiento de los datos, se procedió a realizar análisis estadístico univariado por cada una de las variables utilizando software R Studio el cual se detalla en el numeral 5.2. También se generó un análisis bivariado teniendo como punto de referencia la condición de egreso, donde se encuentra la pérdida de seguimiento. A las variables en las que se identificó que la proporción era más alta para esta condición de egreso, se les aplicó test de Chi cuadrado e índice de Cramer para determinar la fuerza de asociación (p valor <= 0.05), utilizando el software R Studio, los resultados del análisis se detallan en el numeral 5.3. De esta manera, se hizo una primera selección de variables consensuado con la directora del proyecto, que incluyó 17 predictores a utilizar en la fase de modelado, que describen en el numeral 5.4.

Posteriormente, al aplicar la discretización de las variables categóricas fue necesario realizar reducción de dimensionalidad, para ello se hizo una revisión metodológica determinando como apropiados el uso de los siguientes métodos: Group LASSO, Regresión logística combinado con *Fordward, Backward* y *Stepwise,* así como el método RFE, los resultados se detallan en el numeral 6.2. En esta fase se utilizó Pyhton con IDE Colab de Google.

4.4 Modelado

Esta fase es importante ya que a partir de ella se logra cumplir el objetivo principal de construir un modelo de predicción, a partir de la clasificación de etiquetas basado en el aprendizaje automático.

Teniendo en cuenta el desbalance de clases, fue necesario aplicar diferentes técnicas de remuestreo en la fase de entrenamiento para garantizar un adecuado desempeño en las métricas acorde con las clases de la variable objetivo (pérdida de seguimiento), dentro de las que se aplicaron: *Oversampling,* SMOTE, *Undersampling, Tomek* y combinaciones de éstas. Los resultados se detallan el numeral 6.3. Para esta fase se utilizó Pyhton con IDE Colab de Google, dado debido a que cuenta con un mayor número de librerías disponibles que el software R Studio.

Se utilizaron 4 tipos de algoritmos supervisados: Random Forest [16], Regresión Logística [16], XGBoost [21] y Naive Bayes [20], en el cual se usaron las variables seleccionadas a partir de las técnicas de reducción de variables aplicadas y las técnicas de balanceo de clases. Para optimizar los parámetros de los modelos, se empleó un proceso de validación

cruzada de diez pasos combinado con una búsqueda en grilla. Los resultados de cada uno de los modelos se describen en el ítem 6.4. Para esta fase se utilizó Pyhton con IDE Colab de Google, dado debido a que cuenta con un mayor número de librerías disponibles que el software R Studio.

4.5 Evaluación

En esta fase se tuvo en cuenta el desempeño de los modelos con las diferentes métricas obtenidas a partir del modelamiento. Teniendo en cuenta el gran desbalance de clases en la base de datos, se definió utilizar la métrica Sensibilidad (*Recall)* como base para determinar el desempeño en la adecuada clasificación del modelo, esto teniendo en cuenta que la exhaustividad resume qué tan bien se predijo la clase positiva y corresponde al mismo cálculo para la sensibilidad (verdaderos positivos), es decir que permite identificar a los pacientes que realizan pérdida de seguimiento en el programa distrital de tuberculosis. La exhaustividad se define como la relación de los verdaderos positivos sobre la suma de verdaderos positivos y falsos negativos [51] a partir de la matriz de confusión y se calcula de la siguiente forma:

Sensibilidad (Recall) = Verdaderos Positivos / (Verdaderos Positivos + Falsos Negativos)

El valor obtenido puede variar de 0 a 1 y por lo general, se expresa de forma porcentual sobre el total de observaciones de la matriz de confusión (n). Los resultados de las métricas de cada uno de los modelos se describen en el numeral 7. Para esta fase se utilizó Pyhton con IDE Colab de Google, dado debido a que cuenta con un mayor número de librerías disponibles que el software R Studio.

Con el fin de realizar una validación más exhaustiva se usó un tercer conjunto de datos (validación) a partir del mismo conjunto de datos, conteniendo el 20% del total de base de datos como se describe en el numeral 6.2. Adicionalmente, se hizo una segunda validación externa utilizando el conjunto de la base de datos del programa preliminar del año 2023, la cual contiene 1.953 observaciones y se realizó el mismo procedimiento de preprocesamiento ya descrito. En esta base de datos se cuenta con 137 pérdidas de seguimiento, que corresponde a la clase 1.

4.6 Despliegue

El archivo que contiene el modelo seleccionado será cedido a la entidad a través de un contrato de cesión de derechos patrimoniales a cargo de la Subdirección de contratación. El cual podrá ser utilizado por la referente distrital del programa de tuberculosis para el despliegue a los equipos locales que permitan la identificación temprana de pacientes con alto riesgo de pérdida de seguimiento, al ingreso del programa. Por último, se aclara que se encuentra fuera de alcance del presente proyecto, la integración del mejor modelo en un sistema de producción de la entidad.

5. ANÁLISIS EXPLORATORIO DE DATOS

Del apartado de datos generales se usaron 17 variables (Tabla 2):

Tabla 2. Variables del componente datos generales base TB sensible requeridas. (Fuente: construcción propia).

Nombre variable	Definición	Posibles valores	Tipo variable	Formato	Importancia
IPS DE DIAGNÓSTICO	Nombre del prestador que realizó el diagnóstico del caso de tuberculosis.	Múltiples	Nominal	String	Alta
IPS DE SEGUIMIENTO DE TRATAMIENTO	Nombre del prestador donde actualmente está recibiendo el tratamiento el usuario.	Múltiples	Nominal	String	Alta
FECHA DE INICIO DE SÍNTOMAS (dd/mm/aaaa)	Fecha de inicio de síntomas presuntivos en la persona que fue confirmada posteriormente con la TB	Fecha	Fecha	Date	Alta
INGRESA A TRATAMIENTO	Seleccionar si el paciente ingresó o no al tratamiento de la tuberculosis.	Si, No	Nominal	Binaria	Alta
FECHA DE INICIO DE TRATAMIENTO (dd/mm/aaaa)	Fecha de ingreso a tratamiento del paciente	Fecha	Fecha	Date	Alta
SEXO	Seleccionar el sexo	Femenino, Masculino	Nominal	Binaria	Alta
EDAD (EN AÑOS)	Si es un niño o niña con meses de nacimiento, se selecciona <1año	Numérico	Discreta	Interger	Alta
PERTENENCIA ÉTNICA	Diligenciar la pertenencia étnica acorde con autorreconocimiento	Indígena, afrodescendiente, rom, raizal o gitano	Nominal	String	Alta
PUEBLO INDÍGENA	Se la persona tiene pertenencia étnica como indígena se debe reportar el pueblo al cual pertenece	Múltiples	Nominal	String	Alta
BARRIO DE RESIDENCIA	Barrio en áreas urbanas o en zonas rurales la vereda de la persona afectada por tuberculosis	Múltiples	Nominal	String	Media
COMUNA/LOCALIDAD DE RESIDENCIA	comuna o localidad de residencia de la persona afectada por tuberculosis	Múltiples	Nominal	String	Alta
RÉGIMEN DE AFILIACIÓN	desplegable el régimen al que pertenece el usuario, según corresponda:	Contributivo, subsidiado, no asegurado, especial, excepción	Ordinal	String	Alta
EAPB	desplegable la EAPB según corresponda.	Múltiples	Nominal	String	Media

Nombre variable	Definición	Posibles valores	Tipo variable	Formato	Importancia
GRUPO POBLACIONAL	Diligenciar el grupo poblacional al que pertenece el paciente, desplegable:	discapacidad, desplazado, migrante, población carcelaria, gestante, habitante de calle, población infantil a cargo del ICBF, madres comunitarias, desmovilizados, población en centros psiquiátricos, LGBTIQ+, víctima de la violencia armada, trabajador de la salud	Nominal	String	Alta
TIPO TUBERCULOSIS	localización anatómica de la enfermedad	Tuberculosis pulmonar, extrapulmonar	Nominal	String	Alta
LOCALIZACIÓN DE LA TB EXTRAPULMONAR	desplegable el tipo de tuberculosis extrapulmonar.	· Meníngea · Peritoneal · Ganglionar · Renal · Intestinal · Osteoarticular · Genitourinaria · Pericárdica · Cutánea · Pleural · Otro	Nominal	String	Alta
CONDICIÓN DE INGRESO	desplegable la condición de ingreso	Nuevo, Reingreso tras recaída, Reingreso tras fracaso, Reingreso tras pérdida en el seguimiento, Otros previamente tratados, Remitido	Nominal	String	Alta

Del componente diagnóstico de la tuberculosis, se usaron 3 variables (ver Tabla *3*):

Tabla 3. Variables del componente diagnóstico de la tuberculosis, base TB sensible requeridas. (Fuente: construcción propia).

Nombre variable	Definición	Posibles valores	Tipo variable	Formato	Importancia
BK	resultado de la baciloscopia	negativo, (1 a 9) BAAR,(+),(++),(+++), No realizado(NR).	Ordinal	String	Alta
CULTIVO LÍQUIDO	resultado de cultivo de diagnóstico	Negativo, (1 a 19) BAAR,(+),(++),(+++), No Realizado (NR), contaminado, en proceso.	Ordinal	String	Alta
PRUEBA MOLECULAR	resultado de la prueba molecular	Detectado, No detectado, No interpretable, Contaminado	Ordinal	String	Alta

De las actividades colaborativas tuberculosis y VIH, se utilizaron 6 variables (ver Tabla 4):

Tabla 4. Variables del componente COINFECCIÓN TB/VIH, base TB sensible requeridas. (Fuente: construcción propia).

Nombre variable	Definición	Posibles valores	Tipo variable	Formato	Importancia
SE REALIZÓ APV	Usuario recibió asesoría para Prueba de VIH, desplegable	Si, No, paciente no acepta, VIH + previo.	Ordinal	String	Media
SE REALIZÓ PRUEBA	Usuario se realizó la prueba de VIH desplegable,	Si, No, paciente no acepta, VIH + previo.	Ordinal	String	Media
RESULTADO PRUEBA	desplegable el resultado de la prueba presuntiva del VIH	Si, No, paciente no acepta, VIH + previo.	Ordinal	String	Alta
PRUEBA CONFIRMATORIA ACORDE A LA NORMA	Solo se diligencia si el resultado de la prueba presuntiva fue positiva o el paciente es VIH + Previo	Si, No, paciente no acepta, VIH + previo.	Ordinal	String	Alta
RECIBE TAR	siempre y cuando se confirme el resultado positivo de la prueba de VIH o el caso sea VIH + PREVIO,	Si, No	Nominal	Binaria	Media
RECIBE TRIMETOPRIM (TMSX)	siempre y cuando se confirme el resultado positivo de la prueba de VIH o el caso sea VIH + PREVIO,	Si, No	Nominal	Binaria	Media

Del control bacteriológico, se utilizaron las 4 variables de la base de datos (ver Tabla 5):

Tabla 5. Variables del componente control bacteriológico, base TB sensible requeridas. (Fuente: construcción propia).

Nombre variable	Definición	Posibles valores	Tipo variable	Formato	Importancia
BK (Final 1ª Fase)	desplegable el resultado de BK seriado de control de final de la primera fase	Negativo, 1 a 9 BAAR, (+), (++), (+++), no realizado	Ordinal	String	Media
BK (Mitad de la 2ª Fase)	desplegable el resultado de BK seriado de control de mitad de la segunda fase	Negativo, 1 a 9 BAAR, (+), (++), (+++), no realizado	Ordinal	String	Media
BK (Final del tto)	desplegable el resultado de BK seriado de final de tratamiento	Negativo, 1 a 9 BAAR, (+), (++), (+++), no realizado	Ordinal	String	Media
CULTIVO AL FINAL DEL TRATAMIENTO	desplegable el resultado de cultivo de final del tratamiento	Negativo, 1 a 9 BAAR, (+), (++), (+++), no realizado	Ordinal	String	Media

De la susceptibilidad a fármacos, se utilizaron 2 variables de la base de datos (ver Tabla 6):

Tabla 6. Variables del componente susceptibilidad a fármacos, base TB sensible requeridas. (Fuente: construcción propia).

Nombre variable	Definición	Posibles valores	Tipo variable	Formato	Importancia
PRUEBA DE SUSCEPTIBILIDAD A FÁRMACOS	desplegable el tipo de prueba de susceptibilidad a fármacos antituberculosos realizada,	· Nitrato reductasa · proporciones en LJ · BACTEC MGIT · proporciones en agar · LIPA (Genotype, Anyplex) · PCR en T' real (Genexpert) · no realizada	Nominal	String	Media
TIPO DE FARMACORRESISTENCIA	desplegable el tipo de farmacorresistencia	Monorresistencia: Polirresistencia, MDR, RR, PreXDR, XDR	Ordinal	String	Alta

Del componente condición de egreso, se utilizaron 2 variables (ver

Tabla *7*):

Tabla 7. Variables del componente condición de egreso, base TB sensible requeridas. (Fuente: construcción propia).

Nombre variable	Definición	Posibles valores	Tipo variable	Formato	Importancia
CONDICIÓN DE EGRESO	Resultado del tratamiento de las personas afectadas por tuberculosis sensible	Curado, Tratamiento Terminado, Fracaso, Pérdida de Seguimiento, Fallecido, No Evaluado, Descartado	Nominal	String	Alta
FECHA DE EGRESO (dd/mm/aaaa)	fecha de egreso del tratamiento	Fecha	Fecha	Date	Alta

De la condición de comorbilidades, se utilizaron las 3 variables (ver Tabla 8):

Tabla 8. Variables del componente comorbilidades, base TB sensible requeridas. (Fuente: construcción propia).

Nombre variable	Definición	Posibles valores	Tipo variable	Formato	Importancia
COMORBILIDAD 1 **COMORBILIDAD 2** **COMORBILIDAD 3**	Seleccionar del desplegable la primera comorbilidad asociada a la persona afectada por tuberculosis,	· Alcoholismo · Cáncer · Consumo SPA · Desnutrición · Diabetes Mellitus · Enfermedad autoinmune · Enfermedad hepática · Enfermedad Renal Crónica · EPOC · Ninguna · Silicosis · Tabaquismo · COVID-19 Otras	Nominal	String	Alta

Del componente de seguimiento al tratamiento directamente observado, se utilizaron 6 variables de la base de datos (ver Tabla 9):

Tabla 9. Variables del componente datos de seguimiento al tratamiento directamente observado, base TB sensible requeridas. (Fuente: construcción propia).

Nombre variable	Definición	Posibles valores	Tipo variable	Formato	Importancia
MODALIDAD DE TRATAMIENTO DIRECTAMENTE OBSERVADO	desplegable el tipo de TDO en el que se le está administrando el tratamiento	TDO en IPS TDO Domiciliario TDO Comunitario TDO Hospitalario TDO virtual	Nominal	String	Alta
TIPO DE PROGRAMAS DE PROTECCIÓN SOCIAL QUE RECIBE	desplegable el tipo de programas de protección social que recibe el paciente con tuberculosis	· Subsidio alimentario · Subsidio de vivienda · Subsidio de desempleo · Subsidio educativo · Subsidio monetario · Cuenta con varios subsidios de apoyo · No aplica a subsidios	Nominal	String	Alta
REACCIONES ADVERSAS AL TRATAMIENTO	desplegable si el paciente presenta reacciones adversas al tratamiento.	Leve Moderado Grave Ninguna	Ordinal	String	Alta
METODOLOGÍA DE CAPTACIÓN DEL CASO	desplegable la metodología por el cual se captó el caso de tuberculosis	· Búsqueda activa institucional · Búsqueda activa derivado de agente comunitario · Búsqueda activa derivado de trabajador de la salud. · Remitido por el Centro Nacional de Enlace-CNE · Durante estudio de contactos	Nominal	String	Alta
OBSERVACIONES	observaciones relacionadas con el paciente de tuberculosis	Múltiples	Nominal	String	Baja
ÁREA GEOGRÁFICA	Zona geográfica urbana, rural o rural dispersa. Subred de residencia	Subred Norte Subred Centro Oriente Subred Sur Subred suroccidente	Nominal	String	Baja

5.1 Limpieza y transformación de datos

Una vez se cuenta con el conjunto de datos, se verificó calidad en el diligenciamiento de la base en formato XLS, verificando:

- Se verificó en cada variable que estuviera en el formato correspondiente: fechas, variables categóricas y numéricas, acorde con el diccionario de datos.
- Registros vacíos dependiendo de la variable, se diligenció a partir de otras variables como edad a partir de fecha de nacimiento. Subred de residencia a partir de la dirección, barrio o municipio en el caso de Fuera de Bogotá.
- Coherencia: se verificaron variables relacionadas como tipo de TB y localización anatómica, en los casos en los que se identificó inconcordancias se solicitó la revisión a la referente distrital del programa quien realizó cruces de información con otras fuentes entregando el dato verificado casos inconcordantes de VIH, grupos poblacionales y comorbilidades principalmente.
- Estandarización de variables: se realizó depuración de errores de digitación en cada una de las variables. Se estandarizaron las variables a partir de tablas de referencia para las variables: localidad de residencia, localidad de diagnóstico, tipos de farmacorresistencia, comorbilidades, localización TB, grupo poblacional e IPS de diagnóstico.
- Recodificación de variables: se realizó con las variables de resultados de laboratorios para BK y cultivo +, ++, +++ y 1 A 9 BAAR se recategorizaron a Positivo, - a Negativo, NR a No realizado y los vacíos se manejaron como Sin dato (SD). En la prueba molecular Detectado se recategorizó a Positivo y No detectado como Negativo, los vacíos se diligenciaron como No realizado. En VIH las categorías positivo y VIH previo, se recategorizaron como positivo, Paciente no acepta como No realizado.
- Nuevas variables: se generó subred de residencia y subred de diagnóstico con el fin de agrupar para el análisis debido a la gran cantidad de categorías (20 localidades). Se generó la variable objetivo en formato binario pérdida de seguimiento (Si, No), a partir de la condición de egreso, pérdida en el seguimiento.
- No se realizó imputación de datos, ni se eliminaron registros.

Posterior a la limpieza de la base de datos, se seleccionaron las variables relevantes y susceptibles de análisis acorde con el conocimiento que se cuenta sobre el evento de tuberculosis y la experiencia en el manejo de la base del programa. Producto de este proceso, se obtuvo una base de datos con un total de 60 variables y 10.102 observaciones, con la cual se realizó el análisis descriptivo cuyos resultados se muestran en el apartado 5.2.

5.2 Análisis Descriptivo

El conjunto de datos utilizados para la resolución de este proyecto consta de 10.102 observaciones y 60 atributos. De los cuales 2 son variables cuantitativas (Edad y año), 3 corresponden a fechas y las demás (n=55) corresponden a variables categóricas. A continuación, se presenta el análisis de las principales variables, el análisis completo del conjunto de datos se encuentra disponible en el link de Rpubs.

La distribución de casos por año se muestra en la Figura 3, donde se evidencia que el menor número de casos se presentó en 2016 con 1.311 casos, y el mayor número en 2022 con 1.808 casos. La media de casos por año fue de 1.443 casos con una desviación estándar de 189 casos por año.

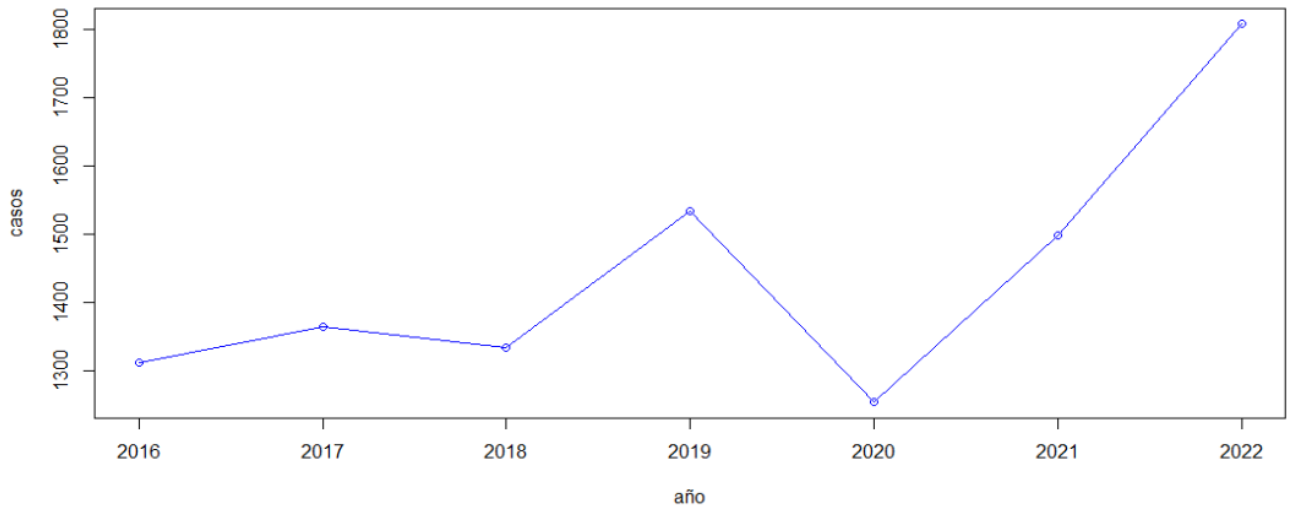

Figura 3. Distribución de casos de TB por año por año (2016-2022) en el distrito capital (Fuente: construcción propia. Bases distritales del programa de TB 2016-2022)

Con relación a la edad, la media y mediana es de 50 años con un máximo de 109 años y un mínimo, de menor de 1 año o 12 meses de edad. La variable no tiene una distribución normal, sino trimodal (2, 28 y 69 años) como se observa en la Figura 4.

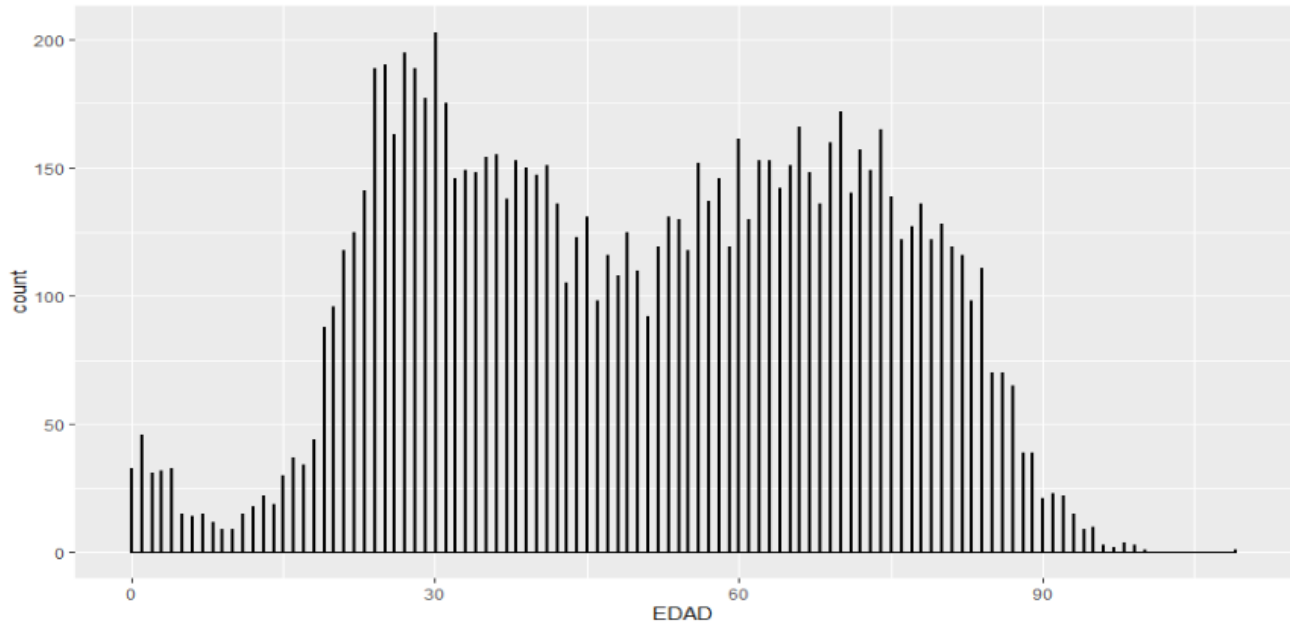

Figura 4. Distribución de la edad en casos de TB (2016-2022) en el distrito capital (Fuente:

construcción propia. Bases distritales del programa de TB 2016-2022)

El 66.2 % de los pacientes que ingresaron al programa de TB entre 2016 y 2022 son hombres (6.687), el porcentaje restante corresponde a mujeres. La razón hombre/mujer es de 1.95, es decir que por cada dos hombres que padecen de TB se encuentra una mujer en el distrito capital. En la Figura 5, se observa el comparativo de casos distribuido por año y por sexo.

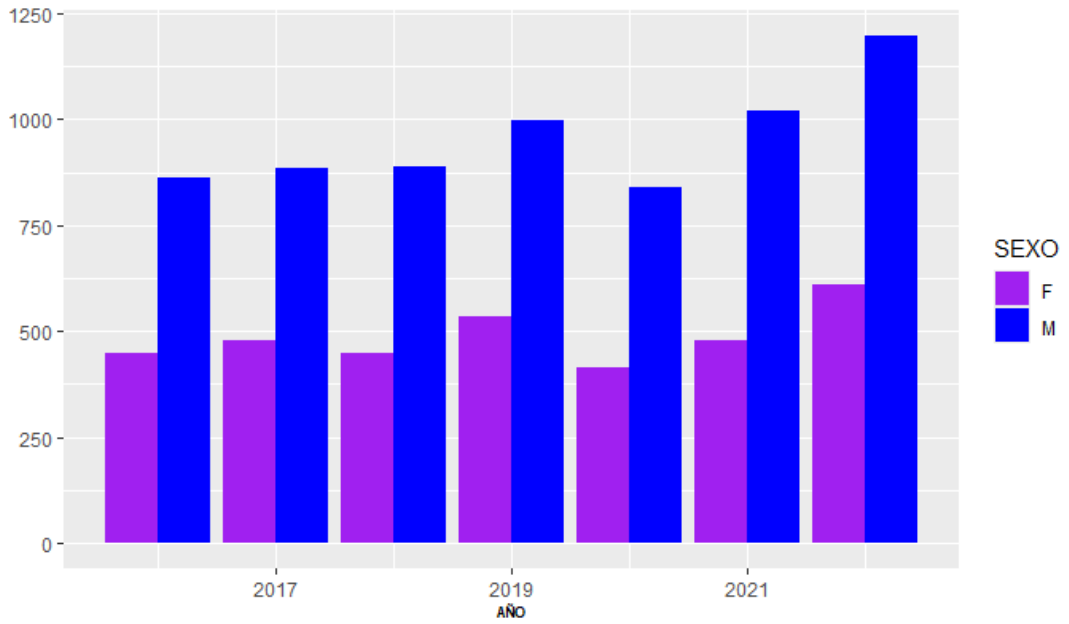

Figura 5. Distribución de casos de TB por año y sexo (2016-2022) en el distrito capital (Fuente: construcción propia. Bases distritales del programa de TB 2016-2022)

El 95.7 % de las personas que ingresaron al programa distrital de TB entre 2016 y 2022, iniciaron tratamiento tetraconjugado. Del porcentaje restante de quienes no iniciaron tratamiento, el 72 % fallecieron o el diagnóstico fue *post mortem* y en el 14.3 % se descartó la tuberculosis como enfermedad causante del cuadro clínico. En el 8 % no alcanzaron a iniciar tratamiento porque corresponden a pérdidas en el seguimiento.

Con relación a la pertenencia étnica se encuentra que la mayoría corresponde a indígenas 1.5 %, seguido de población afrocolombiana 0.8 % y ROOM 0.12 %. El 97.4 % no se autorreconoce como parte de una etnia. Dentro de los pacientes pertenecientes a etnia indígena se encuentra que los pueblos más frecuentes son: Embera y Embera-Katío con el 45% de la representación de los pueblos indígenas.

En cuanto a los grupos poblacionales especiales, se encuentra que el 1.33 % de los pacientes con TB refieren tener una condición de discapacidad (n=135), el 0.78 % refieren ser víctimas del desplazamiento forzado (n=79), el 5.2 % corresponden a población migrante (n=523), el 3.4 % de los pacientes son personas privadas de la libertad (n=339), el 0.6 % son gestantes (n=20), el 5.4 % a habitantes de calle (n=541), el 0.14 % a menores

en protección del ICBF (n=14), el 0.009 % madres comunitarias (n=1), el 0.11 % personas reportadas como población psiquiátrica (n=11), el 0.16 % son personas víctimas de la violencia por conflicto armado (n=16), el 2 % se reportan como trabajadores de la salud (n=204). No se reportan desmovilizados dentro de los grupos poblacionales.

A su vez, la localidad de diagnóstico se evidencia que los casos son diagnosticados en su mayoría en IPS de la Subred Norte la cual concentra el 34.7 % de los casos diagnosticados en el distrito. Esto se debe a la oferta de servicios por parte de los prestadores de servicios de salud en Bogotá, se concentra en las localidades de Chapinero 12.6 %, Usaquén 12 % y Teusaquillo 10.07 %. La tercera localidad con mayor número de casos diagnosticados que no pertenece a la subred Norte es Los Mártires con 11.3 % y la cuarta localidad es San Cristóbal con 9.4 %, seguido de Kennedy con 8 %, esto se debe a que en estas localidades se encuentran instituciones de alta complejidad con buena capacidad de diagnóstico para TB.

De igual manera, al analizar la localidad de residencia como se observa en la Figura 6, se encuentra que el 19.2 % (n=1.942 casos) pese a ser diagnosticados en Bogotá residen fuera de la ciudad, el 80.8 % de los casos restantes (n=8.160) residen en localidades de Bogotá, siendo la localidad de Suba 12 % donde más residen, seguido de Kennedy 11.8 %, Engativá 9 %, Ciudad Bolívar 8.2 %, Rafael Uribe Uribe 8.1 % y Bosa 7 %, las que concentran la mayor cantidad de casos de los últimos 7 años.

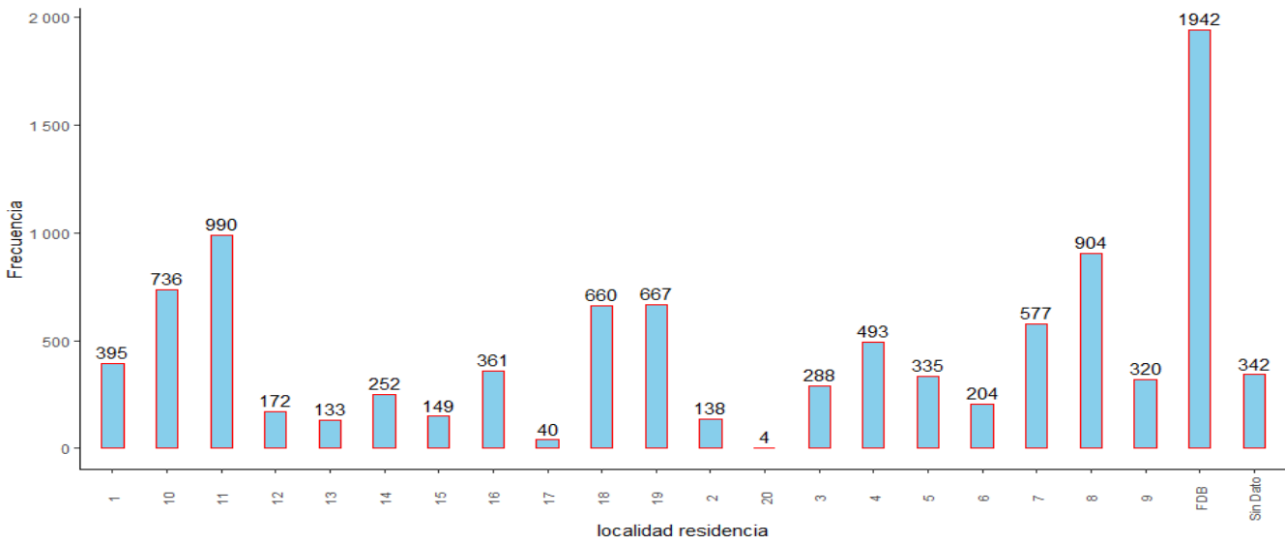

Figura 6. Distribución de casos de TB por localidad de residencia (2016-2022) en el distrito capital (Fuente: construcción propia. Bases distritales del programa de TB 2016-2022)

Con relación al aseguramiento, se observa en la Figura 7 que el 55.4 % de los pacientes diagnosticados con TB en Bogotá pertenecen al régimen contributivo (n=5.599), seguido del régimen subsidiado con 30.6 % (n=3.088), en el 6.6 % los pacientes no se encuentran asegurados al sistema de salud (n=669), en el 5.1 % pertenecen a regímenes especiales (n=516) y en el 2.3 % a régimen de excepción (n=230).

En cuanto al tipo de TB, se encuentra que el 69 % de los casos confirmados en el distrito son TB pulmonares (n=6.940) mientras que el 31 % corresponden a formas extrapulmonares. De las formas extrapulmonares, la localización anatómica más frecuente es pleural 32.4 % (n=1.023), seguida de meníngea 26.8 % (n=848), ganglionar 11.6 % (n=366) entre otras formas.

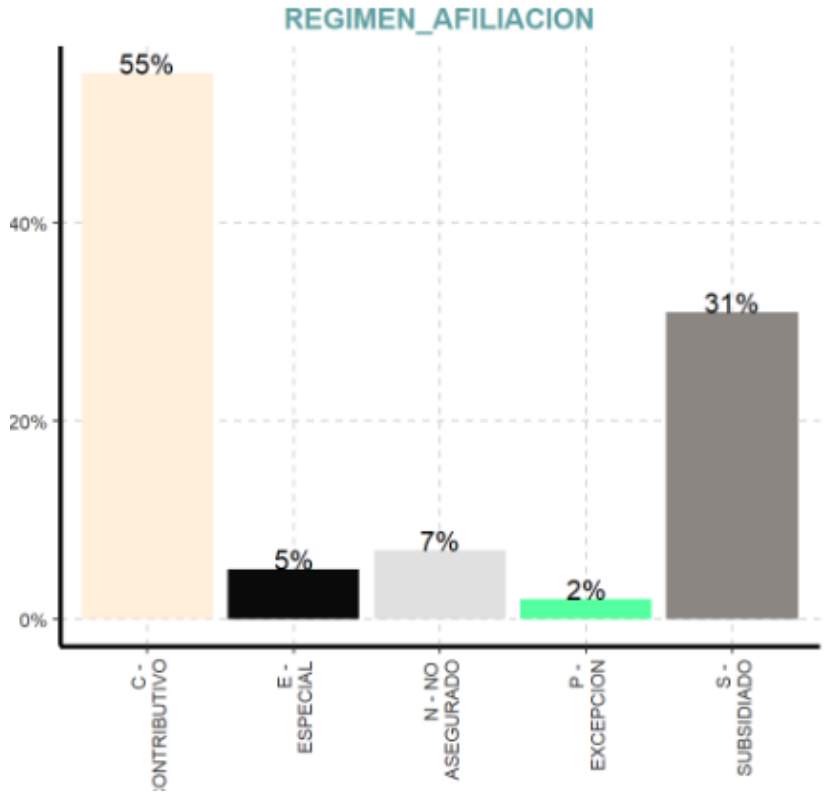

Figura 7. Distribución de casos de TB según régimen de afiliación (2016-2022) en el distrito capital (Fuente: construcción propia. Bases distritales del programa de TB 2016-2022)

Así mismo, en cuanto a la condición de ingreso de los pacientes al programa distrital de TB se encuentra que el 93.4 % corresponden a casos nuevos, mientras que el 3.1 % son previamente tratados, de los que no se tiene traza de los tratamientos previos recibidos. Por otro lado, de quienes es posible verificar el tratamiento previo, el 2.2 % de los casos corresponden a recuperados tras pérdida en el seguimiento, 1.06 % reingresos por recaídas y 0.25 % reingreso tras fracaso del tratamiento, como se observa en la Figura 8.

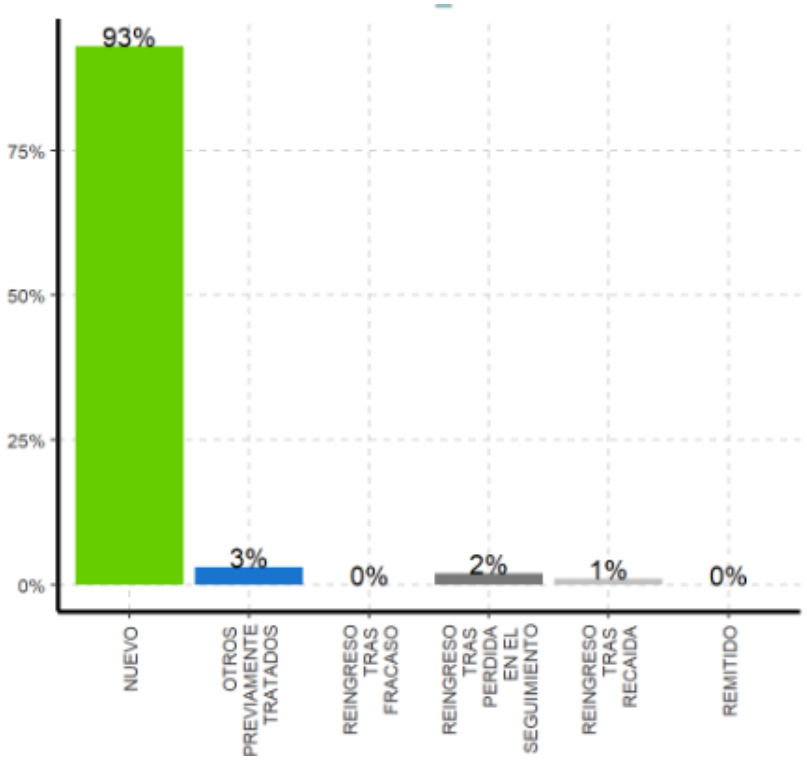

Figura 8. Distribución de casos de TB según condición de ingreso al programa (2016-2022) en el distrito capital (Fuente: construcción propia. Bases distritales del programa de TB 2016-2022)

La condición de egreso de los pacientes adscritos al programa de TB del distrito (variable objetivo), el 61 % corresponde a curaciones (para el caso de TB pulmonar) y tratamientos finalizados (TB extrapulmonar); es decir, el éxito programático del distrito de los últimos 7 años. Como se observa en la Figura 9, en el 21.5 % los pacientes fallecieron durante o previo al tratamiento (diagnóstico post mortem), es decir que el diagnóstico de la enfermedad fue tardío. En el 5.2 % de los pacientes se descartó el diagnóstico de tuberculosis, esto sucede cuando se identifican que se trata de micobacterias que no hacen parte del complejo tuberculosis (no tuberculosas). La pérdida de seguimiento en los últimos 7 años para el distrito corresponde al 7.3 %. En el 3 % corresponde a no evaluados, esto se da porque no fue posible identificar si el paciente logró terminar su tratamiento de manera exitosa, por lo general corresponde a pacientes que residen fuera de Bogotá y no se obtiene realimentación del ente territorial frente a la finalización del tratamiento. En menor proporción se encuentran los fracasos terapéuticos 0.6 %, en los cuales la enfermedad persiste frente al tratamiento de primera línea, así como exclusiones por resistencias a medicamentos de primera línea para TB 0.92 %, en este caso hacen parte de formas resistentes de la enfermedad y se debe instaurar tratamiento de segunda línea y se excluyen de esta cohorte.

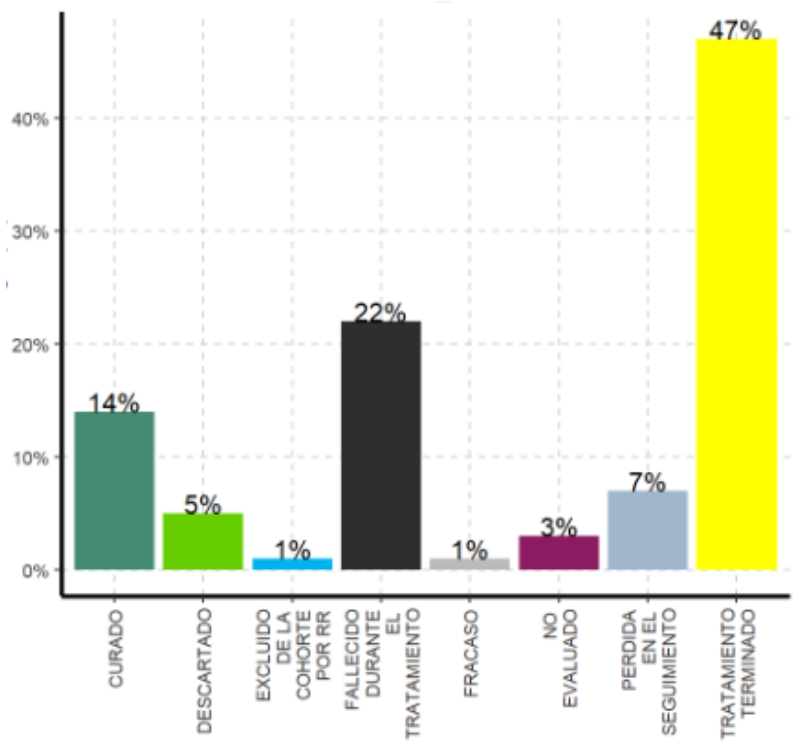

Figura 9. Distribución de casos de TB según condición de egreso del programa (2016-2022) en el distrito capital (Fuente: construcción propia. Bases distritales del programa de TB 2016-2022)

Por otro lado, en cuanto a la confirmación por laboratorio clínico se encuentra en el 54 % las baciloscopias (BK) fueron negativas, en el 30.7 % estas tuvieron un resultado positivo, en el 14 % de los casos no se realizó BK y en el 1.3 % no se cuenta con información de si se realizó esta prueba en el sistema de información. El 30.54 % de los casos cuenta con cultivo negativo para micobacterias, mientras que en el 41.3 % tuvieron un resultado positivo, no se realizó cultivo en el 22.5 % de los casos y en el 5.7 % no se cuenta con información de si se realizó esta prueba. Con relación a pruebas moleculares para micobacterias, se encuentra que en el 45.3 % de los casos cuenta con prueba molecular positiva, mientras que en el 15.1 % tuvieron un resultado negativo, no se realizó esta prueba al paciente en el 0.44 % de los casos y en el 39.2 % no se cuenta con información de si se realizó o no esta prueba en el sistema de información. Al respecto, es importante mencionar que a partir del año 2020 se modificaron los algoritmos diagnósticos para tuberculosis en nuestro país (Resolución 227 de 2020), debido a las limitantes en sensibilidad que representan las baciloscopias y los cultivos en medio sólido. A partir de ese año, se implementaron las pruebas moleculares y los cultivos en medio líquido, con el fin de disminuir los tiempos de diagnóstico de la enfermedad y garantizar tratamientos oportunos con una buena adherencia por parte de los pacientes.

Para la identificación de resistencias a antibióticos, se encuentra que la prueba de susceptibilidad a fármacos se le practicó al 48.2 % de los pacientes (n= 4.867). Es importante mencionar que esta prueba se priorizaba a personas que cumplían criterios específicos, por tal razón el porcentaje de no realización es del 51.8 %. No obstante, con

los nuevos algoritmos diagnósticos esta prueba se realiza de manera simultánea con la prueba molecular para la detección de genes de resistencia junto con el material genético de la bacteria. De las 4.867 pruebas de resistencia realizadas, no se encontró resistencia en el 48.5 % de los casos. De los tipos de resistencia más frecuentes en el distrito se encuentran: monorresistencia isoniacida 52 %, seguido de resistencia a rifampicina (RR) 32 %, multidrogorresistencia (MDR) 11.9 %, y en menor proporción se encuentran otro tipo de monorresistencias y polirresistencias a otros antibióticos.

Se encuentra que el 19.7 % de los pacientes presentan coinfección con VIH, mientras que en el 74.7 % no presentan coinfección y en el 5.6 % el estado de la coinfección es desconocido, esto puede deberse a que el paciente fallece o no accede a realizarse las pruebas colaborativas entre TB y VIH. De los pacientes con coinfección, se encuentra que el 77 % está recibiendo terapia antirretroviral (TAR), el 21.7 % no la recibe y en el 1.1 % no se cuenta con esta información.

En cuanto a las comorbilidades de los pacientes que ingresaron al programa de TB en los últimos 7 años, se encuentra que el 0.32 % de los pacientes refieren antecedente de alcoholismo (n= 33), el 5.6 % de los pacientes refieren presentar cáncer (n=562), el 0.22 % presenta antecedente de enfermedades cardiovasculares (n=22), el 1.23 % se reportan como consumidoras de sustancias psicoactivas (SPA) (n=124), el 0.99 % de los pacientes cursan con cuadro de COVID-19 de forma simultánea (n=100), el 17.4 % de los pacientes presentan desnutrición (n=1.760), en el 8.97 % se reportan con diabetes (n=906), el 0.11 % presentan enfermedades mentales (no se incluye adicciones) (n=11), el 2.52 % de los pacientes presentan enfermedades autoinmunes (n=255), en el 1.60 % presentan enfermedad hepática (n=162), en el 8.08 % reportan enfermedad renal (n=816), en el 10.71 % presentan Enfermedad Pulmonar Obstructiva Crónica (EPOC) (n=1.082), en el 1.36 % cursan con silicosis (n=137), en el 0.79 % se reporta tabaquismo (n=78), en el 10.76 % los pacientes presentan hipotiroidismo (n=1.087), en el 10.63 % los pacientes reportan otro tipo de comorbilidades infrecuentes (n=1.074). Es importante tener en cuenta que un mismo paciente puede presentar más de una comorbilidad asociada, adicional a la tuberculosis.

Con relación a la modalidad de tratamiento, esta variable se comenzó a diligenciar por parte del programa de TB desde el año 2020 (año pandémico dadas las dinámicas que esto generó a la atención de pacientes desde el sistema de salud); por esta razón el 55 % de los registros no cuenta con esta información. No obstante, para los últimos dos años el TDO en IPS representa el 34 % de las modalidades tratamiento seguido del TDO hospitalario 7.8 % y TDO virtual 1.33 %.

A su vez, en el 22.1 % de los casos los pacientes no recibieron ningún tipo de subsidio, seguido de no aplica a subsidio 20.9 %. De quienes si reciben algún tipo de apoyo el más frecuente es el subsidio alimentario 0.92 %, seguido de otros subsidios 0.5 % como transporte, vivienda y desempleo.

Por otro lado, las reacciones adversas al tratamiento tetraconjugado, se describen un total de 49, de las cuales las reacciones graves son las más frecuentes con 0.41 %, seguido de

moderadas 0.39 %. Lo que permite deducir que estos medicamentos son bastantes seguros ya que el 98.9 % de los pacientes que ingresaron al programa en los últimos 7 años, no reportaron ningún tipo de reacción adversa al tratamiento.

El 98.9 % de los pacientes diagnosticados se captaron a través de Búsqueda Activa Institucional (n=4.512). El 0.92 % por búsqueda activa derivada del trabajador de la salud (n= 42) y en menor proporción se reporta durante el estudio de contactos 0.08 % (n=4) y remitido por el Centro Nacional de Enlace (CNE) 0.02 % (n=1).

5.3 Análisis Bivariado

Se realizaron análisis bivariados teniendo en cuenta todas las condiciones de egreso de los pacientes en general (Anexo 6), así como un análisis solamente teniendo en cuenta la condición de egreso igual a pérdida de seguimiento (Anexo 7), en el link de Rpubs se encuentran los análisis estadísticos con sus respectivas gráficas.

Para el grupo de pérdida de seguimiento se encuentra un total de 739 pacientes entre 2016 y 2022 (n=10.102). De los cuales el 95.1% lograron ingresar al programa distrital de tuberculosis, es decir que iniciaron tratamiento.

Con relación a la variable sexo, en los pacientes con pérdida de seguimiento se evidencia que el 72.9 % corresponde a hombres, un porcentaje ligeramente mayor al compararlo con otras condiciones de egreso 65.7 %. Se evidenció significancia estadística de chi cuadrado entre el sexo y la pérdida de seguimiento (p valor =7.528e-05).

En cuanto a la edad, se evidencia que de forma general los pacientes enferman de TB en promedio sobre los 51 años con una mediana de 52 años; no obstante, para el grupo de pérdida de seguimiento, la mediana se ubica sobre los 36 años con una media de 41 años. Es decir, las personas más jóvenes y económicamente activas, quienes no tienen adherencia al tratamiento.

A su vez, frente al régimen de afiliación al SGSSS, se encuentra que las personas del régimen contributivo tienen mayor adherencia terapéutica con el 57.4 %, lo que demuestra que las EAPB realizan acciones de seguimiento a su población afiliada diagnosticada para generar adherencia terapéutica. De igual manera, las personas sin aseguramiento al SGSSS tienen un porcentaje mayor de pérdida de seguimiento 18.2 % respecto a las otras condiciones de egreso 5.7 %. En el régimen subsidiado es donde se concentra la mayor proporción de pacientes que no tienen adherencia al tratamiento de TB 46.7 %, respecto al total de condiciones de egreso 29.3 %. Se obtuvo significancia estadística de chi cuadrado entre el régimen de afiliación y la pérdida de seguimiento (p valor = < 2.2e-16).

Al analizar el tipo de tuberculosis se identifica que las formas pulmonares predominan en las pérdidas de seguimiento 76.8 %, respecto a las otras condiciones de egreso 68.1 %. dentro de las formas extrapulmonares, se evidencia diferencia para el grupo con pérdida de

seguimiento únicamente en la forma ganglionar 4.6 % frente a las demás condiciones de egreso 3.5 %. Se observó significancia estadística de chi cuadrado entre el tipo de tuberculosis y la pérdida de seguimiento (p valor = 9.308e-07).

En cuanto a la condición de ingreso al programa de TB, para los pacientes con pérdida de seguimiento, se encuentra que el 11.1 % corresponden a reingresos por pérdida de seguimiento frente a 1.4 % de otras condiciones de egreso; y el 6.5 % a otros previamente tratados versus 2.9 % de las otras condiciones de egreso, es decir que quienes ya han estado en tratamiento para TB y no tuvieron adherencia, tienen mayor probabilidad que vuelvan a tener pérdida de seguimiento en futuros tratamientos. Se evidenció significancia estadística de chi cuadrado entre la condición de ingreso y la pérdida de seguimiento (p valor < 2.2e-16).

Se identifica frente a la coinfección con VIH que los pacientes que presentan coinfección en el grupo de pérdida de seguimiento son mayores 32.4 %, que en el resto de las condiciones de egreso 18.8 %. Se encontró significancia estadística de chi cuadrado entre la coinfección con VIH y la pérdida de seguimiento (p valor < 2.2e-16).

Frente a las prueba diagnosticas de baciloscopia, cultivo y prueba molecular, se observa un porcentaje ligeramente mayor para positividad en estas pruebas que entre el grupo de pérdida de seguimiento y las otras condiciones de egreso (baciloscopia positiva 39.2 % vs 30.1 %, cultivo 45.8 % vs 40.9 %, prueba molecular 51.4 % vs 44.8 %). Se obtuvo significancia estadística de chi cuadrado entre la pérdida de seguimiento y el resultado de baciloscopia (p valor = 7.151e-08) y resultado de prueba molecular (p-valor = 0.01178).

En la prueba de susceptibilidad a fármacos se evidencian ligeras diferencias entre el tipo de prueba realizada al grupo de pérdida de seguimiento con LiPA (13.4 % vs 9.3 %) y PCR-TR (41.3 % vs 35.5 %) frente a pacientes con otras condiciones de egreso. Se observó significancia estadística de chi cuadrado entre la prueba de susceptibilidad a fármacos y la pérdida de seguimiento (p valor = 5.994e-07). Con relación a los tipos de farmacorresistencia, no se evidencia diferencia entre grupos. Es importante aclarar que los casos de TB farmacorresistente, el seguimiento se realiza en una base de datos diferente a la de TB sensible. Por tanto, las pérdidas de seguimiento de casos con TB resistente no son evaluados en el presente proyecto.

Dentro de las comorbilidades se evidenció mayor porcentaje para el grupo de pérdida de seguimiento frente a las otras condiciones de egreso en las siguientes: Consumo de SPA 6.6 % vs 0.8 %, Enfermedad mental 0.5 % vs 0.1 %, Desnutrición 22 % vs 17.1 %, Tabaquismo 1.6 % vs 0.7 %. No se encontró significancia estadística de chi cuadrado entre la pérdida de seguimiento y alcoholismo (p valor = 0.4655). Si se evidenció significancia estadística entre pérdida de seguimiento y las siguientes comorbilidades: consumo de SPA (p valor < 2.2e-16), desnutrición (p valor = 0.0009044), enfermedad mental (p valor = 0.001773), tabaquismo (p valor = 0.0129).

En cuanto a la modalidad del tratamiento directamente observado (TDO), el 36.7 % de los

pacientes que tienen pérdida de seguimiento lo realizan en la modalidad supervisión en IPS, un poco mayor frente a las demás condiciones de egreso 33.6 %. No se tuvo en cuenta para análisis bivariado dado que el diligenciamiento de esta variable inició en 2020.

Dentro de los programas de protección social se evidencia que las personas con pérdida de seguimiento tuvieron acceso a subsidio alimentario en 1.9 % mayor, frente al resto de condiciones de egreso 0.8 %, al igual que vivienda 0.5 % vs 0.3 % y otro tipo de subsidios 1.2 % vs 0.4 %. No se tuvo en cuenta para análisis bivariado dado que el diligenciamiento de esta variable inició en 2020.

En cuanto a la descripción de reacciones adversas de los pacientes al tratamiento tetraconjugado, se evidencia un mayor porcentaje de reacciones graves en el grupo con pérdida de seguimiento 0.4 % vs 0.2 % y moderadas 0.5%, frente a pacientes de otras condiciones de egreso 0.1 %. No se observó significancia estadística entre la pérdida de seguimiento y las reacciones adversas (p valor = 0.06234).

Frente a la metodología de captación, se evidencia una ligera diferencia porcentual para el grupo de pacientes con pérdida de seguimiento 46.9 % identificados por Búsqueda Activa Institucional (BAI) frente a las demás condiciones de egreso 44.5 %. No se tuvo en cuenta para análisis bivariado dado que el diligenciamiento de esta variable inició en 2020.

Dentro de la pertenencia étnica se identifica que los indígenas 3 % y afrocolombianos 1.6 %, presentan un mayor porcentaje en pérdida de seguimiento frente a las demás condiciones de egreso (1.5 % y 0.8 %, respectivamente). En los indígenas este porcentaje solamente se supera para el grupo de no evaluados 4.2 %, dado que regresan a sus comunidades. Mientras que, en los afrocolombianos se reporta un mayor porcentaje frente al grupo de excluidos por Farmacorresistencia y es igual el porcentaje frente al grupo de no evaluados y fracasos. Si se evidenció significancia estadística entre pérdida de seguimiento y pertenencia étnica (p valor = 0.002056).

Se evidencia un mayor porcentaje de pérdida de seguimiento en las siguientes poblaciones especiales frente a las demás condiciones de egreso: desplazados 2.2 % vs 0.7 %, migrantes 10.2 % vs 4.8 %, carcelarios 4.3 % vs 3.3 %, habitantes de calle 27.6 % vs 3.6 %, población ICBF 0.3 % vs 0.1 %, población psiquiátrica 0.5 % vs 0.1 %, víctima de la violencia de conflicto armado y gestante con 0.5 % vs 0.2 %, en igual proporción para ambos grupos. No se encontró significancia estadística de chi cuadrado entre la pérdida de seguimiento y los siguientes grupos poblacionales: discapacidad (p valor = 0.1464), carcelarios (p valor = 0.1528), gestante (p valor = 0.07948) y población a cargo del ICBF (p valor = 0.6238). Si se evidenció significancia estadística entre pérdida de seguimiento y los siguientes grupos poblacionales: desplazados (p valor = 2.414e-05), migrante (p valor = 3.786e-10), habitante de calle (p valor < 2.2e-16) y población psiquiátrica (p valor = 0.001773).

Dentro de la subred de residencia de los pacientes con pérdida de seguimiento, se evidencia una mayor pérdida de seguimiento para quienes residen en: Subred centro

oriente 24.5 % vs 18.2 % y Subred Sur 15 % vs 11.7 %, también es más alto en quienes no se cuenta con dato de residencia 10.2 % vs 2.9 %. Dentro de estas subredes se encuentran localidades donde se concentran en su mayoría habitantes de calle o población flotante. Se observó significancia estadística de chi cuadrado entre la localidad de residencia y la pérdida de seguimiento (p valor < 2.2e-16).

Este mismo comportamiento se observa al realizar el análisis por Subred de diagnóstico, en la subred centro oriente se evidencia una proporción más alta 44 % para las personas con pérdida de seguimiento vs otras condiciones de egreso 34.5 %, al igual que para la subred sur 9.5 % para el grupo de pérdida de seguimiento vs 5.8 % otras condiciones de egreso. Se encontró significancia estadística de chi cuadrado entre la Subred de diagnóstico y la pérdida de seguimiento (p valor = 4.691e-14). No obstante, la subred de diagnóstico no fue tenida en cuenta en el modelado, dado que para el distrito los seguimientos de los pacientes desde el área de salud pública se realizan por residencia.

5.4 Selección de predictores

Pérdida seguimiento vs Edad: La variable edad muestra diferencias para el grupo con pérdida de seguimiento con una mediana de 36.0 años y media de 40.8 años con una desviación estándar de 18 años versus una media de 51 años con una desviación estándar de 21.7 años y mediana de 52 años para el grupo No pérdida de seguimiento. Es decir, que las personas que no tienen adherencia terapéutica tienen entre 36 y 41 años, como se muestra en la Figura 10.

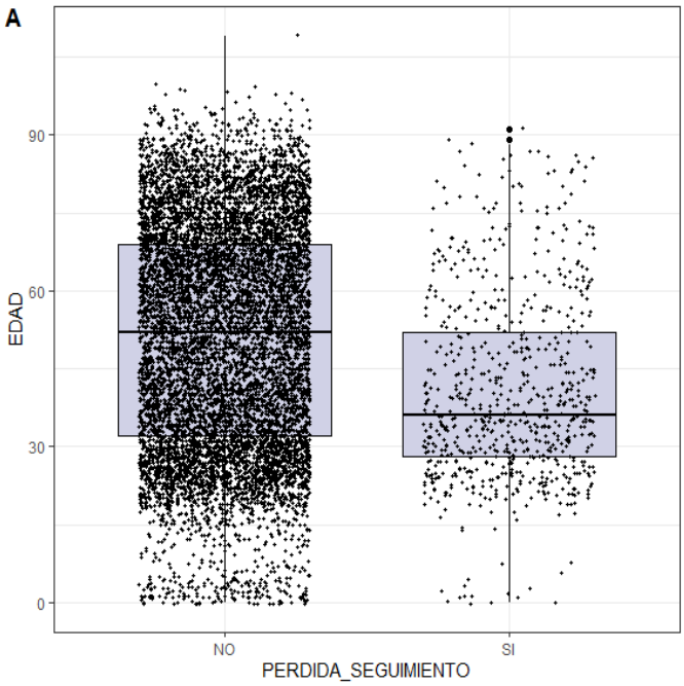

Figura 10. Comparativo de la edad en la pérdida de seguimiento (2016-2022) en el distrito capital (Fuente: construcción propia. Bases distritales del programa de TB 2016-2022)

Se aplicó prueba de Mann–Whitney–Wilcoxon (WMW), también conocido como Wilcoxon rank-sum se trata de un test no paramétrico, que contrasta si dos muestras proceden de poblaciones equidistribuidas. La hipótesis nula establece que no hay diferencia entre las medianas de los dos grupos mientras que la hipótesis alternativa, indica que sí existe una diferencia entre las medianas. Al aplicar la prueba a ambos grupos (con pérdida y sin pérdida de seguimiento) se encuentra que si hay diferencia entre ambos con significancia estadística (p valor < 2.2e-16), es decir que la mediana del grupo sin pérdida de seguimiento es mayor (52 años) respecto a la mediana del grupo con pérdida de seguimiento (36 años).

A continuación, en la
Tabla 10 se muestran los resultados de las variables categóricas que presentaron una proporción más alta para la condición de egreso con pérdida en el seguimiento que tuvieron significancia estadística con la prueba de chi cuadrado p (<=0.05) y su respectivo índice de Cramer.

Tabla 10. Variables con significancia estadística (Chi-cuadrado) para la pérdida de seguimiento de pacientes con TB (2016-2022) en el distrito capital (Fuente: construcción propia).

Variables contrastadas con pérdida de seguimiento	P-valor X2	Índice de Cramer
Régimen de afiliación	< 2.2 e-16	0.18
Tipo de Tuberculosis	< 2.2 e-16	0.19
Coinfección con VIH	< 2.2 e-16	0.10
Condición de ingreso	< 2.2 e-16	0.19
Comorbilidad Consumo de SPA	< 2.2 e-16	0.14
Grupo poblacional habitante de calle	< 2.2 e-16	0.28
Grupo poblacional migrante	3.786 e-10	0.06
Grupo poblacional desplazado	2.414 e-05	0.04
Sexo	7.528 e-05	0.04
Comorbilidad Desnutrición	0.0009	0.03
Pertenencia étnica	0.0020	0.04
Comorbilidad Tabaquismo	0.0129	0.03
Resultado baciloscopia	7.151 e-08	0.06
Subred residencia	< 2.2 e-16	0.13
Comorbilidad enfermedad mental	0.0020	0.04

Se identifica que las variables predictoras para la pérdida de seguimiento en pacientes del programa de TB distrital, que presentan significancia estadística se relacionan con población que presentan condiciones de vulnerabilidad. Ya sea por que pertenecen a un grupo poblacional especial: migrantes, desplazados, habitantes de calle, pertenencia étnica; quienes en su mayoría no cuentan con afiliación al sistema de seguridad social en salud (No afiliados o pertenecen al régimen subsidiado), se identifica que corresponden a hombres con edades entre los 36 y 41 años. El tipo de tuberculosis que predomina en la pérdida de seguimiento es la TB pulmonar que se relaciona con los resultados de baciloscopia, lo cual representa un problema a nivel de salud pública ya que al ser bacilífero el paciente logra transmitir la enfermedad a otros. Las principales comorbilidades asociadas a la pérdida de seguimiento son VIH, desnutrición, tabaquismo, enfermedad mental y consumo de SPA. Se evidencia que los pacientes que han tenido abandonos previos tienen

mayor probabilidad de no tener adherencia terapéutica. En cuanto a la residencia de los usuarios, se identificó que la mayoría residen en las localidades de la Subred Centro Oriente y Sur, zonas de alta vulnerabilidad a nivel socioeconómico y que puede estar relacionado con estados de mal nutrición y zonas de consumo y expendio de sustancias psicoactivas.

5.5 Preprocesamiento del conjunto de datos

Para esta fase del proyecto, se utilizó el software Python a través del uso de la plataforma Google Colab recursos básicos. Dado que se cuenta con una variable numérica y las demás categóricas, incluyendo la variable objetivo, se procedió a realizar la conversión de las variables categóricas a numéricas, para ello se separaron del conjunto de datos y se recodificaron usando la función OneHotEncoder de la librería sklearn.preprocessing, generando nuevas columnas por cada categoría de la variable. Por otro lado, la variable objetivo se trabajó de forma binaria como pérdida de seguimiento Si=1 y No=0, en una misma columna. La variable numérica edad fue normalizada utilizando la función StandardScaler de la librería con el mismo nombre. Posteriormente, se unificaron las variables en el mismo conjunto de datos obteniendo un total de 50 atributos que contienen la información de las 17 variables predictoras y 10.102 observaciones.

El conjunto de datos para realizar las fases de modelado y validación quedó conformado por las 17 variables que mostraron significancia estadística respecto a la pérdida de seguimiento, de las cuales muchas contienen varias categorías como se describe en la Tabla 11.

Tabla 11. Variables utilizadas en el modelamiento (Fuente: construcción propia).

Variable	Nombre variable Dataset	Tipo de variable	Valores variable
Sexo paciente	Sexo	Categórica	F=Femenino M=Masculino
Edad del paciente en años	Edad	Numérica continúa	0-107
Pertenencia étnica	Pertenencia_etnica	Categórica	"Indigena", "Afro", "Otro", "Palenquero", "Raizal", "Room"
Grupo poblacional desplazado	gp_desplaz	Categórica	No Si
Grupo poblacional migrante	gp_migrant	Categórica	No Si
Grupo poblacional habitante de calle	gp_indigen	Categórica	No Si
Localidad de residencia del paciente	Loc_res	Categórica	"Sin Dato", "NORTE", "CO", "SO", "SUR" "FDB"
Régimen de afiliación al SGSSS	Regimen_afiliacion	Categórica	"C", "E", "N", "P", "S"
Tipo de tuberculosis	Tipo_TB	Categórica	"extrapulmonar", "pulmonar"
Condición de ingreso al programa	Condicion_ingreso	Categórica	"Nuevo", "OPT", "RTF", "RTPS", "RTR", "remitido"
Resultado de baciloscopia	Resultado_BK_recod	Categórica	"negativo", "NR", "positivo", "SD"
Condición de coinfección con VIH	Condicion_VIH	Categórica	"desconocido", "negativo", "positivo"
Consumo de sustancias psicoactivas	Consumidor_SPA	Categórica	No Si

Variable	Nombre variable Dataset	Tipo de variable	Valores variable
Comorbilidad Desnutrición	Desnutricion	Categórica	No Si
Comorbilidad Tabaquismo	Tabaquismo	Categórica	Si No
Comorbilidad Enfermedad mental	Enf_Mental	Categórica	No SI
Pérdida del seguimiento por parte del paciente	Pérdida_Seguimiento	Categórica- Variable objetivo	No Si

6. ENTRENAMIENTO DE MODELOS DE MACHINE LEARNING

A continuación, se describen las fases realizadas al conjunto de datos para el proceso de modelamiento.

6.1 Partición del conjunto de datos

Se procedió a particionar el conjunto de datos en tres partes: entrenamiento, pruebas y validación, utilizando la función train_test_split de la librería sklearn.model_selection. La partición se generó de la siguiente forma: 60% entrenamiento, 20% pruebas y 20% validación y quedaron conformados como se observa en la Tabla 12.

Tabla 12. Comparativo de la partición del conjunto de datos: entrenamiento, pruebas y validación. (Fuente: construcción propia)

Conjunto de datos	% partición	Clase 0	Clase 1
Original	100	9.364	738
Entrenamiento	60	5.618	443
Prueba	20	1.873	148
Validación	20	1.873	147

6.2 Reducción de variables predictoras

Inicialmente se obtuvo 17 variables predictoras para la pérdida de seguimiento de pacientes a partir de la aplicación de la prueba de Chi cuadrado; con el fin de determinar si existía multicolinealidad entre las mismas, se realizó un comparativo del estadístico Chi-cuadrado entre ellas (exceptuando la variable edad) generando una matriz con estos valores calculados, como se observa en la Figura 11.

Se aprecia en esta figura que la variable sexo tiene significancia con la mayoría de las variables excepto grupo migrante, grupo desplazado y tipo de TB. La pertenencia étnica presenta significancia estadística para la mayoría de las variables excepto grupo migrante, tipo de TB, resultado de baciloscopia, comorbilidades tabaquismo y enfermedad mental. El grupo de desplazado tiene significancia estadística con pertenencia étnica, localidad de residencia, régimen de afiliación, tipo de TB y pérdida de seguimiento. El grupo migrante no presenta significancia estadística con las siguientes variables: sexo, pertenencia étnica, grupo desplazado, condición de ingreso, consumo de SPA y enfermedad mental. El grupo habitante de calle presenta significancia con la mayoría de las variables excepto grupo desplazado, tabaquismo y enfermedad mental. La localidad de residencia presenta significancia estadística con la mayoría de las variables excepto enfermedad mental. El régimen de afiliación presenta significancia estadística con todas las variables excepto tabaquismo. El tipo de tuberculosis presenta significancia estadística con la mayoría de las variables a excepción de Sexo, pertenencia étnica, tabaquismo y enfermedad mental. La condición de ingreso al programa presenta significancia estadística excepto con grupo desplazados, grupo migrantes, tabaquismo y enfermedad mental. El resultado de

baciloscopia (BK) presenta significancia estadística con las demás variables excepto por pertenencia étnica, grupo desplazado, tabaquismo y enfermedad mental. La condición de VIH presenta significancia estadística con la mayoría de las variables exceptuando el grupo de desplazados. El consumo de SPA no presenta significancia estadística con las variables grupo desplazado y migrante, con el resto de las variables si presenta significancia. La comorbilidad desnutrición presenta significancia estadística con la mayoría de las variables excepto por grupo desplazado, tabaquismo y enfermedad mental. La comorbilidad tabaquismo solamente presenta significancia estadística con sexo, grupo migrante, localidad de residencia, condición VIH, consumo de SPA y pérdida de seguimiento. A su vez, la comorbilidad enfermedad mental solo presenta significancia estadística con las variables: régimen de afiliación, condición VIH, consumo de SPA y pérdida de seguimiento. Por su parte, la variable pérdida de seguimiento presenta significancia estadística con las demás variables, confirmado los resultados del análisis bivariado realizado.

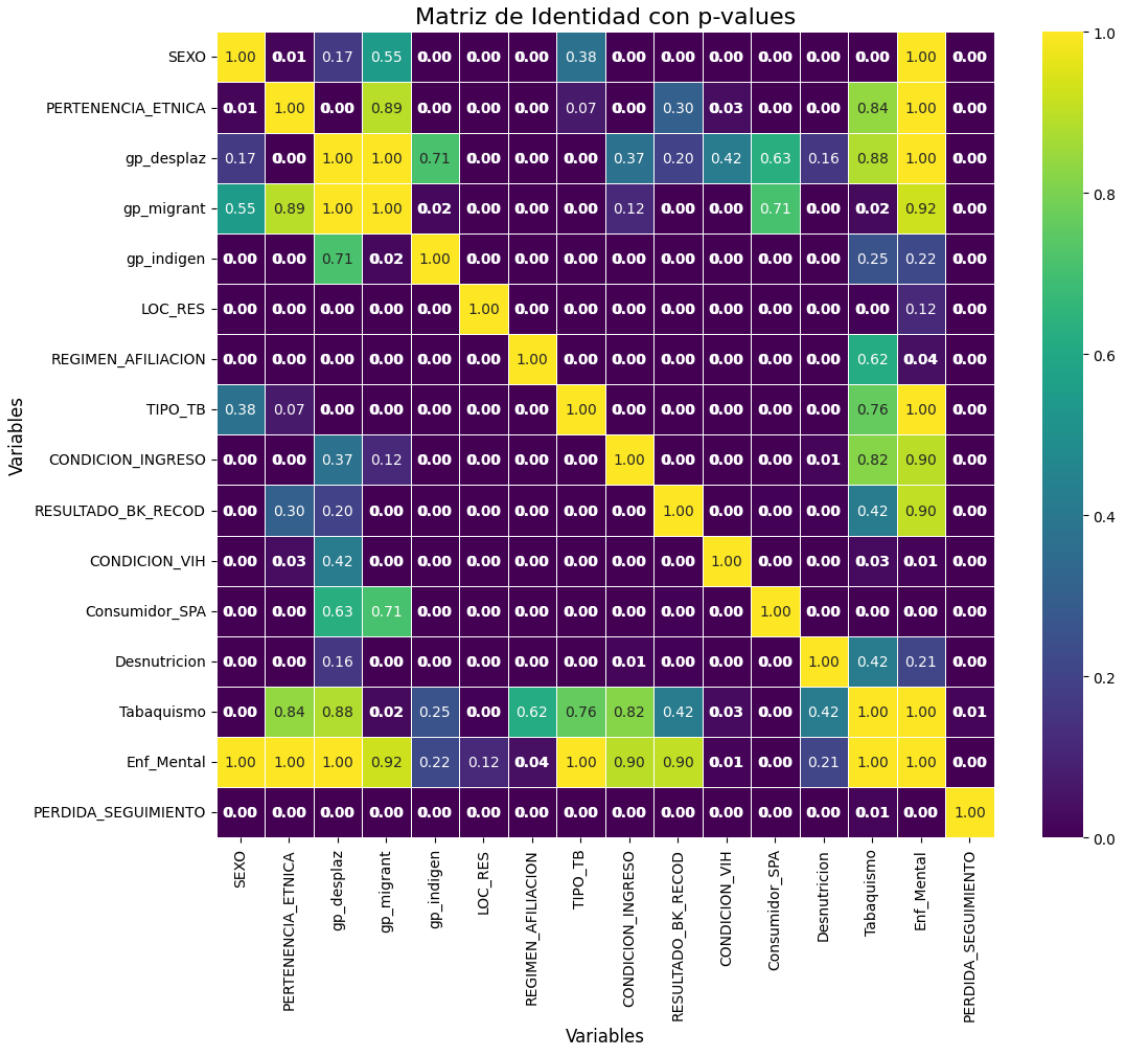

Figura 11. Matriz con p-valor de chi cuadrado variables categóricas. (Fuente: construcción propia)

Una vez discretizadas estas 17 variables (dummy), se obtuvo un total de 49 atributos o características, por lo cual se hizo necesario aplicar técnicas para reducir la cantidad de variables a utilizar en la fase de modelado. Para ello se aplicaron las siguientes técnicas de reducción de atributos; el script de los modelos generados se encuentra disponibles para su consulta en el repositorio de GitHub en el siguiente link: Features.ipynb

Se aplicó la reducción de variables por el método Group LASSO utilizando la librería Celer [52] dado no está disponible en la implementación de scikit-learn. Se generaron varios modelos Group LASSO variando el coeficiente alfa [0.001, 0.01, 0.1, 1, 10, 50, 100]. Con el valor de 0.001 se logró reducir a 38 atributos y con el valor de 0.01 se redujeron a 8 atributos como se muestra en la Tabla 13,mientras que con los otros valores alfa indicados se penalizó la totalidad de variables.

Aplicando el modelo base de regresión logística estableciendo un valor de p significante por debajo o igual a 0.05, se tuvo inicialmente una reducción de 49 a 31 características como se observa en la Tabla 13. Al incorporar variante con selección Forward se obtuvo un total de 14 características como se evidencia en la Tabla 13, mientras que con la incorporación Backward se obtuvo un total de 15 características seleccionadas como se observa en la misma tabla. A su vez, la selección por pasos (Stepwise) se obtuvo 3 características, como se muestra en la Tabla 13.

En cuanto a los resultados con el método de RFE se obtuvo un total de 12 características con un p valor significante menor o igual a 0.05, como se muestra en la Tabla 13.

Tabla 13. Comparativo variables relevantes por cada método empleado (Fuente: construcción propia).

Variables	Group LASSO 0.001:	Group LASSO 0.01	LOGIT	LOGIT + forward	LOGIT + backward	LOGIT + stepwise	RFE
Edad'	X	X	X	X	X		X
Sexo F	X		X				
Sexo M	X			X			
Pertenencia Étnica Afro	X						
Pertenencia Étnica Indígena	X					X	
Pertenencia Étnica Otro	X		X				
Pertenencia Étnica Palenquero							
Pertenencia Étnica Raizal	X						
Pertenencia Étnica Room (Gitano)	X					X	
Gp desplaz NO			X		X		X
Gp desplaz SI			X	X			

Variables	Group LASSO 0.001:	Group LASSO 0.01	LOGIT	LOGIT + forward	LOGIT + backward	LOGIT + stepwise	RFE
Gp migrant NO			X				
Gp migrant SI							
Gp indigen NO	X	X	X				
Gp indigen SI	X	X	X	X	X		X
Loc Res CO	X		X				
Loc Res FDB	X		X	X	X		X
Loc Res Norte	X		X				
Loc Res SO	X						
Loc Res Sur	X		X	X	X		
Loc Res Sin Dato	X		X				
Régimen Afiliación C	X	X	X	X			
Régimen Afiliación E	X	X					
Régimen Afiliación N	X	X	X	X	X	X	X
Régimen Afiliación P	X	X	X				
Régimen Afiliación S	X	X	X	X	X		X
Tipo TB Extrapulmonar	X		X				
Tipo TB Pulmonar	X						
Condición Ingreso Nuevo	X		X	X			X
Condición Ingreso OPT	X				X		
Condición Ingreso Remitido	X						
Condición Ingreso RTF	X				X		
Condición Ingreso RTPS	X		X	X	X		X
Condición Ingreso RTR	X				X		
Resultado BK Negativo	X						
Resultado BK No realizado	X		X				
Resultado BK Positivo	X		X				
Resultado BK SD	X		X				
Condición VIH Desconocido	X		X	X			X
Condición VIH Negativo	X		X		X		
Condición VIH Positivo	X				X		
Consumidor SPA NO	X		X	X	X		X
Consumidor SPA SI	X		X				
Desnutrición NO							

Variables	Group LASSO 0.001:	Group LASSO 0.01	LOGIT	LOGIT + forward	LOGIT + backward	LOGIT + stepwise	RFE
Desnutrición SI			X				
Tabaquismo NO			X				
Tabaquismo SI			X				
Enfermedad Mental NO			X	X	X		X
Enfermedad Mental SI							X
TOTAL	**38**	**8**	**31**	**14**	**15**	**3**	**12**

Para determinar el método más adecuado en cuanto al grupo de variables a seleccionar, se tuvo en cuenta los valores de: Criterio de Información Bayesiano (BIC) y el Criterio de Información de Akaike (AIC) por cada uno de los modelos, como se observa en la Tabla 14.

Tabla 14. Comparativo BIC y AIC por cada método empleado (Fuente: construcción propia).

Estimador	LASSO 0.001:	LASSO 0.01	LOGIT	LOGIT + forward	LOGIT + backward	LOGIT + stepwise	RFE
AIC	300.14	300.14	3232.22	4533.35	4535.69	3094.16	4541.36
BIC	533.51	533.51	3424.41	4641.66	4651.22	3123.04	4635.23

De esta manera, se establece que el grupo de variables a seleccionar corresponde al método de Group Lasso con valores de alfa 0.001 y 0.01, los cuales presentan los menores valores de los estimadores tanto para BIC como AIC. Se decide modelar con ambos valores de alfa para determinar el mejor desempeño de las métricas.

6.3 Aplicación de modelos de aprendizaje automático

El aprendizaje automático ofrece un enfoque más que se adapta a las particularidades de esta enfermedad y permite emplear métricas personalizadas las cuales se adaptan mejor a las necesidades y limitaciones específicas del manejo de la TB [42]. En esta fase se aplicaron 4 tipos de algoritmos utilizados ampliamente en problemas de clasificación binaria como en este caso (pérdida de seguimiento Si y No): Bosques Aleatorios, Regresión logística, XGBoost y Naive Bayes; escogidos acorde con la revisión realizada en el marco teórico, que permitió resolver problemas de clasificación al mostrar buenos desempeños para discriminar de manera correcta las clases. Adicionalmente, estos modelos se reconocen por su capacidad para manejar diversos tipos de datos, su robustez frente al sobreajuste y su alto rendimiento en términos de velocidad y precisión [42].

Dado que existe un marcado desbalanceo de clases en el conjunto de datos generado por que la variable objetivo contiene el 7.3% (n=738 casos) con pérdida de seguimiento y 92.7% (n= 9.364) sin pérdida de seguimiento; se hizo necesario aplicar técnicas de remuestreo a cada conjunto de entrenamiento con el fin de equilibrarlas. Cada una de estas técnicas se

aplicó al conjunto de datos de entrenamiento de cada uno de los modelos de aprendizaje supervisado descritos en el numeral 6.5, comparando las métricas obtenidas con un modelo sin ningún tipo de remuestreo, con el fin de generar una línea base de comparación que permitió seleccionar el modelo con mejor desempeño. En cada script por tipo de modelo se aprecia cada una de las técnicas utilizadas.

Por otro lado, se definió utilizar la sensibilidad o *Recall* como punto de referencia con el fin de priorizar la correcta clasificación de la clase positiva pérdidas de seguimiento. Es decir, la proporción de verdaderos positivos sobre total de verdaderos positivos y Falsos negativos. De esta manera, el *Recall* resume qué tan bien se predijo la clase positiva respecto a las predicciones [53] hechas por el modelo. Esto es importante dado que, para la fase de despliegue del modelo, se pretende realizar un seguimiento personalizado a los pacientes desde su ingreso al programa de TB, por lo cual es importante que la métrica utilizada discrimine muy bien a este tipo de pacientes en riesgo de pérdida de seguimiento de tal manera que se logre optimizar de la mejor forma al personal de salud de los equipos locales encargados de esta actividad.

En cada modelo generado se realizó optimización de hiperparámetros utilizando la función GridsearchCV y se hizo validación cruzada con la función Cross_val_score ambas pertenecientes a la librería sklearn de Pyhton, esta última permite realizar una validación cruzada a partir de la partición del conjunto de datos mientras que la primera permite encontrar los parámetros óptimos en cada modelo, acorde con los valores para *Recall* según la combinación de éstos [54].

Los tiempos de cómputo de todos los modelos generados fueron medidos en un equipo que cuenta con las siguientes especificaciones técnicas: Core i7 octava generación, 16 RAM, Disco duro de estado sólido de 1 Tera con acceso a canal de internet dedicado ETB 10 MB.

6.3.1 Bosques Aleatorios

De acuerdo con la revisión realizada sobre remuestreo, se procedió a generar diferentes modelos aplicando cada una de estas técnicas y de forma combinada, y posteriormente, se compararon las métricas obtenidas con el modelo sin ningún tipo de remuestreo obteniendo los resultados de la Tabla 15. El script de los modelos generados se encuentra disponibles para su consulta en el repositorio de GitHub en el siguiente link: Random Forest models.ipynb

Tabla 15. Comparativo de métricas en modelos de Random Forest aplicando diferentes técnicas de remuestreo. (Fuente: construcción propia).

Random Forest con	Recall	Precision	F1 Score	Accuracy	AUC
Submuestreo (Undersampling)	0.675676	0.182149	0.286944	0.754082	0.717977

Random Forest con	Recall	Precision	F1 Score	Accuracy	AUC
Balanceo de clases (Class weights)	0.594595	0.221662	0.322936	0.817417	0.714809
Sobremuestreo (Oversampling)	0.581081	0.212871	0.311594	0.811974	0.705650
SMOTE + Tomek	0.547297	0.217158	0.310940	0.822365	0.695699
SMOTE Sobremuestreo (Oversampling)	0.527027	0.222857	0.313253	0.830777	0.690903
Tomek submuestreo (Undersampling)	0.135135	0.416667	0.204082	0.922823	0.560094
Ninguno	0.094595	0.437500	0.155556	0.924790	0.542492

Los hiperparámetros optimizados para estos modelos fueron los siguientes: número de árboles del bosque (n_estimators), la profundidad máxima del árbol (max_depth) y aleatoriedad del muestreo buscando la mejor división en cada nodo (random_state). La combinación de los mejores hiperparámetros para cada uno de los modelos, se encuentran en el Script de Github.

En cuanto a la correcta clasificación de las clases de los modelos Random Forest, se describe en la Tabla 16, los resultados de las matrices de confusión generadas con el conjunto de prueba (contiene 1873 observaciones de la clase 0 no perdida de seguimiento y 148 de la clase 1 pérdida de seguimiento); se evidencia que el modelo de Random Forest que presenta una mejor identificación de las pérdidas de seguimiento es submuestreo, sin embargo presenta un alto número de falsos positivos, seguido de remuestreo con Balanceo de clases que logra identificar 88 verdaderos positivos con un alto número de falsos positivos 309.

Tabla 16. Comparativo de matriz de confusión en modelos de Random Forest (Fuente: construcción propia).

Random Forest con	Verdaderos positivos	Verdaderos negativos	Falsos positivos	Falsos negativos
Submuestreo (Undersampling)	100	1424	449	48
Balanceo de clases (Class weights)	88	1564	309	60
Sobremuestreo (Oversampling)	86	1555	318	62
SMOTE + Tomek	81	1581	292	67
SMOTE Sobremuestreo (Oversampling)	78	1601	272	70
Tomek submuestreo (Undersampling)	15	1384	21	96
Ninguno	14	1855	18	134

También se midió el tiempo de cómputo (segundos) en la generación de cada modelo el cual se muestra de forma comparativa en la Tabla 17, acorde con las especificaciones ya

descritas.

Tabla 17. Comparativo de tiempos computo en ejecución en modelos de Random Forest. (Fuente: construcción propia).

Modelo	num_estimadores	Mejor_puntaje	Tiempo (s)
Balanceo de clases (Class weights)	50	0.636517	24.07
Ninguno	200	0.094867	30.40
Tomek submuestreo (Undersampling)	50	0.539581	31.07
Submuestreo (Undersampling)	50	0.548570	50.76
SMOTE Sobremuestreo (Oversampling)	50	0.143092	56.14
Sobremuestreo (Oversampling)	50	0.516982	61.38
SMOTE + Tomek	100	0.521527	108.90

6.3.2 Regresión Logística

Se generaron varios modelos de Regresión Logística uno base y otros aplicando diferentes técnicas de remuestreo, obteniendo las métricas descritas en la Tabla 18. El script de los modelos generados se encuentra disponibles para su consulta en el repositorio de GitHub en el siguiente link: LOGIT.ipynb

Tabla 18. Comparativo de métricas en modelos de Regresión Logística aplicando diferentes técnicas de remuestreo. (Fuente: construcción propia).

Logit con	Recall	Precision	F1 Score	Accuracy	AUC
Sobremuestreo (Oversampling)	0.689189	0.155251	0.253416	0.702622	0.696437
SMOTE Sobremuestreo (Oversampling)	0.682432	0.150746	0.246944	0.695200	0.689321
submuestreo (Undersampling)	0.682432	0.156347	0.254408	0.707076	0.695728
SMOTE + Tomek	0.675676	0.148148	0.243013	0.691737	0.684341
Tomek submuestreo (Undersampling)	0.108108	0.533333	0.179775	0.927759	0.550317
Ninguno	0.081081	0.545455	0.141176	0.927759	0.537871

Los hiperparámetros optimizados para estos modelos fueron los siguientes: Inverso de la fuerza de regularización (C) debe ser un número flotante positivo, algoritmo a utilizar en el problema de optimización (solver) y penalización (penalty). La combinación de los mejores hiperparámetros para cada uno de los modelos, se encuentran en el Script de Github.

En cuanto a la correcta clasificación de las clases de los modelos de regresión logística, se describen en la Tabla 19 los resultados de las matrices de confusión generadas con el conjunto de prueba (contiene 1873 observaciones de la clase 0 no perdida de seguimiento

y 148 de la clase 1 pérdida de seguimiento); se evidencia que el modelo de Regresión logística que presenta una mejor identificación de las pérdidas de seguimiento es sobremuestreo con identificación correcta de 102 casos del total de 148 pérdidas de seguimiento; sin embargo presenta un alto número de falsos positivos (555), los otros modelos que muestran buen desempeño son SMOTE con sobremuestreo y Submuestreo ambos clasifican correctamente 101 perdidas de seguimiento de las 148 pero presentan un alto número de falsos positivos 569 y 545, respectivamente.

Tabla 19. Comparativo de matriz de confusión en modelos de Regresión logística (Fuente: construcción propia).

Logit con	Verdaderos positivos	Verdaderos negativos	Falsos positivos	Falsos negativos
Sobremuestreo (Oversampling)	102	1318	555	46
SMOTE Sobremuestreo (Oversampling)	101	1304	569	47
Submuestreo (Undersampling)	101	1328	545	47
SMOTE + Tomek	100	1298	575	48
Tomek submuestreo (Undersampling)	16	1859	14	132
Ninguno	12	1863	136	10

Con estos modelos también se midió el tiempo de cómputo (segundos) en la generación de cada modelo el cual se muestra de forma comparativa en la Tabla 20, acorde con las especificaciones ya descritas.

Tabla 20. Comparativo de tiempos computo en ejecución en modelos de Regresión Logística. (Fuente: construcción propia).

Modelo	C_regresión logistica	Mejor_puntaje	Tiempo (s)
Ninguno	1.000	0.081231	1.01
Submuestreo (Undersampling)	0.001	0.688788	42.40
Tomek submuestreo (Undersampling)	100.000	0.110606	55.11
SMOTE + Tomek	0.001	0.695707	78.86
Sobremuestreo (Oversampling)	0.001	0.684394	84.47
SMOTE Sobremuestreo (Oversampling)	0.001	0.695707	372.17

6.3.3 XGBoost

De igual manera, se generaron varios modelos XGBoost uno de línea base, otro favoreciendo el peso de clase positiva (pérdida de seguimiento), los otros con técnicas de remuestreo over y under sampling, obteniendo las métricas descritas en la Tabla 21. El script de los modelos generados se encuentra disponibles para su consulta en el repositorio de GitHub en el siguiente link: XGBOOST.ipynb

Tabla 21. Métricas del modelo XGBoost con diferentes técnicas de remuestreo. (Fuente: construcción propia).

XGBoost con	Recall	Precision	F1 Score	Accuracy	AUC
Submuestreo (Undersampling)	0.716216	0.144809	0.240909	0.669471	0.690997
Balanceo de clases (Class weights)	0.655405	0.139769	0.230404	0.679367	0.668333
Sobremuestreo (Oversampling)	0.254237	0.214900	0.232919	0.877753	0.590546
Ninguno	0.167421	0.349057	0.226300	0.916529	0.571433
SMOTE + Submuestreo (Undersampling)	0.159322	0.251337	0.195021	0.903984	0.560974

Los hiperparámetros optimizados para estos modelos fueron los siguientes: La tasa de aprendizaje reduce la contribución de cada árbol mediante *learning_rate*. Existe un equilibrio entre *learning_rate* y *n_estimators*. La profundidad máxima de los estimadores de regresión individuales (max_depth). Esta profundidad limita el número de nodos en el árbol. Ajusta este parámetro para obtener el mejor rendimiento, ya que el valor óptimo depende de la interacción entre las variables de entrada. El número de etapas de refuerzo (*boosting*) a realizar (n_estimators). El *gradient boosting* es bastante resistente al sobreajuste, por lo que un número grande de estimadores suele mejorar el rendimiento. El parámetro submuestreo por árbol subsample (colsample_bytree) es la proporción de columnas que se utiliza al construir cada árbol. El muestreo aleatorio (*subsampling*) ocurre una vez por cada árbol que se construye. La combinación de los mejores hiperparámetros para cada uno de los modelos, se encuentran en el Script de Github.

En cuanto a la correcta clasificación de las clases de los modelos de XGBoost, se describen en la Tabla 22 los resultados de las matrices de confusión generadas con el conjunto de prueba (contiene 1873 observaciones de la clase 0 no perdida de seguimiento y 148 de la clase 1 pérdida de seguimiento); se evidencia que el modelo XGBoost que presenta una mejor identificación de las pérdidas de seguimiento es submuestreo con identificación correcta de 106 casos del total de 148 pérdidas de seguimiento; sin embargo presenta un alto número de falsos positivos (626) pero bajo número de falsos negativos, el siguiente modelo es con balanceo de clases el cual clasifica correctamente 97 pérdidas de seguimiento de las 148 pero presenta un alto número de falsos positivos 597 y bajo de falsos negativos.

Tabla 22. Comparativo de matriz de confusión en modelos de XGBoost (Fuente: construcción propia).

XGBoost con	Verdaderos positivos	Verdaderos negativos	Falsos positivos	Falsos negativos
Submuestreo (Undersampling)	106	1247	626	42
Balanceo de clases (Class weights)	97	1276	597	51
Sobremuestreo (Oversampling)	75	3472	274	220
Ninguno	37	2741	69	184

SMOTE + Submuestreo (Undersampling)	47	3606	140	248

Para estos modelos también se midió el tiempo de cómputo (segundos) en la generación de cada modelo el cual se muestra de forma comparativa en la Tabla 23, acorde con las especificaciones ya descritas.

Tabla 23. Comparativo de tiempos computo en ejecución en modelos de XGBoost (Fuente: construcción propia).

Modelo	num_estimadores	Mejor_puntaje	Tiempo (s)
Ninguno	200	0.125641	0.00
XG Submuestreo (Undersampling)	200	0.688458	0.18
XG Sobremuestreo (Oversampling)	200	0.996974	175.29
XG SMOTE + Submuestreo (Undersampling)	200	0.920971	496.80

6.3.4 Naive Bayes

Se generaron varios modelos de tipo bayesiano uno de línea base, otro favoreciendo el peso de clase positiva (pérdida de seguimiento), los otros con técnicas de sobremuestreo y submuestreo, obteniendo las métricas descritas en la Tabla 24. El script de los modelos generados se encuentra disponibles para su consulta en el repositorio de GitHub en el siguiente link: Bayesiano.ipynb

Tabla 24. Métricas de modelos bayesianos con diferentes técnicas de remuestreo. (Fuente: construcción propia).

Naive Bayes con	Recall	Precision	F1 Score	Accuracy	AUC
Balanceo de clases (Class weights)	0.972973	0.081172	0.149844	0.191489	0.551356
Sobremuestreo (Oversampling)	0.972973	0.083045	0.153029	0.211282	0.562034
SMOTE	0.972973	0.077670	0.143856	0.151905	0.530000
Ninguno	0.898649	0.094527	0.171061	0.362197	0.609228
Submuestreo (Undersampling)	0.364865	0.284211	0.319527	0.886195	0.646127

Para los modelos Gaussian Naive Bayes la optimización de hiperparámetros son estimados utilizando el máximo de verosimilitud por el mismo modelo, por lo cual no se optimiza a través de *GridSearchCV,* como se observa en el script.

En cuanto a la correcta clasificación de las clases de los modelos de Naive Bayes, se describen en la Tabla 25 los resultados de las matrices de confusión generadas con el conjunto de prueba (contiene 1873 observaciones de la clase 0 no perdida de seguimiento y 148 de la clase 1 pérdida de seguimiento); se evidencia que estos modelos aunque

identifican correctamente 144 casos como perdidas de seguimiento de los 148, presentan un alto número de falsos positivos por lo cual no pueden ser tenidos en cuenta dentro del modelamiento.

Tabla 25. Comparativo de matriz de confusión en modelos de Naive Bayes (Fuente: construcción propia).

Naive Bayes con	Verdaderos positivos	Verdaderos negativos	Falsos positivos	Falsos negativos
Balanceo de clases (Class weights)	144	243	1630	4
Sobremuestreo (Oversampling)	144	283	1590	4
SMOTE	144	163	1710	4
Ninguno	133	599	1274	15
Submuestreo (Undersampling)	54	1737	136	94

También se midió el tiempo de cómputo (segundos) en la generación de cada modelo el cual se muestra de forma comparativa en la Tabla 26, acorde con las especificaciones ya descritas.

Tabla 26. Comparativo de tiempos computo en ejecución en modelos Bayesianos (Fuente: construcción propia).

Modelo	Mejor_puntaje	Tiempo (s)
NB Ninguno	0.820253	0.00
NB Submuestreo (Undersampling)	0.820253	0.01
NB Sobremuestreo (Oversampling)	0.908125	0.02
Balanceo de clases (Class weights)	0.960124	0.02
NB + SMOTE	0.340606	0.05

En general, el algoritmo con mejor desempeño es Regresión Logística que muestra métricas de Recall entre 67 y 69 % con AUC entre 68 y 69 %, aplicando diferentes técnicas de remuestreo. De los modelos de XGBoost solamente dos presentan métricas aceptables entre 66 % y 72 % de Recall, priorizando peso de clases y con submuestreo, respectivamente. En cuanto al algoritmo de Random Forest solamente uno tiene buen desempeño y es con submuestreo 67.5 %. Mientras que los modelos bayesianos presentan en general buenas métricas de Recall entre 89 % y 98 % aplicando diferentes técnicas de remuestreo; no obstante, la exactitud desmejora entre 15 % y 21%; al verificar los resultados de las matrices de confusión Tabla 25, se encuentra que la mayoría de los registros se clasifican como falsos positivos, lo cual explica estos resultados.

En cuanto a los tiempos de cómputo de los modelos de bosques aleatorios se ejecutan con optimización de hipeparámetros entre 24 y 109 segundos, los modelos de regresión logística se ejecutan entre 1 y 372 segundos incluyendo optimización de hiperparámetros y validación cruzada con 10 folds; los modelos de XGBoost se ejecutan entre 0.00 y 497

segundos e incluye la optimización y validación cruzada con 10 folds. A su vez, los modelos bayesianos presentan los menores tiempos de ejecución, menor a un segundo e incluye optimización de hiperparámetros y validación cruzada con 10 folds.

De los modelos implementados se encuentra que la mejor métrica de Recall se obtiene con XGBoost con submuestreo 71.6 % (exactitud 66.9 %) el Modelo de Regresión logística con SMOTE + Tomek 70 % (exactitud 67 %), seguido de varios modelos de regresión logística: sobremuestreo 69 % de Recall (exactitud 71 %), SMOTE sobremuestreo 68 % Recall (exactitud 69.5 %), con submuestreo 68 % (exactitud 71 %); y tercer lugar Random Forest con submuestreo 67 % (exactitud 75 %). Los tres presentan tiempos de ejecución de 0.17 segundos (XGBoost con submuestreo), los modelos de regresión logística presentan tiempos de 42, 78, 84 y 372 segundos (submuestreo, SMOTE + Tomek, submuestreo y SMOTE + sobremuestreo) y 50.76 segundos (Random Forest con submuestreo).

Posteriormente, se decide realizar el modelamiento con los algoritmos seleccionados que mostraron mejor desempeño en la métrica Recall, aplicando el conjunto de datos de entrenamiento y prueba como se muestra en la Tabla 27. Adicional, se realiza medición de tiempos de cómputo de estos modelos como se visualiza en Tabla 28

Tabla 27. Comparativo de métricas en modelos finales con remuestreo. (Fuente: construcción propia).

Modelo	Recall	Precision	F1 Score	Accuracy	AUC
XGBoost con submuestreo (Undersampling)	0.714932	0.126097	0.214383	0.617948	0.662626
Random Forest con submuestreo (Undersampling)	0.687783	0.137931	0.229781	0.663807	0.674852
Logit con sobremuestreo (Oversampling)	0.674208	0.150810	0.246485	0.699439	0.687816
Logit con SMOTE + Tomek	0.669683	0.146245	0.240065	0.690861	0.681105
Logit con submuestreo (Undersampling)	0.665158	0.136111	0.225980	0.667766	0.666565

Al comparar los modelos generados se encuentra que XGBoost con submuestreo presenta las mejores métricas en Recall 71 %, exactitud 62 % y AUC 66%, seguido del modelo de bosques de árboles aleatorios con submuestreo con 69 % de Recall, 66 % de exactitud y AUC de 67%. En tercer lugar, se ubica el modelo de regresión logística con sobremuestreo que muestra un Recall 67 %, exactitud de 69 % y AUC 69%, en cuarto lugar, se encuentra el modelo de regresión logística con remuestreo por SMOTE + Tomek con Recall de 7 % exactitud de 69 % y AUC 68%, y en último lugar se ubica el modelo de regresión logística con submuestreo el cual presenta un Recall de 67 %, exactitud de 67% y AUC de 67 %.

En cuanto a los tiempos de ejecución de los modelos se observa en la Tabla 28 que el modelo con menor tiempo de ejecución es XGBoost con submuestreo con un tiempo de 0.09 segundos, seguido por el modelo de regresión logística con sobremuestreo que toma un tiempo de ejecución de 7.5 segundos, seguido del modelo de regresión logística con submuestreo con un tiempo de ejecución de 47.7 segundos, en cuarto lugar se encuentra

el modelo de bosques aleatorios con submuestreo que toma un total de 109 segundos de ejecución, y en último lugar se ubica el modelo de regresión logística con remuestreo por SMOTE+ Tomek que tarda 120 segundos en ejecutarse.

Tabla 28. Comparativo de tiempos de cómputo en modelos finales. (Fuente: construcción propia).

Modelo	num_estimadores	Mejor_puntaje	Tiempo (s)
Logit con SMOTE + Tomek	C= 0.001	0.645928	120.05
Logit con sobremuestreo (Oversampling)	C= 0.001	0.638198	7.54
Logit con submuestreo (Undersampling)	C= 0.001	0.644042	47.67
XGBoost con submuestreo (Undersampling)	50	0.700302	0.089
Random Forest con submuestreo (Undersampling)	100	0.628658	108.83

De acuerdo con lo anterior, se seleccionan como modelos finales XGBoost con submuestreo, Logit con sobremuestreo y Bosques aleatorios con submuestreo, teniendo en cuenta los desempeños similares en las métricas de sensibilidad y exactitud. Con éstos tres modelos se realizó el proceso de validación aplicando las predicciones sobre el conjunto de validación del 20%, cuyos resultados se explican en detalle en el numeral 7.

Acorde con los resultados obtenidos en las métricas de estos modelos al aplicar la regresión LASSO para variables categóricas con valores de alfa 0.01 y 000.1, se evidenció que los modelos de XGBoost y Random Forest mejoran su desempeño cuando se reducen de 38 a 8 atributos predictores (Edad, grupo habitante de calle NO, grupo habitante de calle SI, Régimen afiliación Contributivo, Régimen afiliación Especial, Régimen afiliación No asegurado, Régimen afiliación Especial, Régimen afiliación Subsidiado); es decir que se favorecen con la reducción de la dimensionalidad mientras que los modelos de regresión logística mantienen en general sus métricas. En la Tabla 29 se muestra un comparativo de las métricas con los dos métodos de selección de variables. El script que contiene los modelos generados con Lasso 0.001 se encuentra disponibles para su consulta en el repositorio de GitHub en el siguiente link: Modelo_Final_Lasso_0001.ipynb

Tabla 29. Comparativo de métricas con regresión Lasso alfa 0.001 y alfa 0.01. (Fuente: construcción propia).

Modelo	Regresión LASSO con alfa 0.001					Regresión LASSO con alfa 0.01				
	Recall	Precisión	F1 Score	Accuracy	AUC	Recall	Precisión	F1 Score	Accuracy	AUC
XGBoost con submuestreo (Undersampling)	0.628959	0.125225	0.208866	0.652590	0.641704	0.714932	0.126097	0.214383	0.617948	0.662626
Random Forest con submuestreo (Undersampling)	0.660633	0.125107	0.210375	0.638403	0.648644	0.687783	0.137931	0.229781	0.663807	0.674852
Logit con sobremuestreo (Oversampling)	0.669683	0.148445	0.243021	0.695810	0.683774	0.674208	0.150810	0.246485	0.699439	0.687816

7. EVALUACIÓN DEL RENDIMIENTO DE MODELOS DE MACHINE LEARNING

Se procede a aplicar a los tres modelos seleccionados y optimizados, el conjunto de datos de validación con el fin de evaluar el rendimiento de las métricas y determinar el desempeño de estos. El script de los modelos generados con LASSO alfa 0.01, se encuentra disponibles para su consulta en el repositorio de GitHub en el siguiente link: Modelo_Final_Lasso_001.ipynb

En la Tabla 30 se encuentra el comparativo de las métricas obtenidas con el conjunto de datos de validación.

Tabla 30. Comparativo de métricas en modelos finales con set de validación. (Fuente: construcción propia).

Modelo	Recall	Precisión	F1 Score	Accuracy	AUC
XGBoost con submuestreo (Undersampling)	0.70068	0.12439	0.21128	0.61930	0.67485
Random Forest con submuestreo (Undersampling)	0.68027	0.12269	0.20790	0.62277	0.64926
Logit con sobremuestreo (Oversampling)	0.63945	0.14733	0.23949	0.70445	0.67450

Al aplicar el conjunto de validación 20% compuesto por 147 registros de la case 1 correspondientes a pérdidas de seguimiento y 1873 registros de la clase 0 no pérdidas de seguimiento; dos de los tres modelos muestran buen desempeño para la métrica de Recall 68 y 70 %, muy similares a las obtenidas en la etapa de entrenamiento y prueba; esto demuestra que no hay sobreajuste en éstos y cuentan con la capacidad de predecir de manera adecuada la pérdida de seguimiento de tuberculosis que corresponden a la clase positiva con una exactitud que oscila entre 62 y 70 %, siendo el mejor Logit con sobremuestreo pese a una menor sensibilidad 64 %. A su vez en cuanto a la métrica AUC para los tres modelos oscila entre 65 y 67 %, siendo el más bajo para Random Forest con submuestreo.

Con el fin de comprender mejor los resultados obtenidos, se procede a comparar las matrices de confusión obtenidas por cada uno de los tres mejores modelos en la Tabla 31*Tabla 31*.

Tabla 31. Comparación resultados matriz de confusión modelo set de validación. (Fuente: construcción propia)

Modelo	Verdaderos positivos	Verdaderos negativos	Falsos positivos	Falsos negativos
XGBoost Submuestreo	103	1148	725	44
Random Forest con submuestreo (Undersampling)	100	1158	715	47
LOGIT + sobremuestreo	94	1329	544	53

Se evidencia que los tres modelos logran discriminar correctamente la clase positiva (pérdidas de seguimiento). El modelo de XGBoost con submuestreo logra clasificar de manera adecuada 103 casos de los 147 con pérdida de seguimiento, se clasifican 44 como falsos negativos y 725 como falsos positivos. A su vez, el modelo Random Forest con submuestreo logra clasificar de manera adecuada 100 pacientes con pérdida se seguimiento de 147 y 47 corresponden a falsos negativos con 715 falsos positivos, mientras que con el modelo Logit con sobremuestreo logra clasificar de manera adecuada como pérdidas de seguimiento 94 pacientes de 147, 53 corresponden a falsos negativos y 544 a falsos positivos. De esta manera se concluye que el modelo que presenta las mejores métricas de desempeño para predecir pérdida de seguimiento es XGBoost con submuestreo.

Con el fin de asegurar una segunda validación externa, se utilizó la base preliminar del programa año 2023 con previa autorización de la entidad, la cual se compone de 1.953 registros, de los cuales 137 correspondieron a pérdidas de seguimiento. A esta base de datos se le realizó el mismo preprocesamiento y se generaron predicciones sobre este conjunto de datos, los resultados se muestran en la Tabla 32.

Tabla 32. Métricas obtenidas con el conjunto de datos de validación externa. (Fuente: construcción propia)

Modelo	Recall	Precisión	F1 Score	Accuracy	AUC
XGBoost con Submuestreo (Undersampling)	0.69343	0.11757	0.20105	0.61341	0.65040

Se obtuvo un Recall del 69 % similar al obtenido en la fase de prueba y validación, evidenciando que el modelo no presenta sobreajuste y logra predecir de forma correcta 95 de las 137 pérdidas de seguimiento para 2023, mientras que 42 fueron clasificadas como falsos negativos, es decir, corresponden a pérdidas de seguimiento, pero el modelo no las clasificó como tal. Por su parte, 713 registros fueron clasificados como pérdidas de seguimiento sin llegar a serlo en la realidad (falsos positivos).

Con el fin de analizar como inciden los 8 atributos de las 3 variables predictoras en el modelo de XGBoost, se utilizó la técnica de SHAP (SHapley Additive exPlanations), este método muestra la contribución o la importancia de cada característica en la predicción del modelo basado en la teoría de juegos cooperativos y se utilizan para aumentar la transparencia y la interpretabilidad de los modelos de aprendizaje automático [42], los resultados se muestran en la Figura 12.

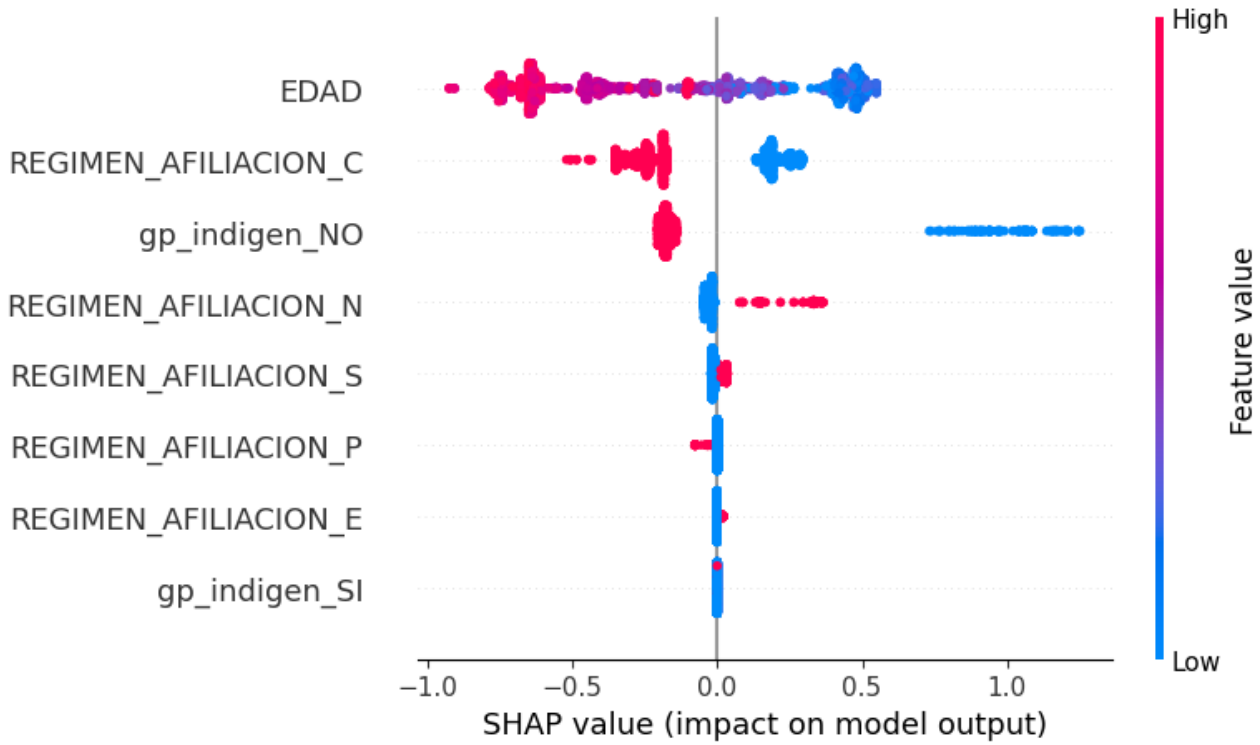

Figura 12. Valores de SHAP para los ocho atributos del modelo XGBoost con submuestreo (Fuente: construcción propia)

Como se observa el modelo permite predecir la pérdida de seguimiento en TB con un 69 % de sensibilidad y una exactitud del 61 %, a partir de 8 atributos generados por 3 variables Edad, régimen de afiliación y el grupo poblacional habitante de calle; estas últimas se subdividen en sus correspondientes categorías (N = No asegurado, C = contributivo, S = Subsidiado, P = Especial, E = Excepción).

En la gráfica de SHAP todas las variables se muestran en orden de importancia de característica global, siendo la primera la más importante y la última la menos importante. La variable edad tiene una contribución positiva muy alta cuando sus valores son bajos, y una contribución negativa baja cuando sus valores son altos. Es decir, que las personas jóvenes inciden más en la predicción de pérdida de seguimiento por parte del modelo. En régimen de afiliación hay una contribución alta cuando no hay aseguramiento o es subsidiado, mientras que tiene una contribución negativa alta cuando pertenece al régimen contributivo. En cuanto al grupo población habitante de calle, cuando pertenece a este grupo poblacional incide en el modelo de forma negativa alta cuando el valor que toma es cercano a cero y cuando se acerca a uno incide de forma positiva baja.

En la Tabla 33 se analizan los valores de Chi cuadrado ($p<=0.05$) para las variables habitante de calle (gp_indigen) y régimen de afiliación, las cuales se relacionan fuertemente con la mayoría de las demás variables. El grupo habitante de calle no tiene significancia estadística con la variable tabaquismo, grupo poblacional desplazado y enfermedad mental.

Por otro lado, la variable régimen de afiliación presenta fuerte correlación con la mayoría de las variables a excepción de la comorbilidad tabaquismo.

Tabla 33. Valores de chi Cuadrado para las variables predictoras habitante de calle y régimen de afiliación (Fuente: construcción propia).

Variables contrastadas	Habitante de calle (p-valor)	Régimen de afiliación (p-valor)
Sexo	9.59 e-13	1.26 e -21
Pertenencia étnica	4.71 e-06	1.77 e -40
Grupo poblacional desplazado	0.71	2.31 e -16
Grupo poblacional migrante	0.02	0.0
Grupo poblacional habitante de calle	-	2.95 e -163
Subred residencia	1.49 e -188	1.58 e -209
Régimen de afiliación	2.95 e -163	-
Tipo de Tuberculosis	2.86 e -22	5.94 e -18
Condición de ingreso	1.03 e -133	4.16 e -27
Resultado baciloscopia	2.23 e -21	5.42 e -41
Coinfección con VIH	1.12 e -21	5.55 e -76
Comorbilidad Consumo de SPA	7.02 e -145	3.75 e -20
Comorbilidad Desnutrición	1.17 e -14	9.32 e -25
Comorbilidad enfermedad mental	0.22	0.03
Pérdida de seguimiento	1.15 e -170	1.39 e -69
Comorbilidad Tabaquismo	0.25	0.61

Estos resultados son esperados si se tiene en cuenta que las personas habitantes de calle, por lo general no cuentan con aseguramiento en el sistema de salud o pertenecen a régimen subsidiado; muchos de ellos son pacientes previamente tratados para tuberculosis, se encuentran en estado de mal nutrición e incluso presentan coinfección con VIH, la gran mayoría presenta formas pulmonares de la enfermedad, tienen afectaciones en salud mental dado por adicciones, la más frecuente es el consumo abusivo de sustancias psicoactivas adicionalmente se localizan o ubican en determinadas zonas de la ciudad donde es fácil acceder a este tipo de sustancias. A su vez, el régimen de afiliación presenta una fuerte asociación con la subred de residencia y el grupo habitante de calle, por las razones ya mencionadas. La pertenencia étnica y el grupo desplazado presentan relación al igual que el grupo de migrantes teniendo en cuenta que por ejemplo estos últimos no cuentan con afiliación al sistema de salud, los desplazados, indígenas y afrocolombianos, en su mayoría pertenecen al régimen subsidiado o no se encuentran afiliados.

Los resultados obtenidos se asemejan a los reportados por Chinagudaba y otros [42] en 2024, en el cual utilizaron varios algoritmos de aprendizaje automático (Naive Bayes, árboles de decisión, Random Forest y k-Nearest Neighbors (k-NN), Gradient Boosting Machine (GBM), XGBoost, LightGBM y CatBoost), para predecir los resultados del tratamiento de la tuberculosis en pacientes de Karnataka (India). En ese trabajo el conjunto de datos usados fue dividido 70:15:15. No obstante, la prevalencia de pérdida del seguimiento era mucho más alta (22%) que la nuestra (7%). Para balancear las clases también utilizaron técnicas de balance SMOTE y Random Oversampling y midieron el desempeño de los modelos utilizando como métrica la sensibilidad (Recall) y promedio de sensibilidad (AvRecall), al igual que en el presente proyecto. Nuestros resultados se

asemejan dado que obtuvieron los mejores resultados con XGBoost y Random Forest con SMOTE que tuvieron 97% de Recall para ambos, mientras que con sobremuestreo 95 % y 98 %, respectivamente. En nuestro caso, el mejor Recall se obtiene con XGBoost con submuestreo 70 % y Random Forest con submuestreo 68 %. De los modelos probados, encontraron que XGBoost y GBM demostraron resultados equilibrados en todas las técnicas de codificación, lo que indica su robustez y versatilidad. Sin embargo, el modelo LightGBM fue particularmente notable, superando a todos los demás modelos en términos de Recall. En cuanto a las variables que inciden en la pérdida de seguimiento para TB, encuentran a partir del análisis SHAPA que el estado del paciente (comorbilidades), genero, condición de VIH influyen positivamente en la predicción, mientras que el consumo de alcohol y el diagnóstico de PHI influyen negativamente [42]. En nuestro caso aplicando el mismo análisis SHAP se encuentra que la edad, el régimen de afiliación y pertenecer al grupo poblacional habitante de calle son las variables más influyentes en el modelo, dado que las demás se encuentran fuertemente correlacionadas con ésta.

Nuestros resultados son similares a los reportados por Ferreira y colaboradores en 2021 [55], quienes desarrollaron y evaluaron modelos de aprendizaje automático (SVM, RF, XGBoost y regresión logística) para predecir la pérdida de seguimiento en pacientes con tuberculosis (LTFU) en Brasil, encontrando una prevalencia más alta 12,51% (n = 3.036, de cohorte integral de 24.265 pacientes) que la encontrada para el distrito capital en el presente proyecto. No obstante, separaron a estos pacientes en dos grupos: las pérdidas de seguimiento con y sin inicio de tratamiento, evaluando ambos grupos por separado. La selección de variables también se realizó con Lasso y otras técnicas (SFS, SFFS, SBS y SBFS) al igual que en nuestro proyecto y se balancearon clases con métodos de submuestreo. Se seleccionaron 17 variables y se modelaron 9 algoritmos, en todos igual que en el presente proyecto se optimizaron hiperparametros y se usó validación cruzada con 10 folds. Compararon las métricas usando datos balanceados y no balanceados. En los resultados, XGBoost superó en sensibilidad a los otros modelos alternativos Random Forest 55 % LTFU antes del tratamiento y 76 % durante el tratamiento; el modelo de regresión logística 59 % LTFU antes del tratamiento y 64 % durante el tratamiento; mostrando una mayor sensibilidad (0,81), la puntuación F1 (0,85) y el AUC (0,921) para el grupo LTFU antes del tratamiento. Comparado con los resultados obtenidos en el presente proyecto la sensibilidad para regresión logística fue mayor en nuestros modelos (68 y 70 %) y para Random Forest con submuestreo, las métricas de sensibilidad nuestras son ligeramente superiores 68%. No obstante, en el estudio mencionan como limitante, no realizar un análisis general de la pérdida de seguimiento sin hacer distinción sobre el inicio del tratamiento lo que no permite capturar el impacto general de la LTFU, como si se describe en nuestros resultados. Por otro lado, los pacientes que no inician tratamiento en el distrito capital, en la mayoría de los casos corresponden a diagnósticos tardíos (post mortem). En el estudio mencionado no es claro el motivo por el cual no inician tratamiento. Dentro de los predictores más influyentes en la pérdida de seguimiento describen, el nivel de educación, el historial de hospitalización, el consumo de alcohol, el ingreso ambulatorio y el historial previo de tuberculosis surgieron como precursores de LTFU previo al tratamiento. La situación laboral, la admisión ambulatoria, la presencia de hepatitis

crónica/cirrosis, las reacciones adversas a los medicamentos, la disponibilidad de contactos alternativos y la cobertura del seguro médico ejercieron una influencia sustancial en la LTFU en la fase de tratamiento. Con relación a esto en el presente proyecto se evidenció que, muchos de estos predictores se encuentran fuertemente relacionados con tres variables la edad, el régimen de afiliación y pertenecer al grupo poblacional habitante de calle.

En un estudio similar realizado por Chen y otros en 2024 [40], desarrollaron y evaluaron modelos de aprendizaje automático (SVM, RF, XGBoost y regresión logística) para predecir la pérdida de seguimiento en pacientes con tuberculosis en China a partir de datos de 2017 a 2021, con una cohorte integral de 24.265 pacientes y una prevalencia de LTFU del 12,51 % (n = 3.036). También se dividió en dos grupos a esta población quienes no iniciaron tratamiento y quienes hicieron LTFU durante el tratamiento, entendido como una interrupción de este por más de dos meses a diferencia de nuestra definición mayor a un mes. Al igual que en nuestro proyecto se utilizaron pruebas de chi-cuadrado y regresión logística para las comparaciones entre grupos. Seleccionaron variables con la técnica de LASSO, hicieron validación cruzada con 10 folds en los algoritmos seleccionados. No obstante, no se mencionan uso de técnicas de balanceo para las clases. Compararon el desempeño de los modelos basado en AUC y curva ROC. XGBoost logró el AUC promedio más alto (0,921), superando a Random Forest (0,828), Logistic Regression (0,736) y SVM (0,677) para LTFU antes de tratamiento. En el conjunto de validación, XGBoost mostró la mayor sensibilidad (0,81), puntuación F1 (0,85) y AUC (0,921) respecto a los otros modelos igual que en nuestro proyecto. La secuencia descendente de rendimiento para los otros modelos fue Logístico (0,811), RF (0,755) y SVM (0,712), similar a los resultados obtenidos por nuestro proyecto. En cuanto a la métrica de sensibilidad nuestros modelos de regresión logística muestran mejor desempeño 68% (0.64 para el grupo LTFU sin tratamiento y 0.59 para el grupo con tratamiento), mientras que Random Forest tuvo un comportamiento similar con Recall entre 52-68 % (0.76 para el grupo LTFU sin tratamiento y 0.55 para el grupo con tratamiento). A su vez, el Recall para XGBoost muestra un resultado dispar respecto a nuestros mejores resultados 65-71 % (0.81 para el grupo LTFU sin tratamiento y 0.53 para el grupo con tratamiento). Es posible que la disparidad se deba a las técnicas usadas para balancear las clases dado que en el estudio reportan mejor sensibilidad para el grupo que realiza pérdida de seguimiento durante el tratamiento y a la aplicación diferenciada de los modelos para ambos grupos. El nivel de educación, el historial de hospitalización, el consumo de alcohol, el ingreso ambulatorio y el historial previo de tuberculosis surgieron como precursores de LTFU previo al tratamiento. La situación laboral, la admisión ambulatoria, la presencia de hepatitis crónica/cirrosis, las reacciones adversas a los medicamentos, la disponibilidad de contactos alternativos y la cobertura del seguro médico ejercieron una influencia sustancial en la LTFU en la fase de tratamiento. En nuestro caso se evidenció que, muchos de estos predictores se encuentran fuertemente relacionados con otras tres variables la edad, el régimen de afiliación y pertenecer al grupo poblacional habitante de calle.

En otro estudio realizado por Moreno *et al* en 2024 [39], desarrollaron una puntuación para predecir el riesgo de LTFU durante el tratamiento a partir de una cohorte nacional de casos 2015-2022 en Brasil. La prevalencia de LTFU fue del 17% (n= 41.373 de un total de 243.726

casos incluidos), mucho mayor que la encontrada en el distrito capital para un período de tiempo similar. Los algoritmos utilizados fueron regresión logística, bosque aleatorio y refuerzo de gradiente ligero, particionando la data en 80% entrenamiento y 20% prueba. Utilizaron técnicas de submuestreo para balanceo de clases y RFECV. Las variables predictoras seleccionadas fueron 8: TB previa, uso de drogas, edad, sexo, infección por VIH y nivel de escolaridad. No obstante, para el estudio excluyeron niños (<18 años), grupos vulnerables o TB resistente a medicamentos. En nuestro caso, si se incluyeron a los menores (grupos priorizados) y a los grupos vulnerables, acorde con los hallazgos del análisis bivariado realizado, donde la enfermedad es más prevalente en algunos de estos grupos sociales vulnerables. El modelo de regresión logística al igual que en el presente proyecto, demostró sus máximas capacidades predictivas con una fuerte regularización, en particular C = 0,01. El modelo RF logró su mejor desempeño al establecer la profundidad máxima en 8 y utilizó un conjunto de 500 árboles de decisión. En nuestro caso, la mejor profundidad fue de 10 con un conjunto de 50 árboles de decisión, lo que significa que cada uno de los árboles de decisión del modelo puede tomar decisiones hasta en 10 niveles de profundidad. No obstante, para el estudio mencionado no tuvo el mejor desempeño ya que el modelo subestimó la probabilidad real de la clase positiva, en nuestro caso esta situación se presentó con los modelos bayesianos. Esos sistemas de puntuación de predicción exhibieron un área bajo la curva (AUC) que oscilaba entre 0,71 y 0,72, no se mencionan los resultados de las otras métricas como sensibilidad por lo cual no es posible realizar una comparación detallada con nuestros resultados. Sin embargo, concluyen que el modelo Light Gradient Boosting resultó en el mejor rendimiento de predicción, especificidad de ponderación y sensibilidad. Dentro del análisis SHAP, la tuberculosis previa fue la característica más importante seguida del consumo de drogas. En nuestro caso la tuberculosis previamente tratada, se relaciona fuertemente con el grupo poblacional habitante de calle y el consumo de SPA también se correlaciona con este grupo poblacional. Dentro de las limitantes mencionadas indican que la mayoría de las comorbilidades y características clínicas fueron autoinformadas (sesgo de clasificación errónea), solo incluyeron casos de tuberculosis pulmonar y, en consecuencia, no puede aplicarse a la tuberculosis extrapulmonar o diseminada; a diferencia del presente proyecto que si incluye la TB en todas las formas clínicas y las comorbilidades son identificadas por el personal de salud y/o a través de cruces de información con otras fuentes de información.

8. CONCLUSIONES Y TRABAJOS FUTUROS

8.1 CONCLUSIONES

El modelo con mejor desempeño corresponde a XGBoost con submuestreo dado que obtuvo la mejor métrica de sensibilidad para identificar pacientes con pérdida de seguimiento, lo que es similar a estudios en países con alta carga de la enfermedad.

Dentro de las variables predictoras que influyen en la pérdida de seguimiento de los pacientes en el distrito capital, se encuentra el pertenecer a un grupo poblacional vulnerable, no contar con afiliación al sistema de seguridad social en salud, tener entre 36 y 41 años, presentar alguna de las siguientes comorbilidades VIH, desnutrición, tabaquismo, enfermedad mental, consumo de SPA y contar con antecedente de pérdida de seguimiento previo. Para el modelo final, tres variables edad, régimen de afiliación y grupo poblacional habitante de calle; distribuidas en 8 atributos (Edad, grupo habitante de calle No, grupo habitante de calle Si, régimen afiliación contributivo, régimen afiliación especial, régimen afiliación no asegurado, régimen afiliación excepción, régimen afiliación subsidiado) son las que permiten realizar la predicción.

Para el entrenamiento de los modelos es indispensable identificar las variables más relevantes a través de diferentes técnicas y posteriormente, realizar la selección teniendo en cuenta los resultados de los estimadores BIC y AIC. La mejor técnica de selección de variables en nuestro caso fue Group Lasso, reportada también en otros estudios similares. En los conjuntos de datos donde las clases se encuentran muy desbalanceadas, se debe realizar un tratamiento especial con el conjunto de entrenamiento, siendo la técnica de submuestreo (undersampling) la que presentan un mejor resultado de *Recall*, respecto al modelo base (sin balanceo de clases).

Las métricas para utilizar en modelos con clases desequilibradas son más difíciles ya que se requiere métricas de rendimiento que se centren en la clase minoritaria. En este caso la sensibilidad o Recall, permite identificar los modelos que mejor discriminan la predicción en la clase minoritaria y que corresponde, a las personas que tienen mayor probabilidad de tener pérdida en el seguimiento de tuberculosis.

8.2 TRABAJOS FUTUROS

El despliegue del modelo se encuentra a discreción de la entidad, no obstante, la implementación exitosa de este modelos podría marcar un hito importante en la lucha contra la tuberculosis fortaleciendo la adherencia terapéutica por parte de los usuarios, demostrando así el potencial transformador del aprendizaje automático en la atención médica y en el abordaje de problemas de salud pública.

Con las pérdidas de seguimiento identificadas al ingreso del programa de TB en el distrito

capital, se pueden enfocar aún más los esfuerzos del personal de salud para contribuir en la gestión del riesgo, focalizando de esta manera las actividades existentes para los equipos locales del programa de tuberculosis, en articulación con otros actores del sector salud.

El modelo se podría implementar como complemento al sistema de información del programa distrital de tuberculosis y se puede incorporar técnicas procesamiento del lenguaje natural (NLP) para mejorar el aprendizaje de modelos. Incluso en otros estudios reportan el desarrollo de aplicaciones útiles para identificar rápidamente a los pacientes con posibles pérdidas de seguimiento, calculadora web de fácil uso por parte del personal de salud (https://tbprediction.herokuapp.com/) [30].

9. REFERENCIAS BIBLIOGRÁFICAS

[1] Organization World Health, «Tuberculosis,» 2023. [En línea]. Available: https://www.who.int/es/news-room/fact-sheets/detail/tuberculosis. [Último acceso: 13 11 2023].

[2] Ministerio de Salud y protección Social, «Enfermedades transmisibles,» 2023. [En línea]. Available: https://www.minsalud.gov.co/salud/publica/PET/Paginas/Tuberculosis.aspx. [Último acceso: 2023 Noviembre 13].

[3] A. B. Ortiz, «Detección temprana de fracasos a tratamiento en pacientes con tuberculosis pulmonar,» *Rev Med Hered,* vol. 18, nº 3, pp. 123-128, 2007.

[4] O. A. Cruz, «Informe de evento Tuberculosis año 2022,» Ministerio de Salud y protección Social, Bogotá DC, 2022.

[5] G. Lastre Amell, M. Suarez Villa, J. Rodríguez Lopez, D. Martínez Sierra y M. Navarro Agamez, «Social determinants of health and the loss of follow-up to the treatment of pulmonary,» *Salus,* vol. 24, nº 1, pp. 26-32, 2020.

[6] C. M. Sauer, D. Sasson, K. Paik, N. McCague, A. Celi, I. Sánchez Fernández y B. Illigens, «Feature selection and prediction of treatment failure in tuberculosis,» *PLoS ONE,* vol. 13, nº 11, pp. 1-14, 2018.

[7] I. Montiel, E. Alarcón, S. Aguirre, G. Sequera y D. Marín, «Factores asociados al resultado de tratamiento no exitoso de pacientes con tuberculosis sensible en Paraguay,» *Rev Panam Salud Publica,* vol. 44, nº 89, pp. 1-9, 2020.

[8] T. Portos Pereiro, «Factores asociados al abandono del tratamiento de la tuberculosis en el adulto,» *Publicaciones didacticas,* nº 96, pp. 274-279, 2018.

[9] C. L. Perlaza, F. E. Cruz Mosquera, L. M. Ramirez Murillo, V. Becerra Sepulveda y C. D. Córdoba Arenas, «Factores de abandono al tratamiento de la tuberculosis en la red pública de salud,» *Rev Saude Publica.,* vol. 57, nº 8, pp. 1-8, 2023.

[10] P. Molina Chailan , S. Mendoza Parra , K. Saez y S. Cabrera, «Biopsychosocial profile of the patient with tuberculosis and factors associated with therapeutic adherence,» *Revista chilena de enfermedades respiratorias,* vol. 36, nº 2, pp. 100-108, 2020.

[11] V. Túñez Bastidas, M. García Ramosa, M. Pérez del Molinoa y F. Lado Ladoa, «Epidemiología de la tuberculosis,» *Medicina Integral,* vol. 39, nº 5, pp. 172-180, 2002.

[12] Instituto Nacional de Salud, «Protocolo de vigilancia de Tuberculosis,» 22 03 2022. [En línea]. Available: https://www.ins.gov.co/buscador-eventos/Lineamientos/Pro_Tuberculosis%202022.pdf. [Último acceso: 17 11 2023].

[13] Secretaria Distitral de Salud, «SaluData,» Secretaría Distrital de Salud, 2023. [En línea]. Available: https://saludata.saludcapital.gov.co/osb/index.php/datos-de-salud/enfermedades-trasmisibles/enfermedadesmicobacterias/. [Último acceso: 17 11 2023].

[14] Ministerio de Salud y Proteccion Socia *Resolución 227 de 2020,* Bogota, 2020.

[15] N. Tatés-Ortega, J. Álvarez, L. López, A. Mendoza y E. Alarcón, «Pérdida en el seguimiento de pacientes tratados por tuberculosis resistente a rifampicina o multidrogorresistente en Ecuador,» *Rev Panam Salud Publica,* vol. 43, p. 91, 2019.

[16] IBM, «¿Qué es machine learning?,» IBM, [En línea]. Available: https://www.ibm.com/mx-es/topics/machine-learning. [Último acceso: 29 Noviembre 2023].

[17] A. Lugo, «¿Qué es el Machine Learning?,» INVID, 2020. [En línea]. Available: https://invidgroup.com/es/machine-learning-metodos/. [Último acceso: 29 Noviembre 2023].

[18] Mathworks, «¿Qué es Machine Learning?,» Mathworks, 2023. [En línea]. Available: https://la.mathworks.com/discovery/machine-learning.html#:~:text=Machine%20Learning%20emplea%20dos%20tipos,en%20los%20datos%20de%20entrada.. [Último acceso: 29 Noviembre 2023].

[19] K. Huh, «Surviving in a Random Forest with Imbalanced Datasets,» Medium, 13 02 2021. [En línea]. Available: https://medium.com/sfu-cspmp/surviving-in-a-random-forest-with-imbalanced-datasets-b98b963d52eb. [Último acceso: 20 08 2024].

[20] IBM, «¿Qué son los clasificadores Naïve Bayes?,» IBM, [En línea]. Available: https://www.ibm.com/mx-es/topics/naive-bayes#:~:text=%C2%BFQu%C3%A9%20son%20los%20clasificadores%20Na%C3%AFve,para%20realizar%2

0tareas%20de%20clasificaci%C3%B3n.. [Último acceso: 26 09 2024].

[21] Lumivero, «Extreme Gradient Boosting (XGBOOST),» XLSTAT, [En línea]. Available: https://www.xlstat.com/es/soluciones/funciones/extreme-gradient-boosting-xgboost. [Último acceso: 26 09 2024].

[22] J. Roque López, «Técnicas de seleccion de variables en regresión lineal múltiple,» Universidad Internacional de Andalucía, Andalucía, 2021.

[23] Datacamp, «Tutorial de Lasso y regresión Ridge en Python,» Datacamp, 03 05 2024. [En línea]. Available: https://www.datacamp.com/es/tutorial/tutorial-lasso-ridge-regression. [Último acceso: 26 09 2024].

[24] K. Kipsang, «Feature Selection; Stepwise Regression (Forward Selection and Backward Elimination) with Python,» Medium, 09 09 2023. [En línea]. Available: https://medium.com/@kelvinsang97/feature-selection-stepwise-regression-forward-selection-and-backward-elimination-with-python-d53230be995c. [Último acceso: 26 09 2024].

[25] L.-H. Hsu, «Feature Selection with “Recursive Feature Elimination” (RFE) for Parisian Bike Count Data,» Medium, 19 02 2024. [En línea]. Available: https://medium.com/@hsu.lihsiang.esth/feature-selection-with-recursive-feature-elimination-rfe-for-parisian-bike-count-data-23f0ce9db691. [Último acceso: 06 09 2024].

[26] J. García Abad, «Comparativa de técnicas de balanceo de datos. Aplicacion a un caso real para la predicción de fuga de clientes,» Universidad de Oviedo, Oviedo, 2021.

[27] D. Mondal, «Imbalanced data classification: Oversampling and Undersampling,» Medium, 05 02 2023. [En línea]. Available: https://medium.com/@debspeaks/imbalanced-data-classification-oversampling-and-undersampling-297ba21fbd7c. [Último acceso: 20 08 2024].

[28] M. Adel, «Tomek Links,» Medium, 14 03 2023. [En línea]. Available: https://medium.com/@mahmoudadel200215/tomek-links-948ea097199e. [Último acceso: 20 08 2024].

[29] R. A. A. Viadinugroho, «Imbalanced Classification in Python: SMOTE-Tomek Links Method,» Medium, 18 04 2021. [En línea]. Available: https://towardsdatascience.com/imbalanced-classification-in-python-smote-tomek-links-method-6e48dfe69bbc. [Último acceso: 06 09 2024].

[30] J. D. Rojas Sanchez, «Desarrollo de un modelo de Machine Learning para la clasificación de tipos de dengue de acuerdo a su nivel de severidad: Un estudio de caso de Bucaramanga, Colombia,» Universidad El Bosque, Bogotá, 2023.

[31] E. T. Arias Zuluaga, «Desarrollo de un modelo predictivo con inteligencia artificial para establecer clasificación ASA a pacientes en una consulta preanestésica,» Universidad de Antioquia, Medellín, 2020.

[32] A. Naidu, S. S. Nayak, V. Sundararajan y S. Lulu, «Advances in computational frameworks in the fight against TB: The way forward,» *Front. Pharmacol,* vol. 14, pp. 1-24, 2023.

[33] R. Chikhale , S. Pawar , M. Kolpe , O. Shinde , K. Dahlous , S. Mohammad , P. Patil y S. Bhowmick , «Identification of mycobacterial Thymidylate kinase inhibitors: a comprehensive pharmacophore, machine learning, molecular docking, and molecular dynamics simulation studies,» *Mol Divers,* vol. 4, nº 28, pp. 1947-1964, 2024.

[34] Y. Lin , Y. Zou , M. Karlsson y E. Svensson , «A pharmacometric multistate model for predicting long-term treatment outcomes of patients with pulmonary TB.,» *J Antimicrob Chemother,* vol. 74, nº 10, pp. 2561-2569, 2024.

[35] L. Peetluk, F. Ridolfi, P. Rebeiro, D. Liu, V. Rolla y T. Sterling, «Systematic review of prediction models for pulmonary tuberculosis treatment outcomes in adults,» *BMJ Open,* vol. 11, pp. 1-17, 2021.

[36] W. S. Van , . H. H. Lin y M. M. Claassens , «A systematic review of prediction models for prevalent pulmonary tuberculosis in adults,» *Int J Tuberc Lung Dis,* vol. 21, p. 405–411, 2017.

[37] J.-B. Ma, L.-C. Zeng, F. Ren, L.-Y. Dang, H. Luo, Y.-Q. Wu, X.-J. Yang, R. Li, H. Yang y Y. Xu, «Development and validation of a prediction model for unsuccessful treatment outcomes in patients with multi-drug resistance tuberculosis,» *BMC Infectious Diseases,* vol. 23, nº 289, pp. 1-9, 2023.

[38] E. W. Pefura-Yone, A. D. Balkissou, V. Poka-Mayap, H. K. Fatime-Abaicho, P. T. Enono-Edende y A. P. Kengne, «Development and validation of a prognostic score during tuberculosis treatment,» *BMC Infectious Diseases,* vol. 17, nº 251, pp. 1-9, 2017.

[39] M. S. Moreno , B. Barreto Duarte, C. L. Vinhaes, M. Araújo Pereira, E. R. Fukutani , K. Bone Bergamaschi, A. Kristki , M. Cordeiro Santos , V. C. Rolla, T. R. Sterling, A. T. Queiroz y B. B. Andrade, «Machine learning algorithms using national registry data to predict loss to follow-up during tuberculosis treatment,» *BMC Public Health,* vol. 24, nº 1, p. 1385, 2024.

[40] J. Chen, Y. Jiang , L. Zhihuan , M. Zhang, L. Linlin , A. Li y L. Hongzhou , «Predictive machine learning models

for anticipating loss to follow-up in tuberculosis patients throughout anti-TB treatment journey,» *Sci Rep,* vol. 14, nº 1, pp. 1-9, 2024.

[41] O. Hussain y K. Junejo, «Predicting treatment outcome of drug-susceptible tuberculosis patients using machine-learning models,» *Informatics for health & Social Care,* vol. 44, nº 2, pp. 135-151, 2019.

[42] S. N. Chinagudaba, D. Gera, K. K. Vamsi, U. Shankar, K. Kiran , A. Singarajpure, U. Shivayogappa, N. Somashekar , V. K. Chadda y B. Sharath, «Predictive Analysis of Tuberculosis Treatment Outcomes Using Machine Learning : A Karnataka TB Data Study at a Scale,» *Next Research,* vol. 1, nº 1, pp. 3050-4759, 2024.

[43] L. Peetluk, P. Rebeiro, F. Ridolfi, B. Andrade, M. Cordeiro-Santos, A. Kritski, B. Durovni, S. Calvacante, M. Figuereido, D. Haas, D. Liu, V. Rolla y T. Sterling, «A Clinical Prediction Model for Unsuccessful Pulmonary Tuberculosis Treatment Outcomes,» *Clinical Infectious Diseases,* vol. 74, pp. 973-982, 2022.

[44] M. Kulkarni, S. Golechha, R. Raj, J. K. Sreedharan, A. Bhardwaj, S. Rathod, B. Vadera, J. Kurada, S. Mattoo, R. Joshi, K. Rade y A. Raval, «Predicting Treatment Adherence of Tuberculosis Patients at Scale,» *Machine Learning for Health,* vol. 193, p. 35–61, 2022.

[45] M. Kheirandish, D. Catanzaro, V. Crudu y S. Zhang, «Integrating landmark modeling framework and machine learning algorithms for dynamic prediction of tuberculosis treatment outcomes,» *Journal of the American Medical Informatics Association,,* vol. 29, nº 5, p. 900–908, 2022.

[46] A. Díaz Pérez, C. Roldan Menco, J. Muñoz Baldiris, A. Giraldo, E. García Caro, J. Llanos Perdomo y F. Campo Peñaloza, «Factores predictivos para el riesgo de tuberculosis en población vulnerable: clasificación del riesgo por medio del uso de red neuronal artificial,» *Revista Ciencia y Salud Virtual,* vol. 4, nº 1, pp. 62-79, 2012.

[47] O. F. Bedoya Leiva, H. S. Guarín Aristizábal y J. Z. Agudelo Delgado, «Aplicación de técnicas de inteligencia artificial para la detección de tuberculosis pulmonar en Colombia,» *Revista EIA,* vol. 20, nº 39, pp. 1-23, 2023.

[48] I. C. Sánchez Vega, «Predicción de series temporales de incidencia de tuberculosis pulmonar a partir de algoritmos de inteligencia computacional,» Universidad Antonio Nariño, Bogotá, 2023.

[49] A. D. Orjuela Cañón, A. L. Jutinico, C. Awad, A. Palencia y E. Vergara, «Generación de modelos alternativos basados en inteligencia computacional para tamización y diagnóstico de Tuberculosis pulmonar,» Escuela de Medicina y Ciencias de la Salud Universidad del Rosario GiBiome, Bogotá, 2022.

[50] C. Roberto, «Crisp-DM: las 6 etapas de la metodología del futuro,» MBA USP ESALQ, 31 05 2022. [En línea]. Available: https://blog.mbauspesalq.com/es/2022/05/31/crisp-dm-las-6-etapas-de-la-metodologia-del-futuro/. [Último acceso: 26 09 2024].

[51] J. Brownlee, «Tour of Evaluation Metrics for Imbalanced Classification,» Machine Learning Mastery, 01 05 2021. [En línea]. Available: https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/. [Último acceso: 2024 09 26].

[52] B. Moufad, «Leveraging linear regression for feature selection of categorical and continuous variables,» 09 07 2022. [En línea]. Available: https://towardsdatascience.com/beyond-linear-regression-467a7fc3bafb. [Último acceso: 20 11 2024].

[53] J. Brownlee, «Tour of Evaluation Metrics for Imbalanced Classification,» machinelearningmastery, 21 05 2021. [En línea]. Available: https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/. [Último acceso: 20 08 24].

[54] K. t. school, «¿Qué es GridSearchCV?,» Keepcoding tech school, 16 04 2024. [En línea]. Available: https://keepcoding.io/blog/que-es-gridsearchcv/. [Último acceso: 26 09 2024].

[55] . L. Ferreira Da Silva , M. Herverton, G. Oliveira Alves, L. M. Florêncio Souza, E. da Silva Rocha, J. F. Lorenzato de Oliveira, T. Lynn, V. Sampaio y P. Takako Endo, «Benchmarking Machine Learning Models to Assist in the Prognosis of Tuberculosis,» *Informatics ,* vol. 8, nº 2, pp. 2-17, 2021.

[56] Secretaria Distitral de Salud, «Rendición de Cuentas vigencia 2022 a la Contraloría de Bogotá D.C. Balance Social,» Secretaría Distrital de Salud, Bogotá DC, 2022.

[57] L. E. Ramírez Bejarano y J. E. Chamorro Ortega, «Drug resistant tuberculosis,» *Revista Colombiana de Neumología,* vol. 25, nº 3, pp. 170-173, 2013.

[58] O. Bernal, R. López, E. Montoro, P. Avedillo, K. Westby y M. Ghidinelli, «Determinantes sociales y meta de tuberculosis en los Objetivos de Desarrollo Sostenible en las Américas,» *Rev Panam Salud Publica,* vol. 44, nº 153, pp. 1-8, 2020.

[59] Ministerio de Salud y Protección social, «Biblioteca Digital,» 1993. [En línea]. Available: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/RESOLUCION-8430-DE-1993.PDF. [Último acceso: 26 Diciembre 2023].

[60] P. Huet, «Ética en la Inteligencia Artificial,» openwebinars, 05 Junio 2023. [En línea]. Available:

https://openwebinars.net/blog/etica-en-la-inteligencia-artificial/. [Último acceso: 26 Diciembre 2023].

[61] O. Panamericana de la Salud y Consejo de Organizaciones Internacionales de las C, «Pautas éticas internacionales para la investigación relacionada con la salud con seres humanos,» Consejo de Organizaciones Internacionales, Ginebra, 2016.

[62] S. d. i. y. comercio, «Manejo de información personal, 'Habeas data',» Superintendencia de industria y comercio, [En línea]. Available: https://www.sic.gov.co/manejo-de-informacion-personal#:~:text=El%20principio%20de%20confidencialidad%20en,de%20finalizada%20su%20relaci%C3%B3n%20con. [Último acceso: 29 Diciembre 2023].

[63] R. I. d. p. d. datos, «Recomendaciones generales para el tratamiento de datos en la inteligencia artificial,» Red Iberoamericana de protección de datos, Naucalpan de Juarez, 2019.

[64] M. d. Interior, «derechos de autor,» Direccion Nacional de derechos de autor , 24 01 2024. [En línea]. Available: http://derechodeautor.gov.co:8080/software. [Último acceso: 28 01 2024].

[65] Affirmalegal, «Derechos de Autor para Software en Colombia,» Affirmalegal, 13 02 2023. [En línea]. Available: https://www.affirmalegal.com/blog/derechos-de-autor-para-software-en-colombia/. [Último acceso: 28 01 2024].

[66] A. Guío Español, E. Tamayo Uribe, P. Gómez Ayerbe y M. P. Mujica, Marco ético para la inteligencia artificial en Colombia, Bogotá: Departamento Administrativo de la Presidencia de la República, 2021.

[67] J. Perez Colin, «Los costos de la Inteligencia Artificial pueden irse a las nubes,» 09 09 2021. [En línea]. Available: https://blog.jorgeperezcolin.mx/costos-inteligencia-artificial-pueden-irse-a-las-nubes/. [Último acceso: 29 01 2024].

[68] N. Unidas, Naciones Unidas, [En línea]. Available: https://www.un.org/sustainabledevelopment/es/health/. [Último acceso: 4 Marzo 2024].

[69] J. F. Samaniego, «Inteligencia artificial: hacia un impacto económico, social y medioambiental positivo,» Foretica, 20 noviembre 2023. [En línea]. Available: https://foretica.org/inteligencia-artificial-hacia-un-impacto-economico-social-y-medioambiental-positivo/. [Último acceso: 4 Marzo 2024].

[70] A. d. S. Rezende Moreira, A. Lineu Kritski y A. C. Calçada Carvalho, «Social determinants of health and catastrophic costs associated with the diagnosis and treatment of tuberculosis,» *J Bras Pneumol,* vol. 46, nº 5, pp. 1-5, 2020.

[71] Y. Dahami, «Understanding Ordinary Least Squares (OLS) and Its Applications in Statistics, Machine Learning, and Deep Learning,» Medium, 22 05 2024. [En línea]. Available: https://medium.com/@dahami/understanding-ordinary-least-squares-ols-and-its-applications-in-statistics-machine-learning-ad2c13681501. [Último acceso: 26 09 2024].

[72] IBM, «¿Qué son los clasificadores Naive Bayes?,» IBM, [En línea]. Available: https://www.ibm.com/es-es/topics/naive-bayes. [Último acceso: 26 09 2024].

[73] J. Montomoli, L. Romeo, S. Moccia, M. Bernardini, L. Migliorelli, D. Berardini, A. Donati, A. Carsetti, M. G. Bocci, P. D. Wendel, T. Fumeaux, P. Guerci, R. A. Schüpbach, C. Ince, E. Frontoni y M. P. Hilty, «Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA score at ICU admission in COVID-19 patients,» *Journal of Intensive Medicine ,* vol. 1, p. 110–116 , 2021.

[74] Y. Xiao, Y. Xiao, R. Huang, F. Jiang, J. Zhou y T. Yang, «Interpretable machine learning in predicting drug-induced liver injury among tuberculosis patients: model development and validation study,» *BMC Medical Research Methodology,* vol. 24, nº 92, pp. 2-10, 2024.

[75] C. Tarabanis, E. Kalampokis, M. Khalil, C. Alviar, L. Chinitz y L. Jankelson, «Explainable SHAP-XGBoost models for in-hospital mortality after myocardial infarction,» *Cardiovascular Digital Health Journal,* vol. 4, nº 4, pp. 126-132, 2023.

[76] . A.-z. Peng, X. Kong, S. Liu, H. Zhang , L. Xie, L. Ma, Q. Zhang y Y. Chen, «Explainable machine learning for early predicting treatment failure risk among patients with TB-diabetes comorbidity,» *Scientific Reports ,* vol. 14, nº 6814 , pp. 3-11, 2024.

ANEXOS

Anexo 1. Presentación ante el comité de ética.

En este anexo se encuentran los correos electrónicos enviados hacia comité de ética de la Secretaría Distrital de Salud, en el cual se realiza la postulación del anteproyecto y se agenda espacio por parte del comité para la presentación de este.

RE: presentación anteproyecto

Desde SDS, Comiteetica <Comiteetica@saludcapital.gov.co>
Fecha Lun 15/01/2024 8:23
Para Diana Azucena, Guerrero Barreto <DAGuerrero@saludcapital.gov.co>
CC Ruben Dario, Rodriguez Camargo <RDRodriguez@saludcapital.gov.co>

Estimada Diana:

Recibe un cordial saludo.

Acuso recibo de la documentación. En los próximos días realizaré una revisión preliminar del anteproyecto con el fin de identificar -si fuese necesario- aspectos que requieran modificación o adición antes de la presentación del proyecto ante el Comité de Ética de la Investigación (CEI) de la SDS. En caso de que se requieran cambios, lo notificaré lo más pronto posible, de tal manera que, para el día miércoles 14 de febrero de 2024 se tenga la versión definitiva del protocolo y anexos para ser enviados a los miembros del CEI para su lectura y revisión.

Por otra parte, la presentación del proyecto quedaría programada para la sesión ordinaria del **21 de febrero de 2024**. Días previos a la sesión estaré enviando la respectiva agenda e información adicional.

Quedo atento a tus comentarios o inquietudes.

Cordialmente,

David Bazurto Barragán
Secretario Técnico
Comité de Ética de la Investigación
Dirección Planeación Sectorial
Teléfono 601 3649090 ext. 9078

Cc: Ruben Dario, Rodriguez Camargo <RDRodriguez@saludcapital.gov.co>
Asunto: presentación anteproyecto

Señores:
Comité de ética
Secretaría Distrital de Salud

Cordial saludo,

Por medio de la presente nos permitimos remitir propuesta de anteproyecto y documentos relacionados, para ser evaluados por parte del comité de ética de la institución, con el fin de cumplir los requisitos establecidos. Se anexa:

1. Anteproyecto
2. hojas de vida de imvestigadores principales y directora del proyecto
3. Declaración expresa de conocimiento y cumplimiento de pautas éticas en investigación, debidamente diligenciado y firmado.
4. Declaración de confidencialidad en el manejo de bases de datos, debidamente diligenciado y firmado.
5. Carta de sometimiento de proyectos de investigación al Comité Ética de la Investigación, debidamente diligenciado y firmado.

Quedamos atentos.

Cordialmente,

Diana Guerrero
Profesional especializado
Subdirección de Vigilancia en Salud Pública
3649090 Ext 9343

Anexo 2. Consulta cesión de derechos patrimoniales y de transformación a oficina de asuntos jurídicas de la Secretaría Distrital de Salud.

En este anexo se encuentra la consulta realizada al área jurídica de la entidad solicitando el procedimiento de cesión de derechos patrimoniales y de transformación, distribución y reproducción de la obra; como requisito solicitado por parte del comité de ética

BOGOTÁ | SECRETARÍA DE SALUD

2024-IE-06334

012000
Bogotá D.C., 5 de marzo de 2024

MEMORANDO

PARA: MELISSA TRIANA LUNA
Jefe Oficina Asuntos Jurídicos

DE: DIANE MOYANO ROMERO
Directora de Epidemiología, Análisis y Gestión de Políticas de Salud Colectiva

Asunto: Consulta procedimiento de cesión de derechos patrimoniales y de transformación, distribución y reproducción de la obra

Respetada Doctora Melissa:

Reciba un cordial saludo, por medio de la presente se informa que, en la Dirección de Eppidemiología, Análisis y Gestión de Políticas de Salud Colectiva se llevará a cabo la ejecución del proyecto aplicado "Aplicación de Modelos Machine Learning para predecir el riesgo de pérdida de seguimiento en tuberculosis"; del cual se obtendrá un desarrollo de software, que será entregado por parte de los investigadores principales a la Secretaría Distrital de Salud.

Dicho proyecto, ya fue presentado ante el comité de ética institucional quienes solicitan aclaración y especificación sobre la forma en que se van a transferir los derechos patrimoniales y de transformación a la Secretaría Distrital de Salud. Para ello, se solicita amablemente nos informe el procedimiento establecido (trámites y documentos) para realizar la cesión de derechos patrimoniales y de transformación,

Carrera 32 No. 12 - 81
Teléfono: 3649090
www.saludcapital.gov.co

CO-SC-CER518781

BOGOTÁ | SECRETARÍA DE SALUD

distribución y reproducción de dicho desarrollo, una vez se cuente con él.

Cordialmente,

Carrera 32 No. 12 - 81
Teléfono: 3649090
www.saludcapital.gov.co

ALCALDÍA MAYOR DE BOGOTÁ D.C.

ALCALDÍA MAYOR DE BOGOTÁ D.C.

Anexo 3. Respuesta a consulta cesión de derechos patrimoniales y de transformación a oficina de asuntos jurídicas de la Secretaría Distrital de Salud.

En este anexo se encuentra la respuesta emitida por el área jurídica de la entidad sobre la consulta realizada al área jurídica de la entidad solicitando el procedimiento de cesión de derechos patrimoniales y de transformación, distribución y reproducción de la obra; como requisito solicitado por parte del comité de ética.

BOGOTÁ | SECRETARÍA DE SALUD

2024-IE-10686

000100

Bogotá D.C., 23 de abril de 2024

MEMORANDO

PARA: DIANE MOYANO ROMERO
Directora de Epidemiología, Análisis y Gestión de Políticas de Salud Colectiva

DE: MARISOL CHACON FONTECHA
Jefe de Oficina de Asuntos Jurídicos (E)

Asunto: Concepto sobre el procedimiento de cesión de derechos patrimoniales de transformación, distribución y reproducción del software desarrollado durante la ejecución del proyecto *"Aplicación de Modelos Machine Learning para predecir el riesgo de pérdida de seguimiento en tuberculosis."*

Cordial saludo,

Mediante el memorando identificado bajo la radicación 2024-IE-06334, la Directora de Epidemiología, Análisis y Gestión de Políticas de Salud Colectiva de la Secretaría Distrital de Salud (SDS), menciona que en desarrollo del proyecto denominado *"Aplicación de Modelos Machine Learning para predecir el riesgo de pérdida de seguimiento en tuberculosis"*, se desarrollará un software cuyos derechos patrimoniales, de transformación, distribución y reproducción, serán cedidos por los investigadores principales del proyecto a favor de la SDS, razón por la cual se eleva consulta a la Oficina de Asuntos Jurídicos (OAJ) de la entidad en relación con el procedimiento que se debe seguir para estos efectos, frente a lo cual procedo a rendir mi concepto en los siguientes términos:

Carrera 32 No. 12 - 81 Teléfono: 3649090 www.saludcapital.gov.co

CO-SC-CER115791

BOGOTÁ | SECRETARÍA DE SALUD

1.- LA DEFINICIÓN DE "SOFTWARE" EN EL MARCO DE LA LEGISLACIÓN APLICABLE A LA PROTECCIÓN DE DERECHOS PATRIMONIALES DE AUTOR.

De cara a la consulta, considero menester empezar por precisar la noción de "*software*" que resulta aplicable para efectos de la protección que nuestro marco normativo le otorga a los derechos patrimoniales de sus desarrolladores, también conocidos como autores del programa de ordenador, contenida, hoy día, en el artículo tercero (3º) de la Decisión 351 de 1993 de la Comunidad Andina (Can)[1], a cuyo tenor:

"Artículo 3. A los efectos de esta decisión, se entiende por:

"(...)

"Programa de ordenador (Software): Expresión de un conjunto de instrucciones mediante palabras, códigos, planes o en cualquier otra forma que, al ser incorporadas en un dispositivo de lectura automatizada, es capaz de hacer que un ordenador -un aparato electrónico o similar capaz de elaborar informaciones-, ejecute determinada tarea u obtenga determinado resultado. El programa de ordenador comprende también la documentación técnica y los manuales de uso."

En consecuencia; este concepto resulta de recibo exclusivamente a la cesión de los derechos patrimoniales que los desarrolladores de esta clase de bienes, también definidos como su (s) autor (es), hagan en favor de una tercera (3ª) persona, en este caso la SDS, puesto que si se llegase a tratar de otra clase de obra (s), las consideraciones acá contenidas deberán ser revisadas para efectos de determinar su aplicación en el caso concreto.

2.- QUIENES SON LOS TITULARES DE LOS DERECHOS PATRIMONIALES DEL DEL SOFTWARE Y EL CONTENIDO DE LOS MISMOS.

La mencionada Decisión 351 de 1993 también identifica el concepto de autor de la obra, de significativa relevancia para establecer la titularidad de los derechos patrimoniales derivados de su explotación.

[1] Contentiva del Régimen Común sobre Derechos de Autor y Derechos Conexos, la cual, a voces de su artículo 1, tiene por finalidad *"reconocer una adecuada y efectiva protección a los autores y demás titulares de derechos sobre las obras de ingenio, en el campo literario, artístico o científico, cualquiera que sea el género o la forma de expresión y sin importar el mérito literario o artístico ni su destino"*

En efecto, en el mismo artículo tercero (3º) se define al autor de la obra como la *"persona física que realiza la creación intelectual"*, de modo tal que es esta persona

Carrera 32 No. 12 - 81 Teléfono: 3649090 www.saludcapital.gov.co

ALCALDÍA MAYOR DE BOGOTÁ D.C.

CO-SC-CER115791

ALCALDÍA MAYOR DE BOGOTÁ D.C.

o grupo de personas, o sus derechohabientes, quienes están facultados para autorizar o prohibir el desarrollo de las siguientes actividades:

a) Su reproducción mediante cualquier forma o procedimiento.
b) Su comunicación pública.
c) Su distribución pública.
d) Su importación al territorio de cualquier país.
e) Su traducción, adaptación, arreglo u otra transformación del software.

Sin embargo, el artículo 91 de la Ley 23 de 1982 determina que estos derechos patrimoniales de autor sobre la creación intelectual de empleados o funcionarios públicos, en ejercicio o cumplimiento de las obligaciones a su cargo, son de propiedad de la entidad pública para la cual laboran.

Dice la norma en cita:

"Artículo 91. Los derechos de autor sobre las obras creadas por empleados o funcionarios públicos, en cumplimiento de las obligaciones constitucionales y legales a su cargo, serán de propiedad de la entidad pública correspondiente.

"Se exceptúa de esta disposición las lecciones o conferencias de los profesores.

"Los derechos morales serán ejercidos por los autores, en cuanto su ejercicio no sea incompatible con los derechos y obligaciones de las entidades públicas afectadas."

En ese mismo sentido, los "LINEAMIENTOS DE PROPIEDAD INTELECTUAL DE LA SECRETARÍA DISTRITAL DE SALUD"[1], definen al titular de los derechos de propiedad intelectual en los siguientes términos:

"7.4.1. ¿Qué se entiende por titular de derechos de propiedad intelectual?:

"El titular de derechos de propiedad intelectual es la persona natural o jurídica dueña de los derechos de exclusiva que reconocen los distintos mecanismos antes mencionados sobre productos de carácter inmaterial. A este respecto vale la pena mencionar que (…) estos derechos de exclusiva nacen en la persona natural y/o excepcionalmente, en la persona jurídica generadora del respectivo producto. Es por ello que este tipo de sujetos son conocidos con el nombre de 'titulares originarios'. Sin embargo, otras personas

[1] *Adoptados como obligatorios para todas las dependencias de la SDS y el Fondo Financiero Distrital de Salud mediante la circular 034 del 4 de diciembre de 2023.*

naturales o jurídicas pueden hacerse con todas o algunas prerrogativas de carácter transmisible reconocidas por la normatividad de propiedad intelectual a través de múltiples

INTELECTUAL DE LA SECRETARÍA DISTRITAL DE SALUD" que más adelante se explicarán.

c) Frente a lo normado en el artículo 28 de la Ley 1450 de 2011, tenemos que tratándose de personas vinculadas mediante la celebración de un contrato de prestación de servicios o se trate de un trabajador vinculado a la entidad mediante la celebración de un contrato de trabajo, y en cumplimiento de sus obligaciones contractuales hubiese desarrollado la obra, es decir, se le hubiese encargado, como parte de sus obligaciones, de manera expresa, la creación del mencionado software, se presume que el titular de los derechos patrimoniales de autor es el contratante o el empleador, salvo que en el contrato se hubiese pactado alguna cuestión diferente.

En el evento de ser de recibo la aplicación de la mencionada presunción, no sería necesario la cesión de los derechos patrimoniales de autor a favor de la entidad.

e) Si dicho contratista o trabajador desarrolla el software pero sin que su contrato tenga por objeto principal o accesorio la ejecución de esa actividad, este conserva en su haber los derechos patrimoniales que por ello se derivan, de modo tal que resulta necesario efectuar la cesión de derechos patrimoniales a favor de la SDS.

Lo antes explicado en palabras de la agencia estatal Colombia Compra Eficiente[2]:

"H. ¿Quién es el titular de Derechos Patrimoniales sobre las obras creadas por empleados o funcionarios públicos?

"En este punto debe distinguirse de si se trata de: i] una obra creada por un empleado público en cumplimiento de las funciones del cargo que ocupa en la Entidad o, si se trata de ii] una obra generada por la actividad de la persona del funcionario como producto de una actividad no vinculada a sus funciones.

"Para el primer caso, el artículo 91 de la Ley 23 de 1982, establece de manera expresa que los Derechos de Autor, en este caso los patrimoniales, son de propiedad de la Entidad Estatal, quien tendrá el uso exclusivo y excluyente de hacer los usos que considere necesarios sobre la obra (…) Por el contrario, en el segundo supuesto, al no estar vinculada la creación de la obra a una función pública, los Derechos Morales y Patrimoniales permanecen en cabeza del autor, es decir, no opera una transferencia en favor de la entidad para la cual labora.

[2] Agencia Nacional de Contratación Pública -Colombia Compra Eficiente-, Guía de Propiedad Intelectual en la Contratación Pública, diciembre de 2022, páginas 18 y siguientes.

"(…) Debe precisarse que las relaciones laborales de las Entidades Estatales con personas vinculadas a título de trabajadores oficiales se rigen por las normas de derecho común,

mecanismos como el caso de los contratos, las presunciones de ley y/o los escenarios de herencia en procesos de sucesión. Estos sujetos, en calidad de nuevos titulares de los derechos patrimoniales de exclusiva reconocidos por la normatividad aplicable en materia de propiedad intelectual, son conocidos con el nombre de 'titulares derivados'".

De otro lado; el artículo 28 de la Ley 1450 de 2011, modificatorio del artículo 20 de la Ley 23 de 1982, que se refiere a la titularidad de los derechos patrimoniales sobre las obras creadas en cumplimiento de un contrato de prestación de servicios o de un contrato de trabajo, es decir, obras por encargo, precisa:

"ARTÍCULO 28. PROPIEDAD INTELECTUAL [DE LAS OBRAS] EN CUMPLIMIENTO DE UN CONTRATO DE PRESTACIÓN DE SERVICIOS O DE UN CONTRATO DE TRABAJO. El artículo 20 de la Ley 23 de 1982 quedará así:

"Artículo 20. En las obras creadas para una persona natural o jurídica en cumplimiento de un contrato de prestación de servicios o de un contrato de trabajo, el autor es el titular originario de los derechos patrimoniales y morales; pero se presume, salvo pacto en contrario, que los derechos patrimoniales sobre la obra han sido transferidos al encargante o al empleador, según el caso, en la medida necesaria para el ejercicio de sus actividades habituales en la época de creación de la obra. Para que opere tal presunción se requiere que el contrato conste por escrito. El titular de las obras de acuerdo a este artículo podrá intentar directamente o por intermedia persona acciones preservativas contra actos violatorios de los derechos morales informando previamente al autor o los autores para evitar duplicidad de acciones."

Así las cosas; conforme el marco legal citado, podemos encontrarnos ante las siguientes situaciones en relación con la titularidad de los derechos patrimoniales de autor del software:

a) De cara a lo establecido en el artículo 91 de la Ley 23 de 1982, cuando la obra es creada por una persona que labora para la entidad, vinculado mediante una relación legal y reglamentaria -Vgr. Empleado oficial-, y ello obedece al cumplimiento de las funciones a su cargo, la titularidad de los derechos patrimoniales se encuentra en cabeza de la Entidad para la cual dicha persona labora, caso en el cual no es necesario efectuar alguna cesión de derechos.

b) También la recta interpretación de dicho artículo nos lleva a concluir que cuando la obra es creada por dicho servidor público, pero en desarrollo de una actividad ajena al cumplimiento de sus funciones legales y reglamentarias, los derechos morales y patrimoniales permanecen en cabeza del creador del software, por ende, deberá hacerse la cesión de derechos a favor de la SDS, previo el cumplimiento de los requisitos establecidos en los *"LINEAMIENTOS DE PROPIEDAD*

teniendo aplicación las cláusulas del contrato de trabajo, de las convenciones colectivas y reglamentos internos de trabajo. En ese sentido, los trabajadores oficiales no están sujetos a lo dispuesto en el artículo 91 de la Ley 23 de 1982."

"I. ¿Quién es el titular de Derechos Patrimoniales sobre las obras creadas en cumplimiento de un contrato de prestación de servicios o de un contrato de trabajo [obras por encargo]?.

"El artículo 20 de la Ley 23 de 1982 establece que los Derechos Patrimoniales sobre las obras que hayan sido creadas en cumplimiento de un contrato de prestación de servicios o de un contrato de trabajo, se presumen transferidos al encargante de la obra, esto quiere decir, al contratante o empleador. Es decir, siempre que el contrato tenga por objeto, principal o accesorio, la creación de la obra por parte del contratista o trabajador. Para que opere esta presunción solo se requiere que el contrato conste por escrito."

3.- LA CESIÓN DE LOS DERECHOS PATRIMONIALES DE LOS TITULARES DE LA OBRA A FAVOR DE LA SDS.

Establecidas como quedaron las definiciones de software, de titular de la obra y los derechos patrimoniales que le asisten, así como los casos en que resulta necesaria la cesión de los derechos patrimoniales de autor, procede entrar al análisis de esta transferencia de derechos, en este caso, a favor de la SDS.

La Política de Propiedad Intelectual del Distrito Capital[3], mediante la cual se establecen los principios rectores obligatorios para todas sus entidades, en materia de gestión de los activos de propiedad intelectual, define los principios de protección y gestión de la Titularidad de esta clase de activos, en los siguientes términos:

"3. Protección. Cada Entidad debe tomar medidas necesarias para evitar la pérdida o no consolidación de derechos de PI sobre Activos de PI de importancia operacional, económica o estratégica, y adelantar los trámites para su registro ante las Oficinas Nacionales Competentes cuando la formalidad del registro sea legalmente exigida para adquirir los derechos de PI sobre el activo o cuando la importancia del activo de PI aconseje su registro para dejar constancia pública de su creación y/o titularidad para efectos probatorios."

[3] Política de Propiedad Intelectual del Distrito Capital, Secretaría Jurídica Distrital, año 2018.

"Gestión de la Titularidad: Cada entidad debe tomar las medidas adecuadas para asegurar a su favor la titularidad – o cuando ello no sea posible, las autorizaciones o licencias apropiadas – de la PI cuyo control resulte esencial o estratégico para garantizar el adecuado desarrollo de sus procesos operativos y el cumplimiento de sus funciones. En todas las situaciones en que puedan generarse o adquirirse PI como resultado de actividades

laborales, contractuales, de consultoría, investigación o emprendimiento en colaboración, se deberá especificar a quien corresponderá la titularidad el PI."

En desarrollo de lo anterior, los "LINEAMIENTOS DE PROPIEDAD INTELECTUAL DE LA SECRETARÍA DISTRITAL DE SALUD", se encarga de precisar cuales son los diferentes mecanismos de transferencia de derechos de propiedad intelectual por parte de su creador a favor de la SDS o del FFDS, dentro de estos los siguientes:

a) Transferencia de derechos por virtud de la Ley (la contemplada en el ya analizado artículo 91 de la Ley 23 de 1982).
b) Presunciones de transferencia de derechos de propiedad intelectual (a las que se refiere el mencionado artículo 28 de la Ley 1450 de 2011, modificatorio del artículo 20 de la Ley 23 de 1982)
c) Mecanismos contractuales de transferencia de derechos, a los cuales procedemos a referirnos, bajo el entendido de que se deberá acudir a este mecanismo, únicamente, cuando la transferencia de derechos no opere por virtud legal o por efecto de las presunciones a las que se refiere la norma antes citada, sobre el particular, resulta muy ilustrativo lo dispuesto en el numeral 7.4.8 de los lineamientos, a los cuales nos remitimos en su integridad.

La cesión de tales derechos patrimoniales de autor es un acto permitido por nuestra legislación, que se encuentra previsto en los artículos 29 y siguientes de la decisión de la CAN ya muchas veces citada, en el entendido de que se trata de un acto por medio del cual su titular los transfiere a una tercera persona y, para tales efectos, remite a la legislación interna de cada país miembro de la CAN[3], en nuestro caso, entre otros, al artículo 181 de la Ley 1955 de 2019, modificatorio del artículo 183 de la Ley 23 de 1982, conforme al cual se precisa:

"ARTÍCULO 181. ACUERDOS SOBRE DERECHOS PATRIMONIALES. Modifíquese el artículo 183 de la Ley 23 de 1982, el cual quedará así:

"'Artículo 183. Acuerdos sobre derechos patrimoniales. Los acuerdos sobre derechos patrimoniales de autor o conexos, deberán guiarse por las siguientes reglas:

[3] Dispone el artículo 30 de la Decisión 351 de 1993:
"Artículo 30.- Las disposiciones relativas a la cesión o concesión de derechos patrimoniales y a las licencias de uso de las licencias protegidas, se regirán por lo previsto en las legislaciones internas de los Países Miembros."

"Los derechos patrimoniales de autor o conexos pueden transferirse, o licenciarse por acto entre vivos, quedando limitada dicha transferencia o licencia a las modalidades de explotación previstas y al tiempo y ámbito territorial que se determinen contractualmente.

"La falta de mención en el tiempo limita la transferencia o licencia a cinco (5) años, y la del ámbito territorial, al país en el que se realice la transferencia o licencia.

"Los actos o contratos por los cuales se transfieren, parcial o totalmente, los derechos patrimoniales de autor o conexos, deberán constar por escrito como condición de validez.

"Todo acto por el cual se enajene transfiera, cambie o limite el dominio sobre el derecho de autor, o los derechos conexos, así como cualquier otro acto o contrato que implique exclusividad, deberá ser inscrito en el Registro Nacional del Derecho de Autor, para efectos de publicidad y oponibilidad ante terceros.

"Será inexistente toda estipulación en virtud de la cual el autor transfiera de modo general o indeterminable la producción futura, o se obligue a restringir su producción intelectual o a no producir."

Ahora bien; tratándose de la cesión por parte de funcionarios y empleados públicos a favor de entidades estatales, su autorización se encuentra expresamente consagrada el artículo 1 de la Ley 44 de 1993, que dice:

"ARTÍCULO 1. Los empleados y funcionarios públicos que sean autores de obras protegidas por el Derecho de Autor, podrán disponer contractualmente de ellas con cualquier entidad de derecho público."

De modo tal que; estamos ante una excepción legal al régimen de inhabilidades e incompatibilidades previsto en la Constitución Política y en el Estatuto General de Contratación de la Administración Pública, en particular, frente a la prohibición contenida en su artículo 8 referente a la imposibilidad que a tales servidores les asiste para celebrar contratos con las entidades estatales.

Con fundamento en lo previsto por el artículo 183 de la Ley 23 de 1982, modificada, como ya se advirtió, por el artículo 181 de la Ley 1955 de 2019, en los "LINEAMIENTOS DE PROPIEDAD INTELECTUAL DE LA SECRETARÍA DISTRITAL DE SALUD", a la altura de sus páginas 65 y siguientes, además de señalarse, con toda la razón, que el acto o contrato contentivo de la cesión debe constar por escrito, so pena de su invalidez, deberá ocuparse de precisar los siguientes aspectos:

a) La identificación del (los) cedente (s), que deben ostentar la condición de titulares o desarrolladores del software y los derechos que transfieren.
b) La identificación del Cesionario, que conforme la consulta, sería la SDS, también podría ser el FFDS, a quien se le transferirán los derechos patrimoniales cedidos.
c) La identificación de los derechos patrimoniales y/o modalidades de explotación que le serán cedidos a la SDS.
d) La duración en tiempo y el ámbito territorial dentro del cual se ceden los derechos patrimoniales.

e) Deberá incluirse la autorización por parte del cedente, que le permita al cesionario la divulgación de la obra sin violentar el derecho moral de ineditud.
f) También deberá incluir la correspondiente autorización para efectuar el registro de la obra ante las autoridades correspondientes.
g) Como parte del contrato de cesión, se deberá incluir una cláusula en virtud de la cual el cedente garantice que es el titular de los derechos cedidos, que su creación no vulnera derechos de terceros y se obliga a salir en defensa del cesionario en caso de reclamaciones por parte de esos terceros frente al uso o explotación del PI, como también a reparar íntegramente al cesionario en caso de que este último deba cancelar alguna suma de dinero a favor del reclamante.

Todo lo anterior sin perjuicio de señalar que una vez esta se lleve a cabo, el acto que la contenga debe registrarse en el Registro Nacional de Derechos de Autor.

4.- PROCEDIMIENTO AL INTERIOR DE LA SDS.

En lo que corresponde al procedimiento al interior de la SDS para efectos de llevar a cabo la cesión de los derechos patrimoniales, conforme los "LINEAMIENTOS DE PROPIEDAD INTELECTUAL DE LA SECRETARÍA DISTRITAL DE SALUD", específicamente lo allí contenido a propósito del *"7.5. Lineamientos asociados a la Gestión de la Propiedad Intelectual en la Secretaría Distrital de Salud – Fondo Financiero Distrital de Salud"*, al cual nos remitimos en su integridad, dicho procedimiento está integrado por las siguientes etapas:

4.1. Identificación de los activos intangibles de importancia operacional, económica o estratégica para la SDS - FFDS: Esta etapa estará a cargo de todas las dependencias de la SDS puesto que tienen la obligación de *"reportar los productos de relevancia institucional susceptibles de ser protegidos por la propiedad intelectual."*

Para estos efectos; las dependencias de la SDS cuentan con la asesoría técnica del Comité de Propiedad Intelectual al que se hace referencia en el numeral 7.1 de los lineamientos.

4.2. Consolidación de la información reportada en la etapa de identificación de los activos intangibles de importancia operacional, económica o estratégica para la SDS - FFDS: En la cual solo interviene la Secretaría Técnica del Comité de Propiedad Intelectual, que es la encargada *"consolidar y depurar la información reportada por las distintas dependencias en el proceso de identificación."*

4.3. Priorización y estudio de titularidad de los derechos de propiedad intelectual sobre activos intangibles de importancia operacional, económica o estratégica identificados: En esta etapa se seleccionarán los activos intangibles respecto de los cuales resulta *"pertinente e idóneo iniciar los respectivos trámites de registro y protección ante los organismos competentes."*, además, se evaluará qué productos son realmente de la SDS y/o del FFDS.

Etapa que se encuentra a cargo de los siguientes actores:

- Secretaría Técnica del Comité de Propiedad Intelectual: Que tiene a su cargo "estructurar el informe de priorización y estudio de titularidad de los activos susceptibles de ser protegidos por la propiedad intelectual."
- Asamblea General del Comité de Propiedad Intelectual: Quien tiene la competencia para *"definir y aprobar los activos intangibles que serán priorizados."*

En esta etapa; si el activo PI no es de propiedad de la SDS conforme a las reglas ya explicadas, se deberán celebrar los contratos de cesión de derechos patrimoniales teniendo en cuenta, además, lo dispuesto en el No. 7.7.4 de los lineamientos.

La elaboración del documento contentivo de la cesión de los derechos patrimoniales está a cargo de la Subdirección de Contratación de la SDS, la cual, conforme el artículo 41 del Decreto 507 de 2013, tiene a su cargo las siguientes funciones:

*"**ARTÍCULO 41º. SUBDIRECCIÓN DE CONTRATACIÓN.** Corresponde a la Subdirección de Contratación el ejercicio de las siguientes funciones:*

"1. Realizar la gestión contractual de personas naturales y jurídicas para el apoyo de la gestión institucional independientemente de la cuantía y su naturaleza, con el fin de garantizar que éstos se ajusten al marco legal a los planes, programas y proyectos de la entidad.

"(...)
"3. Tramitar y elaborar las minutas de las novedades que se presenten en la ejecución de los contratos y que soliciten las diferentes dependencias o supervisores de los contratos.

"4. Atender las peticiones que en relación con los procesos de celebración, ejecución de contratos formulen las diferentes dependencias de la Secretaría y los particulares.

4.4. Protección y/o registro de los bienes inmateriales de interés para la propiedad intelectual de la SDS – FFDS: En la cual se realizarán los siguientes dos tipos de acciones:

BOGOTÁ | SECRETARÍA DE SALUD

- Se adelantarán las labores preliminares requeridas para llevar a cabo el registro de los derechos de la propiedad intelectual, a cargo de la Subsecretaría de Planeación y Gestión Sectorial
- Se llevarán a cabo los trámites ente las autoridades competentes para obtener a protección de la titularidad de los derechos patrimoniales de autor, función a cargo de la Oficina de Asuntos Jurídicos de la SDS.

4.5. Inventario y seguimiento de los activos intangibles de importancia operacional, económica o estratégica identificados, evaluados y priorizados, etapa a cargo de la Asamblea General del Comité de Propiedad Intelectual de la SDS, de la Secretaría Técnica del Comité de Propiedad Intelectual de la SDS y de su Subsecretaría de Bienes y Servicios.

4.6. Observancia y Defensa de la propiedad intelectual institucional: a cargo de la Oficina de Asuntos Jurídicos de la SDS, cuando el uso de los activos intangibles de la entidad constituya un uso no autorizado y/o pueda conllevar a la concreción de escenarios de detrimento patrimonial.

Finalmente, me permito puntualizar que, de conformidad con lo previsto en los numerales 1 y 3 del artículo 4 del Decreto 507 de 2013 "*Por el cual se modifica la Estructura Organizacional de la Secretaría Distrital de Salud de Bogotá, D.C.*", dentro de las funciones de esta Oficina se encuentran las de "*asesorar y apoyar en materia jurídica a las distintas dependencias de la Secretaría Distrital de Salud y a las entidades adscritas del sector salud en el Distrito Capital*" y "*emitir conceptos, responder tutelas y absolver consultas y derechos de petición que en materia jurídica formulen los ciudadanos o ciudadanas, las entidades y las autoridades en general que tengan relación con los asuntos de su competencia*", y que de conformidad con lo establecido en el artículo 28 del Código de Procedimiento Administrativo y de lo Contencioso Administrativo, en concordancia con el Consejo de Estado Sala de lo Contencioso Administrativo Sección Primera Consejero Ponente: Rafael E. Ostau de Laffont Pianetta Bogotá D.C., veintidós (22) de abril de dos mil diez (2010) Radicación núm: 11001 0324 000 2007 00050 01 Actor: Jairo José Arenas Romero, los conceptos emitidos por la autoridades públicas como respuestas a peticiones realizadas en ejercicio del derecho a formular consultas, no serán de obligatorio cumplimiento o ejecución, todo lo cual implica que el concepto emitido por esta Oficina Asesora Jurídica a través del presente memorando, constituye sólo un criterio orientador en la interpretación y aplicación de la normatividad aplicable al caso objeto de consulta, conservando la dependencia y autoridad pública consultante, la autonomía en el ejercicio de sus competencias legales y reglamentarias.

BOGOTÁ | SECRETARÍA DE SALUD

Respetada Doctora Diane, en los anteriores términos dejo rendido el concepto solicitado y le reitero mi absoluta disponibilidad para aclarar o completar cualquier aspecto que usted considere necesario.

Cordialmente,

MARISOL CHACON FONTECHA
Jefe de Oficina de Asuntos Jurídicos (E)

Elaboró: Dr. David Fernando Rojas, Abogado contratista OAJ
Revisó: Luz Alba Farfan Casallas,Profesional Universitario OAJ

Carrera 32 No. 12 - 81
Teléfono: 3649090
www.saludcapital.gov.co
icontec ISO 9001 IQNET
CO-SC-CER03781

Anexo 4. Aprobación por parte del comité de ética.

En este anexo se encuentra el correo electrónico enviado por el comité de ética aprobando el uso de la información para el desarrollo del proyecto aplicado.

Outlook

Respuesta del Comité de Ética de la Investigación de la SDS

Desde SDS, Comiteetica <Comiteetica@saludcapital.gov.co>
Fecha Lun 26/02/2024 7:57
Para Diana Azucena, Guerrero Barreto <DAGuerrero@saludcapital.gov.co>; Ruben Dario, Rodriguez Camargo <RDRodriguez@saludcapital.gov.co>

Estimada Diana y estimado Rubén:

Reciban un cordial saludo.

Por medio del presente me permito comunicar las sugerencias y/o solicitudes que el Comité de Ética de la Investigación de la Secretaría Distrital de Salud ha realizado al proyecto: "***Aplicación de Modelos Machine Learning para predecir el riesgo de pérdida de seguimiento en tuberculosis***" antes de la emisión del concepto ético:

- Es necesario especificar dentro del protocolo de investigación que los datos que se emplearán en el proyecto no saldrán de la Secretaría Distrital de Salud y que estos se emplearán únicamente dentro de las instalaciones. Asimismo se recomienda hacer énfasis en que las medidas de seguridad que se tendrán para salvaguardar y proteger la información del proyecto, serán las mismas con las que actualmente cuenta la institución.
- Se solicita que el profesional responsable asignado por la Secretaría Distrital de Salud antes de entregar la información (bases de datos), asigne un código que represente la identidad de la persona titular de la información. En último término, se solicita la anonimización de los datos.
- En el caso de requerir la búsqueda de información faltante en la base de datos, esta se debe solicitar a la persona encargada de la gestión de la información y nunca, en ningún caso se debe obtener por los medio de los investigadores del proyecto.
- Se debe hacer llegar por escrito al Comité:
  - La solicitud de la información, en donde sea evidente que se trata de un documento dirigido al líder de la unidad o dependencia que custodia la información, el nombre de la unidad o dependencia a la que se solicita la información, las variables solicitadas (o relacionar el Anexo 2), el nombre del proyecto y el objetivo del mismo (esto con radicado de recibido o trazabilidad de la entrega al responsable).
  - La respuesta por escrito en donde se accede a la entrega de la información desde el líder de la unidad o la persona encargada.
- Aclarar y especificar de qué manera se van a transferir los derechos patrimoniales y de transformación a la Secretaría Distrital de Salud, independientemente de los derechos de autor que los investigadores tienen sobre su obra.
- Completar la información presentada en el presupuesto del proyecto con relación a los honorarios de los desarrolladores del modelo.

Una vez se hayan hecho las respectivas modificaciones al proyecto, por favor, enviarlas al correo electrónico: comiteetica@saludcapital.gov.co para su revisión.

Con el ánimo de facilitar la revisión de los ajustes al documento, por favor, subrayarlos.

Quedo atento a sus comentarios o inquietudes.

Cordialmente,

ALCALDÍA MAYOR DE BOGOTÁ D.C. | SECRETARÍA DE SALUD | BOGOTÁ

David Bazurto Barragán
Secretario Técnico
Comité de Ética de la Investigación
Dirección Planeación Sectorial
Teléfono 601 3649090 ext. 9078

Anexo 5. Aval de entrega bases de datos.

Este anexo contiene el aval por parte de la Dirección Epidemiología de la Secretaría Distrital de Salud, para la entrega de las bases de datos solicitadas al comité de ética.

BOGOTÁ | SECRETARÍA DE SALUD

SECRETARIA DISTRITAL DE SALUD 10 de marzo de 2024
Al contestar Cite Este No. 2024-EE-36187
Folios: 0 Anexos: 0
ORIGEN: DIANE MOYANO ROMERO - 012000-Dirección De Epidemiología Análisis Y Gestión De Políticas De Salud Colectiva
DESTINO: RUBÉN DARIO RODRIGUEZ CAMARGO - -
TIPO DE DOCUMENTO: Comunicaciones oficiales
ASUNTO: Respuesta solicitud_ENTREGA DE INFORMACION DEL PROYECTO

012000

Señor
RUBÉN DARIO RODRIGUEZ CAMARGO
rubenrodriguezc@javerianacali.edu.co
La ciudad

Asunto: Respuesta solicitud_ENTREGA DE INFORMACION DEL PROYECTO: *"Aplicación de Modelos de Machine Learning para predecir el riesgo de pérdida de seguimiento de tuberculosis"*

Respetado señor Rodriguez:

Reciba un cordial saludo. Por medio de la presente, se informa que una vez surtido el proceso con el comité de ética de la Secretaría Distrital de Salud, el cual mediante resolución 1317 del 2022 se constituye en el espacio que permite la emisión de conceptos técnicos frente a proyectos de investigación. Una vez se cuente con este concepto, el cual es necesario allegar con la correspondiente carta de presentación de la Universidad, se puede hacer entrega de la información requerida para el desarrollo del proyecto a los investigadores principales previa firma de los correspondiente formatos de confidencialidad institucionales. El proceso será acompañado por la referente distrital del programa de Tuberculosis, quien revisará, organizará y entregará la información mediante las cuentas de usuario institucionales. Para surtir con eficiencia el proceso, lo invito a que se detalle la solicitud, mediante una reunión de trabajo, para tener las claridades correspondientes frente a la necesidad de la información.

La información por entregar surtirá los correspondientes procesos de anonimización con la correspondiente codificación represente la identidad de la persona titular de la información, garantizando la confidencialidad. Por lo anterior, en caso de requerir información faltante, se solicite su completitud o verificación al referente distrital del programa de Tuberculosis.

Copio esta comunicación a la Subdirectora de determinantes y la Subdirectora de Vigilancia para su información y conocimiento del proceso surtido.

Finalmente, resalto su interés en los aportes a la investigación, lo cuales la Secretaría Distrital de Salud, estará acompañando muy interesada de los resultados obtenidos, en pro de fortalecer y mejorar permanentemente los procesos y por ende el beneficio de la población.

Sin otro particular.

Cordialmente,

Carrera 32 No. 12 - 81
Teléfono: 3649090
www.saludcapital.gov.co

CO-SC-CER535793

ALCALDÍA MAYOR DE BOGOTÁ D.C.

BOGOTÁ | SECRETARÍA DE SALUD

Firmado digitalmente por: DIANE MOYANO ROMERO
Fecha: 10-03-2024
Hora: 22:56:47

DIANE MOYANO ROMERO

Anexo 6. Tabla de análisis Bivariado todas las condiciones de egreso del programa y determinantes sociales en salud.

La tabla contiene la distribución porcentual entre las condiciones de egreso del programa y el resto de las variables del conjunto de datos, tomado de script de Rstudio.

Variable	categoría	Curado (N=1422)	Descartado (N=526)	Excluido RR (N=93)	Fallecido (N=2172)	Fracaso (N=61)	No Evaluado (N=309)	Pérdida (N=738)	Tratamiento Terminado (N=4781)	Overall (N=10102)
INGRESO_ TTO	SI	1422 (100%)	464 (88.2%)	90 (96.8%)	1856 (85.5%)	58 (95.1%)	304 (98.4%)	702 (95.1%)	4773 (99.8%)	9669 (95.7%)
	NO	0 (0%)	62 (11.8%)	3 (3.2%)	316 (14.5%)	3 (4.9%)	5 (1.6%)	36 (4.9%)	8 (0.2%)	433 (4.3%)
SEXO	F	572 (40.2%)	176 (33.5%)	30 (32.3%)	603 (27.8%)	20 (32.8%)	73 (23.6%)	200 (27.1%)	1741 (36.4%)	3415 (33.8%)
	M	850 (59.8%)	350 (66.5%)	63 (67.7%)	1569 (72.2%)	41 (67.2%)	236 (76.4%)	538 (72.9%)	3040 (63.6%)	6687 (66.2%)
EDAD	Mean (SD)	53.0 (21.4)	47.2 (20.6)	53.0 (22.7)	57.5 (21.0)	51.2 (21.1)	44.7 (20.2)	40.8 (18.0)	48.3 (21.7)	50.3 (21.6)
	Median [Min, Max]	56.0 [0, 97.0]	45.0 [0, 94.0]	56.0 [7.00, 100]	61.0 [0, 109]	51.0 [1.00, 90.0]	40.0 [0, 99.0]	36.0 [0, 91.0]	47.0 [0, 98.0]	50.0 [0, 109]
REGIMEN_ AFILIACION	C	838 (58.9%)	303 (57.6%)	45 (48.4%)	1077 (49.6%)	31 (50.8%)	130 (42.1%)	224 (30.4%)	2951 (61.7%)	5599 (55.4%)
	E	75 (5.3%)	16 (3.0%)	3 (3.2%)	79 (3.6%)	2 (3.3%)	26 (8.4%)	28 (3.8%)	287 (6.0%)	516 (5.1%)
	N	55 (3.9%)	29 (5.5%)	4 (4.3%)	174 (8.0%)	5 (8.2%)	41 (13.3%)	134 (18.2%)	227 (4.7%)	669 (6.6%)
	P	73 (5.1%)	10 (1.9%)	0 (0%)	35 (1.6%)	2 (3.3%)	8 (2.6%)	7 (0.9%)	95 (2.0%)	230 (2.3%)
	S	381 (26.8%)	168 (31.9%)	41 (44.1%)	807 (37.2%)	21 (34.4%)	104 (33.7%)	345 (46.7%)	1221 (25.5%)	3088 (30.6%)
TIPO_TB	Extrapulmonar	4 (0.3%)	234 (44.5%)	21 (22.6%)	726 (33.4%)	12 (19.7%)	114 (36.9%)	171 (23.2%)	1880 (39.3%)	3162 (31.3%)
	Pulmonar	1418 (99.7%)	292 (55.5%)	72 (77.4%)	1446 (66.6%)	49 (80.3%)	195 (63.1%)	567 (76.8%)	2901 (60.7%)	6940 (68.7%)
CONDICION_ INGRESO	Nuevo	1356 (95.4%)	505 (96.0%)	68 (73.1%)	2052 (94.5%)	53 (86.9%)	283 (91.6%)	588 (79.7%)	4532 (94.8%)	9437 (93.4%)
	OPT	31 (2.2%)	19 (3.6%)	10 (10.8%)	59 (2.7%)	5 (8.2%)	14 (4.5%)	48 (6.5%)	131 (2.7%)	317 (3.1%)
	RTF	3 (0.2%)	0 (0%)	6 (6.5%)	3 (0.1%)	0 (0%)	1 (0.3%)	6 (0.8%)	6 (0.1%)	25 (0.2%)
	RTPS	22 (1.5%)	1 (0.2%)	6 (6.5%)	32 (1.5%)	0 (0%)	10 (3.2%)	82 (11.1%)	61 (1.3%)	214 (2.1%)
	RTR	10 (0.7%)	1 (0.2%)	3 (3.2%)	25 (1.2%)	3 (4.9%)	1 (0.3%)	14 (1.9%)	51 (1.1%)	108 (1.1%)
	Remitido	0 (0%)	0 (0%)	0 (0%)	1 (0.0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.0%)
CONDICION_VIH	Desconocido	30 (2.1%)	51 (9.7%)	5 (5.4%)	266	0 (0%)	21 (6.8%)	54 (7.3%)	134 (2.8%)	561 (5.6%)

Variable	categoría	Curado (N=1422)	Descartado (N=526)	Excluido RR (N=93)	Fallecido (N=2172)	Fracaso (N=61)	No Evaluado (N=309)	Pérdida (N=738)	Tratamiento Terminado (N=4781)	Overall (N=10102)
					(12.2%)					
	Negativo	1293 (90.9%)	298 (56.7%)	65 (69.9%)	1296 (59.7%)	52 (85.2%)	211 (68.3%)	445 (60.3%)	3885 (81.3%)	7545 (74.7%)
	Positivo	99 (7.0%)	177 (33.7%)	23 (24.7%)	610 (28.1%)	9 (14.8%)	77 (24.9%)	239 (32.4%)	762 (15.9%)	1996 (19.8%)
RESULTADO_ BK_ RECOD	Negativo	517 (36.4%)	378 (71.9%)	45 (48.4%)	1237 (57.0%)	22 (36.1%)	152 (49.2%)	378 (51.2%)	2722 (56.9%)	5451 (54.0%)
	NR	74 (5.2%)	94 (17.9%)	3 (3.2%)	325 (15.0%)	6 (9.8%)	45 (14.6%)	64 (8.7%)	799 (16.7%)	1410 (14.0%)
	Positivo	828 (58.2%)	39 (7.4%)	45 (48.4%)	582 (26.8%)	31 (50.8%)	93 (30.1%)	289 (39.2%)	1201 (25.1%)	3108 (30.8%)
	SD	3 (0.2%)	15 (2.9%)	0 (0%)	28 (1.3%)	2 (3.3%)	19 (6.1%)	7 (0.9%)	59 (1.2%)	133 (1.3%)
RESULTADO_ CULTIVO _ RECOD	Negativo	313 (22.0%)	320 (60.8%)	8 (8.6%)	682 (31.4%)	9 (14.8%)	91 (29.4%)	226 (30.6%)	1436 (30.0%)	3085 (30.5%)
	NR	249 (17.5%)	94 (17.9%)	17 (18.3%)	517 (23.8%)	6 (9.8%)	71 (23.0%)	134 (18.2%)	1180 (24.7%)	2268 (22.5%)
	Positivo	799 (56.2%)	61 (11.6%)	68 (73.1%)	844 (38.9%)	46 (75.4%)	114 (36.9%)	338 (45.8%)	1902 (39.8%)	4172 (41.3%)
	SD	61 (4.3%)	51 (9.7%)	0 (0%)	129 (5.9%)	0 (0%)	33 (10.7%)	40 (5.4%)	263 (5.5%)	577 (5.7%)
RESULTADO_ PRUEBA_ MOL_RECOD	Negativo	74 (5.2%)	311 (59.1%)	3 (3.2%)	335 (15.4%)	4 (6.6%)	54 (17.5%)	92 (12.5%)	652 (13.6%)	1525 (15.1%)
	NR	4 (0.3%)	0 (0%)	0 (0%)	17 (0.8%)	0 (0%)	0 (0%)	3 (0.4%)	21 (0.4%)	45 (0.4%)
	Positivo	747 (52.5%)	43 (8.2%)	60 (64.5%)	962 (44.3%)	56 (91.8%)	141 (45.6%)	379 (51.4%)	2183 (45.7%)	4571 (45.2%)
	SD	597 (42.0%)	171 (32.5%)	30 (32.3%)	856 (39.4%)	1 (1.6%)	114 (36.9%)	264 (35.8%)	1924 (40.2%)	3957 (39.2%)
	NI	0 (0%)	1 (0.2%)	0 (0%)	2 (0.1%)	0 (0%)	0 (0%)	0 (0%)	1 (0.0%)	4 (0.0%)
PRUEBA_ SUSCEPTIBILIDAD _FARMACOS	BACTEC MGIT	48 (3.4%)	0 (0%)	24 (25.8%)	39 (1.8%)	4 (6.6%)	10 (3.2%)	13 (1.8%)	133 (2.8%)	271 (2.7%)
	LIPA	189 (13.3%)	3 (0.6%)	29 (31.2%)	209 (9.6%)	7 (11.5%)	13 (4.2%)	99 (13.4%)	422 (8.8%)	971 (9.6%)
	NR	670 (47.1%)	361 (68.6%)	8 (8.6%)	1160 (53.4%)	1 (1.6%)	154 (49.8%)	321 (43.5%)	2560 (53.5%)	5235 (51.8%)
	PCR-TR	515 (36.2%)	162 (30.8%)	32 (34.4%)	764 (35.2%)	49 (80.3%)	132 (42.7%)	305 (41.3%)	1666 (34.8%)	3625 (35.9%)
FARMACORRESISTENCIA	Isoniacida	2 (0.1%)	0 (0%)	47 (50.5%)	7 (0.3%)	25 (41.0%)	0 (0%)	1 (0.1%)	6 (0.1%)	88 (0.9%)
	Monoresistencia	2 (0.1%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	0 (0%)	1 (0.1%)	1 (0.0%)	5 (0.0%)
	Ninguna	778 (54.7%)	165 (31.4%)	4 (4.3%)	1040 (47.9%)	8 (13.1%)	167 (54.0%)	421 (57.0%)	2315 (48.4%)	4898 (48.5%)
	NR	639 (44.9%)	361 (68.6%)	0 (0%)	1121 (51.6%)	1 (1.6%)	142 (46.0%)	313 (42.4%)	2459 (51.4%)	5036 (49.9%)
	RR	1 (0.1%)	0 (0%)	26 (28.0%)	3 (0.1%)	22 (36.1%)	0 (0%)	2 (0.3%)	0 (0%)	54 (0.5%)

Variable	categoría	Curado (N=1422)	Descartado (N=526)	Excluido RR (N=93)	Fallecido (N=2172)	Fracaso (N=61)	No Evaluado (N=309)	Pérdida (N=738)	Tratamiento Terminado (N=4781)	Overall (N=10102)
	MDR	0 (0%)	0 (0%)	14 (15.1%)	1 (0.0%)	5 (8.2%)	0 (0%)	0 (0%)	0 (0%)	20 (0.2%)
	Poliresistente	0 (0%)	0 (0%)	1 (1.1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.0%)
Alcoholismo	NO	1417 (99.6%)	521 (99.0%)	92 (98.9%)	2165 (99.7%)	61 (100%)	309 (100%)	734 (99.5%)	4770 (99.8%)	10069 (99.7%)
	SI	5 (0.4%)	5 (1.0%)	1 (1.1%)	7 (0.3%)	0 (0%)	0 (0%)	4 (0.5%)	11 (0.2%)	33 (0.3%)
Cancer	NO	1373 (96.6%)	473 (89.9%)	92 (98.9%)	2006 (92.4%)	58 (95.1%)	291 (94.2%)	710 (96.2%)	4537 (94.9%)	9540 (94.4%)
	SI	49 (3.4%)	53 (10.1%)	1 (1.1%)	166 (7.6%)	3 (4.9%)	18 (5.8%)	28 (3.8%)	244 (5.1%)	562 (5.6%)
Cardiovascular	NO	1415 (99.5%)	526 (100%)	91 (97.8%)	2167 (99.8%)	61 (100%)	309 (100%)	738 (100%)	4773 (99.8%)	10080 (99.8%)
	SI	7 (0.5%)	0 (0%)	2 (2.2%)	5 (0.2%)	0 (0%)	0 (0%)	0 (0%)	8 (0.2%)	22 (0.2%)
Consumidor_ SPA	NO	1415 (99.5%)	518 (98.5%)	90 (96.8%)	2146 (98.8%)	60 (98.4%)	304 (98.4%)	689 (93.4%)	4756 (99.5%)	9978 (98.8%)
	SI	7 (0.5%)	8 (1.5%)	3 (3.2%)	26 (1.2%)	1 (1.6%)	5 (1.6%)	49 (6.6%)	25 (0.5%)	124 (1.2%)
Desnutrición	NO	1249 (87.8%)	396 (75.3%)	83 (89.2%)	1689 (77.8%)	47 (77.0%)	256 (82.8%)	576 (78.0%)	4046 (84.6%)	8342 (82.6%)
	SI	173 (12.2%)	130 (24.7%)	10 (10.8%)	483 (22.2%)	14 (23.0%)	53 (17.2%)	162 (22.0%)	735 (15.4%)	1760 (17.4%)
Diabetes	NO	1283 (90.2%)	480 (91.3%)	88 (94.6%)	1939 (89.3%)	53 (86.9%)	288 (93.2%)	698 (94.6%)	4367 (91.3%)	9196 (91.0%)
	SI	139 (9.8%)	46 (8.7%)	5 (5.4%)	233 (10.7%)	8 (13.1%)	21 (6.8%)	40 (5.4%)	414 (8.7%)	906 (9.0%)
Enf_Mental	NO	1422 (100%)	526 (100%)	93 (100%)	2168 (99.8%)	61 (100%)	309 (100%)	734 (99.5%)	4778 (99.9%)	10091 (99.9%)
	SI	0 (0%)	0 (0%)	0 (0%)	4 (0.2%)	0 (0%)	0 (0%)	4 (0.5%)	3 (0.1%)	11 (0.1%)
Enf_Autoinmune	NO	1408 (99.0%)	501 (95.2%)	92 (98.9%)	2113 (97.3%)	61 (100%)	299 (96.8%)	723 (98.0%)	4650 (97.3%)	9847 (97.5%)
	SI	14 (1.0%)	25 (4.8%)	1 (1.1%)	59 (2.7%)	0 (0%)	10 (3.2%)	15 (2.0%)	131 (2.7%)	255 (2.5%)
Enf_Hepatica	NO	1407 (98.9%)	515 (97.9%)	92 (98.9%)	2110 (97.1%)	61 (100%)	306 (99.0%)	728 (98.6%)	4721 (98.7%)	9940 (98.4%)
	SI	15 (1.1%)	11 (2.1%)	1 (1.1%)	62 (2.9%)	0 (0%)	3 (1.0%)	10 (1.4%)	60 (1.3%)	162 (1.6%)
Enf_Renal	NO	1326 (93.2%)	482 (91.6%)	89 (95.7%)	1935 (89.1%)	53 (86.9%)	289 (93.5%)	707 (95.8%)	4405 (92.1%)	9286 (91.9%)
	SI	96 (6.8%)	44 (8.4%)	4 (4.3%)	237 (10.9%)	8 (13.1%)	20 (6.5%)	31 (4.2%)	376 (7.9%)	816 (8.1%)
EPOC	NO	1294 (91.0%)	460 (87.5%)	88 (94.6%)	1855 (85.4%)	50 (82.0%)	280 (90.6%)	692 (93.8%)	4301 (90.0%)	9020 (89.3%)
	SI	128 (9.0%)	66 (12.5%)	5 (5.4%)	317 (14.6%)	11 (18.0%)	29 (9.4%)	46 (6.2%)	480 (10.0%)	1082 (10.7%)
Silicosis	NO	1398 (98.3%)	519 (98.7%)	93 (100%)	2147 (98.8%)	60 (98.4%)	302 (97.7%)	731 (99.1%)	4715 (98.6%)	9965 (98.6%)
	SI	24 (1.7%)	7 (1.3%)	0 (0%)	25 (1.2%)	1 (1.6%)	7 (2.3%)	7 (0.9%)	66 (1.4%)	137 (1.4%)
Tabaquismo	NO	1418	515	92	2159	60 (98.4%)	305	726	4748 (99.3%)	10023

Variable	categoría	Curado (N=1422)	Descartado (N=526)	Excluido RR (N=93)	Fallecido (N=2172)	Fracaso (N=61)	No Evaluado (N=309)	Pérdida (N=738)	Tratamiento Terminado (N=4781)	Overall (N=10102)
		(99.7%)	(97.9%)	(98.9%)	(99.4%)		(98.7%)	(98.4%)		(99.2%)
	SI	4 (0.3%)	11 (2.1%)	1 (1.1%)	13 (0.6%)	1 (1.6%)	4 (1.3%)	12 (1.6%)	33 (0.7%)	79 (0.8%)
Hipotiroidismo	NO	1270 (89.3%)	459 (87.3%)	88 (94.6%)	1901 (87.5%)	50 (82.0%)	273 (88.3%)	667 (90.4%)	4307 (90.1%)	9015 (89.2%)
	SI	152 (10.7%)	67 (12.7%)	5 (5.4%)	271 (12.5%)	11 (18.0%)	36 (11.7%)	71 (9.6%)	474 (9.9%)	1087 (10.8%)
Otra_Enf	NO	1292 (90.9%)	508 (96.6%)	68 (73.1%)	1853 (85.3%)	61 (100%)	283 (91.6%)	658 (89.2%)	4305 (90.0%)	9028 (89.4%)
	SI	130 (9.1%)	18 (3.4%)	25 (26.9%)	319 (14.7%)	0 (0%)	26 (8.4%)	80 (10.8%)	476 (10.0%)	1074 (10.6%)
MODALIDAD_ TDO	IPS	19 (1.3%)	6 (1.1%)	0 (0%)	4 (0.2%)	0 (0%)	3 (1.0%)	3 (0.4%)	33 (0.7%)	68 (0.7%)
	No Evaluado	883 (62.1%)	120 (22.8%)	92 (98.9%)	1250 (57.6%)	2 (3.3%)	120 (38.8%)	396 (53.7%)	2709 (56.7%)	5572 (55.2%)
	TDO comunitario	7 (0.5%)	1 (0.2%)	0 (0%)	4 (0.2%)	0 (0%)	2 (0.6%)	6 (0.8%)	47 (1.0%)	67 (0.7%)
	TDO Domiciliario	7 (0.5%)	4 (0.8%)	0 (0%)	8 (0.4%)	1 (1.6%)	2 (0.6%)	3 (0.4%)	31 (0.6%)	56 (0.6%)
	TDO en IPS	457 (32.1%)	246 (46.8%)	1 (1.1%)	494 (22.7%)	37 (60.7%)	128 (41.4%)	271 (36.7%)	1779 (37.2%)	3413 (33.8%)
	TDO hospitalario	24 (1.7%)	142 (27.0%)	0 (0%)	406 (18.7%)	19 (31.1%)	47 (15.2%)	51 (6.9%)	102 (2.1%)	791 (7.8%)
	TDO virtual	25 (1.8%)	7 (1.3%)	0 (0%)	6 (0.3%)	2 (3.3%)	7 (2.3%)	8 (1.1%)	80 (1.7%)	135 (1.3%)
PROGRAMAS_ PROTECC_SOCIAL	Alimentario	21 (1.5%)	1 (0.2%)	0 (0%)	6 (0.3%)	2 (3.3%)	1 (0.3%)	14 (1.9%)	48 (1.0%)	93 (0.9%)
	Educativo	1 (0.1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	6 (0.1%)	7 (0.1%)
	Monetario	2 (0.1%)	3 (0.6%)	0 (0%)	0 (0%)	0 (0%)	1 (0.3%)	2 (0.3%)	17 (0.4%)	25 (0.2%)
	NA	301 (21.2%)	192 (36.5%)	1 (1.1%)	435 (20.0%)	35 (57.4%)	95 (30.7%)	145 (19.6%)	906 (19.0%)	2110 (20.9%)
	Ninguno	203 (14.3%)	220 (41.8%)	0 (0%)	485 (22.3%)	19 (31.1%)	87 (28.2%)	173 (23.4%)	1051 (22.0%)	2238 (22.2%)
	No evaluado	883 (62.1%)	109 (20.7%)	92 (98.9%)	1239 (57.0%)	2 (3.3%)	118 (38.2%)	391 (53.0%)	2709 (56.7%)	5543 (54.9%)
	Subsidio de vivienda	4 (0.3%)	0 (0%)	0 (0%)	2 (0.1%)	1 (1.6%)	1 (0.3%)	4 (0.5%)	18 (0.4%)	30 (0.3%)
	Varios	7 (0.5%)	1 (0.2%)	0 (0%)	5 (0.2%)	2 (3.3%)	6 (1.9%)	9 (1.2%)	21 (0.4%)	51 (0.5%)
	Desempleo	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (0.1%)	3 (0.0%)
	Transporte	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (0.0%)	2 (0.0%)
REACCIONES_ ADVERSAS_TTO	Grave	2 (0.1%)	2 (0.4%)	0 (0%)	5 (0.2%)	0 (0%)	1 (0.3%)	3 (0.4%)	6 (0.1%)	19 (0.2%)
	Leve	2 (0.1%)	0 (0%)	0 (0%)	2 (0.1%)	0 (0%)	0 (0%)	1 (0.1%)	7 (0.1%)	12 (0.1%)
	Ninguna	535 (37.6%)	414 (78.7%)	1 (1.1%)	927 (42.7%)	59 (96.7%)	190 (61.5%)	341 (46.2%)	2053 (42.9%)	4520 (44.7%)
	SD	883 (62.1%)	109 (20.7%)	92 (98.9%)	1235 (56.9%)	2 (3.3%)	118 (38.2%)	389 (52.7%)	2705 (56.6%)	5533 (54.8%)
	Moderada	0 (0%)	1 (0.2%)	0 (0%)	3 (0.1%)	0 (0%)	0 (0%)	4 (0.5%)	10 (0.2%)	18 (0.2%)
METODOLOGIA_ CAPTACION	BAI	537 (37.8%)	411 (78.1%)	1 (1.1%)	924 (42.5%)	59 (96.7%)	191 (61.8%)	346 (46.9%)	2043 (42.7%)	4512 (44.7%)
	BTS.	2 (0.1%)	5 (1.0%)	0 (0%)	8 (0.4%)	0 (0%)	0 (0%)	1 (0.1%)	26 (0.5%)	42 (0.4%)

Variable	categoría	Curado (N=1422)	Descartado (N=526)	Excluido RR (N=93)	Fallecido (N=2172)	Fracaso (N=61)	No Evaluado (N=309)	Pérdida (N=738)	Tratamiento Terminado (N=4781)	Overall (N=10102)
	SD	883 (62.1%)	109 (20.7%)	92 (98.9%)	1239 (57.0%)	2 (3.3%)	118 (38.2%)	391 (53.0%)	2709 (56.7%)	5543 (54.9%)
	Contactos	0 (0%)	1 (0.2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (0.1%)	4 (0.0%)
	CNE	0 (0%)	0 (0%)	0 (0%)	1 (0.0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.0%)
PERTENENCIA_ ETNICA	Indígena	13 (0.9%)	2 (0.4%)	1 (1.1%)	36 (1.7%)	1 (1.6%)	13 (4.2%)	22 (3.0%)	65 (1.4%)	153 (1.5%)
	Negro, Mulato, Afrocolombiano	8 (0.6%)	2 (0.4%)	3 (3.2%)	13 (0.6%)	1 (1.6%)	5 (1.6%)	12 (1.6%)	38 (0.8%)	82 (0.8%)
	Otro	1400 (98.5%)	520 (98.9%)	89 (95.7%)	2120 (97.6%)	59 (96.7%)	291 (94.2%)	702 (95.1%)	4665 (97.6%)	9846 (97.5%)
	Room (Gitano)	1 (0.1%)	1 (0.2%)	0 (0%)	2 (0.1%)	0 (0%)	0 (0%)	1 (0.1%)	8 (0.2%)	13 (0.1%)
	Raizal	0 (0%)	1 (0.2%)	0 (0%)	1 (0.0%)	0 (0%)	0 (0%)	1 (0.1%)	4 (0.1%)	7 (0.1%)
	Palenquero	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.0%)	1 (0.0%)
gp_discapa	NO	1407 (98.9%)	518 (98.5%)	91 (97.8%)	2133 (98.2%)	61 (100%)	305 (98.7%)	733 (99.3%)	4719 (98.7%)	9967 (98.7%)
	SI	15 (1.1%)	8 (1.5%)	2 (2.2%)	39 (1.8%)	0 (0%)	4 (1.3%)	5 (0.7%)	62 (1.3%)	135 (1.3%)
gp_desplaz	NO	1409 (99.1%)	525 (99.8%)	91 (97.8%)	2160 (99.4%)	60 (98.4%)	303 (98.1%)	722 (97.8%)	4753 (99.4%)	10023 (99.2%)
	SI	13 (0.9%)	1 (0.2%)	2 (2.2%)	12 (0.6%)	1 (1.6%)	6 (1.9%)	16 (2.2%)	28 (0.6%)	79 (0.8%)
gp_migrant	NO	1366 (96.1%)	489 (93.0%)	91 (97.8%)	2065 (95.1%)	56 (91.8%)	266 (86.1%)	663 (89.8%)	4583 (95.9%)	9579 (94.8%)
	SI	56 (3.9%)	37 (7.0%)	2 (2.2%)	107 (4.9%)	5 (8.2%)	43 (13.9%)	75 (10.2%)	198 (4.1%)	523 (5.2%)
gp_carcela	NO	1321 (92.9%)	519 (98.7%)	92 (98.9%)	2144 (98.7%)	59 (96.7%)	300 (97.1%)	706 (95.7%)	4622 (96.7%)	9763 (96.6%)
	SI	101 (7.1%)	7 (1.3%)	1 (1.1%)	28 (1.3%)	2 (3.3%)	9 (2.9%)	32 (4.3%)	159 (3.3%)	339 (3.4%)
gp_gestan	NO	1417 (99.6%)	526 (100%)	93 (100%)	2170 (99.9%)	61 (100%)	307 (99.4%)	734 (99.5%)	4774 (99.9%)	10082 (99.8%)
	SI	5 (0.4%)	0 (0%)	0 (0%)	2 (0.1%)	0 (0%)	2 (0.6%)	4 (0.5%)	7 (0.1%)	20 (0.2%)
gp_indigen	NO	1374 (96.6%)	512 (97.3%)	80 (86.0%)	2038 (93.8%)	56 (91.8%)	302 (97.7%)	534 (72.4%)	4665 (97.6%)	9561 (94.6%)
	SI	48 (3.4%)	14 (2.7%)	13 (14.0%)	134 (6.2%)	5 (8.2%)	7 (2.3%)	204 (27.6%)	116 (2.4%)	541 (5.4%)
gp_pobicbf	NO	1420 (99.9%)	525 (99.8%)	93 (100%)	2170 (99.9%)	61 (100%)	309 (100%)	736 (99.7%)	4774 (99.9%)	10088 (99.9%)
	SI	2 (0.1%)	1 (0.2%)	0 (0%)	2 (0.1%)	0 (0%)	0 (0%)	2 (0.3%)	7 (0.1%)	14 (0.1%)
gp_psiquia	NO	1422 (100%)	526 (100%)	93 (100%)	2168 (99.8%)	61 (100%)	309 (100%)	734 (99.5%)	4778 (99.9%)	10091 (99.9%)
	SI	0 (0%)	0 (0%)	0 (0%)	4 (0.2%)	0 (0%)	0 (0%)	4 (0.5%)	3 (0.1%)	11 (0.1%)
gp_vic_vio	NO	1421 (99.9%)	526 (100%)	93 (100%)	2169 (99.9%)	61 (100%)	308 (99.7%)	734 (99.5%)	4774 (99.9%)	10086 (99.8%)
	SI	1 (0.1%)	0 (0%)	0 (0%)	3 (0.1%)	0 (0%)	1 (0.3%)	4 (0.5%)	7 (0.1%)	16 (0.2%)
trabajador_ salud	NO	1397	515	88	2163	59 (96.7%)	306	730	4640 (97.1%)	9898

Variable	categoría	Curado (N=1422)	Descartado (N=526)	Excluido RR (N=93)	Fallecido (N=2172)	Fracaso (N=61)	No Evaluado (N=309)	Pérdida (N=738)	Tratamiento Terminado (N=4781)	Overall (N=10102)
		(98.2%)	(97.9%)	(94.6%)	(99.6%)		(99.0%)	(98.9%)		(98.0%)
	SI	25 (1.8%)	11 (2.1%)	5 (5.4%)	9 (0.4%)	2 (3.3%)	3 (1.0%)	8 (1.1%)	141 (2.9%)	204 (2.0%)
gp_otros	NO	176 (12.4%)	50 (9.5%)	9 (9.7%)	234 (10.8%)	10 (16.4%)	52 (16.8%)	247 (33.5%)	433 (9.1%)	1211 (12.0%)
	SI	1246 (87.6%)	476 (90.5%)	84 (90.3%)	1938 (89.2%)	51 (83.6%)	257 (83.2%)	491 (66.5%)	4348 (90.9%)	8891 (88.0%)
LOC_RES	1	48 (3.4%)	29 (5.5%)	7 (7.5%)	61 (2.8%)	2 (3.3%)	14 (4.5%)	17 (2.3%)	217 (4.5%)	395 (3.9%)
	10	113 (7.9%)	38 (7.2%)	11 (11.8%)	138 (6.4%)	3 (4.9%)	19 (6.1%)	55 (7.5%)	359 (7.5%)	736 (7.3%)
	11	155 (10.9%)	64 (12.2%)	3 (3.2%)	194 (8.9%)	6 (9.8%)	27 (8.7%)	54 (7.3%)	487 (10.2%)	990 (9.8%)
	12	29 (2.0%)	15 (2.9%)	1 (1.1%)	29 (1.3%)	0 (0%)	7 (2.3%)	10 (1.4%)	81 (1.7%)	172 (1.7%)
	13	10 (0.7%)	5 (1.0%)	1 (1.1%)	32 (1.5%)	2 (3.3%)	4 (1.3%)	6 (0.8%)	73 (1.5%)	133 (1.3%)
	14	32 (2.3%)	6 (1.1%)	5 (5.4%)	55 (2.5%)	4 (6.6%)	6 (1.9%)	44 (6.0%)	100 (2.1%)	252 (2.5%)
	15	8 (0.6%)	9 (1.7%)	2 (2.2%)	42 (1.9%)	0 (0%)	2 (0.6%)	24 (3.3%)	62 (1.3%)	149 (1.5%)
	16	96 (6.8%)	14 (2.7%)	3 (3.2%)	70 (3.2%)	1 (1.6%)	6 (1.9%)	28 (3.8%)	143 (3.0%)	361 (3.6%)
	17	6 (0.4%)	1 (0.2%)	2 (2.2%)	12 (0.6%)	0 (0%)	1 (0.3%)	4 (0.5%)	14 (0.3%)	40 (0.4%)
	18	129 (9.1%)	14 (2.7%)	7 (7.5%)	115 (5.3%)	2 (3.3%)	7 (2.3%)	39 (5.3%)	347 (7.3%)	660 (6.5%)
	19	80 (5.6%)	35 (6.7%)	7 (7.5%)	141 (6.5%)	2 (3.3%)	10 (3.2%)	71 (9.6%)	321 (6.7%)	667 (6.6%)
	2	17 (1.2%)	9 (1.7%)	2 (2.2%)	31 (1.4%)	0 (0%)	4 (1.3%)	14 (1.9%)	61 (1.3%)	138 (1.4%)
	3	39 (2.7%)	9 (1.7%)	5 (5.4%)	59 (2.7%)	0 (0%)	13 (4.2%)	45 (6.1%)	118 (2.5%)	288 (2.9%)
	4	59 (4.1%)	25 (4.8%)	6 (6.5%)	119 (5.5%)	1 (1.6%)	8 (2.6%)	25 (3.4%)	250 (5.2%)	493 (4.9%)
	5	36 (2.5%)	14 (2.7%)	2 (2.2%)	90 (4.1%)	0 (0%)	2 (0.6%)	20 (2.7%)	171 (3.6%)	335 (3.3%)
	6	18 (1.3%)	8 (1.5%)	0 (0%)	55 (2.5%)	0 (0%)	2 (0.6%)	19 (2.6%)	102 (2.1%)	204 (2.0%)
	7	118 (8.3%)	30 (5.7%)	3 (3.2%)	93 (4.3%)	4 (6.6%)	7 (2.3%)	36 (4.9%)	286 (6.0%)	577 (5.7%)
	8	164 (11.5%)	53 (10.1%)	5 (5.4%)	171 (7.9%)	6 (9.8%)	20 (6.5%)	59 (8.0%)	426 (8.9%)	904 (8.9%)
	9	84 (5.9%)	20 (3.8%)	3 (3.2%)	48 (2.2%)	3 (4.9%)	7 (2.3%)	10 (1.4%)	145 (3.0%)	320 (3.2%)
	FDB	152 (10.7%)	111 (21.1%)	13 (14.0%)	509 (23.4%)	19 (31.1%)	131 (42.4%)	82 (11.1%)	925 (19.3%)	1942 (19.2%)
	Sin Dato	29 (2.0%)	17 (3.2%)	5 (5.4%)	105 (4.8%)	6 (9.8%)	12 (3.9%)	75 (10.2%)	93 (1.9%)	342 (3.4%)
	20	0 (0%)	0 (0%)	0 (0%)	3 (0.1%)	0 (0%)	0 (0%)	1 (0.1%)	0 (0%)	4 (0.0%)
LOC_DX	1	105 (7.4%)	121 (23.0%)	8 (8.6%)	225 (10.4%)	12 (19.7%)	42 (13.6%)	68 (9.2%)	630 (13.2%)	1211 (12.0%)
	10	69 (4.9%)	22 (4.2%)	4 (4.3%)	65 (3.0%)	0 (0%)	14 (4.5%)	38 (5.1%)	159 (3.3%)	371 (3.7%)
	11	101 (7.1%)	16 (3.0%)	2 (2.2%)	75 (3.5%)	1 (1.6%)	12 (3.9%)	19 (2.6%)	212 (4.4%)	438 (4.3%)
	12	57 (4.0%)	9 (1.7%)	1 (1.1%)	77 (3.5%)	4 (6.6%)	7 (2.3%)	32 (4.3%)	186 (3.9%)	373 (3.7%)
	13	103 (7.2%)	68 (12.9%)	9 (9.7%)	215 (9.9%)	5 (8.2%)	38 (12.3%)	26 (3.5%)	553 (11.6%)	1017 (10.1%)
	14	117	48 (9.1%)	9 (9.7%)	302	4 (6.6%)	34 (11.0%)	74	555 (11.6%)	1143

Variable	categoría	Curado (N=1422)	Descartado (N=526)	Excluido RR (N=93)	Fallecido (N=2172)	Fracaso (N=61)	No Evaluado (N=309)	Pérdida (N=738)	Tratamiento Terminado (N=4781)	Overall (N=10102)
		(8.2%)			(13.9%)			(10.0%)		(11.3%)
	15	75 (5.3%)	17 (3.2%)	10 (10.8%)	202 (9.3%)	2 (3.3%)	22 (7.1%)	135 (18.3%)	312 (6.5%)	775 (7.7%)
	16	78 (5.5%)	4 (0.8%)	2 (2.2%)	5 (0.2%)	0 (0%)	1 (0.3%)	13 (1.8%)	49 (1.0%)	152 (1.5%)
	17	5 (0.4%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (0.3%)	2 (0.0%)	9 (0.1%)
	18	96 (6.8%)	25 (4.8%)	4 (4.3%)	104 (4.8%)	2 (3.3%)	4 (1.3%)	31 (4.2%)	320 (6.7%)	586 (5.8%)
	19	35 (2.5%)	7 (1.3%)	6 (6.5%)	65 (3.0%)	0 (0%)	0 (0%)	35 (4.7%)	81 (1.7%)	229 (2.3%)
	2	141 (9.9%)	75 (14.3%)	16 (17.2%)	234 (10.8%)	8 (13.1%)	65 (21.0%)	56 (7.6%)	678 (14.2%)	1273 (12.6%)
	3	16 (1.1%)	2 (0.4%)	2 (2.2%)	13 (0.6%)	0 (0%)	1 (0.3%)	25 (3.4%)	35 (0.7%)	94 (0.9%)
	4	98 (6.9%)	32 (6.1%)	7 (7.5%)	260 (12.0%)	10 (16.4%)	38 (12.3%)	58 (7.9%)	448 (9.4%)	951 (9.4%)
	5	6 (0.4%)	0 (0%)	0 (0%)	1 (0.0%)	0 (0%)	0 (0%)	2 (0.3%)	14 (0.3%)	23 (0.2%)
	6	39 (2.7%)	22 (4.2%)	9 (9.7%)	122 (5.6%)	0 (0%)	1 (0.3%)	33 (4.5%)	135 (2.8%)	361 (3.6%)
	7	48 (3.4%)	1 (0.2%)	0 (0%)	5 (0.2%)	0 (0%)	1 (0.3%)	11 (1.5%)	36 (0.8%)	102 (1.0%)
	8	162 (11.4%)	54 (10.3%)	3 (3.2%)	179 (8.2%)	7 (11.5%)	20 (6.5%)	63 (8.5%)	318 (6.7%)	806 (8.0%)
	9	66 (4.6%)	3 (0.6%)	1 (1.1%)	18 (0.8%)	6 (9.8%)	9 (2.9%)	15 (2.0%)	51 (1.1%)	169 (1.7%)
	Sin Dato	5 (0.4%)	0 (0%)	0 (0%)	5 (0.2%)	0 (0%)	0 (0%)	2 (0.3%)	7 (0.1%)	19 (0.2%)
LOCALIZACION_EXTRA	Genitourinaria	1 (0.1%)	0 (0%)	0 (0%)	7 (0.3%)	1 (1.6%)	4 (1.3%)	0 (0%)	70 (1.5%)	83 (0.8%)
	NA	1418 (99.7%)	292 (55.5%)	72 (77.4%)	1446 (66.6%)	49 (80.3%)	195 (63.1%)	567 (76.8%)	2901 (60.7%)	6940 (68.7%)
	Pleural	3 (0.2%)	51 (9.7%)	5 (5.4%)	182 (8.4%)	1 (1.6%)	36 (11.7%)	46 (6.2%)	699 (14.6%)	1023 (10.1%)
	Cutanea	0 (0%)	7 (1.3%)	0 (0%)	5 (0.2%)	0 (0%)	1 (0.3%)	3 (0.4%)	33 (0.7%)	49 (0.5%)
	Ganglionar	0 (0%)	13 (2.5%)	5 (5.4%)	42 (1.9%)	2 (3.3%)	12 (3.9%)	34 (4.6%)	258 (5.4%)	366 (3.6%)
	Intestinal	0 (0%)	4 (0.8%)	1 (1.1%)	26 (1.2%)	1 (1.6%)	2 (0.6%)	2 (0.3%)	46 (1.0%)	82 (0.8%)
	Meningea	0 (0%)	134 (25.5%)	4 (4.3%)	315 (14.5%)	4 (6.6%)	35 (11.3%)	52 (7.0%)	304 (6.4%)	848 (8.4%)
	Osteoarticular	0 (0%)	3 (0.6%)	3 (3.2%)	21 (1.0%)	1 (1.6%)	17 (5.5%)	13 (1.8%)	151 (3.2%)	209 (2.1%)
	Otro	0 (0%)	12 (2.3%)	1 (1.1%)	62 (2.9%)	2 (3.3%)	3 (1.0%)	8 (1.1%)	135 (2.8%)	223 (2.2%)
	Pericardica	0 (0%)	3 (0.6%)	0 (0%)	19 (0.9%)	0 (0%)	3 (1.0%)	6 (0.8%)	80 (1.7%)	111 (1.1%)
	Peritoneal	0 (0%)	6 (1.1%)	2 (2.2%)	44 (2.0%)	0 (0%)	1 (0.3%)	5 (0.7%)	90 (1.9%)	148 (1.5%)
	Renal	0 (0%)	1 (0.2%)	0 (0%)	2 (0.1%)	0 (0%)	0 (0%)	2 (0.3%)	14 (0.3%)	19 (0.2%)
	Laringea	0 (0%)	0 (0%)	0 (0%)	1 (0.0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.0%)

Anexo 7. Tabla de análisis Bivariado condición de egreso pérdida del seguimiento y determinantes sociales en salud.

En la tabla se encuentra la distribución porcentual entre la pérdida de seguimiento y el resto de las variables del conjunto de datos generado con Rstudio.

Variable	Categoría	NO (N=9364)	SI (N=738)	Overall (N=10102)
INGRESO_TTO	NO	397 (4.2%)	36 (4.9%)	433 (4.3%)
	SI	8967 (95.8%)	702 (95.1%)	9669 (95.7%)
SEXO	F	3215 (34.3%)	200 (27.1%)	3415 (33.8%)
	M	6149 (65.7%)	538 (72.9%)	6687 (66.2%)
EDAD	Mean (SD)	51.0 (21.7)	40.8 (18.0)	50.3 (21.6)
	Median [Min, Max]	52.0 [0, 109]	36.0 [0, 91.0]	50.0 [0, 109]
REGIMEN_AFILIACION	C - CONTRIBUTIVO	5375 (57.4%)	224 (30.4%)	5599 (55.4%)
	E - ESPECIAL	488 (5.2%)	28 (3.8%)	516 (5.1%)
	N - NO ASEGURADO	535 (5.7%)	134 (18.2%)	669 (6.6%)
	P - EXCEPCION	223 (2.4%)	7 (0.9%)	230 (2.3%)
	S - SUBSIDIADO	2743 (29.3%)	345 (46.7%)	3088 (30.6%)
TIPO_TB	EXTRAPULMONAR	2991 (31.9%)	171 (23.2%)	3162 (31.3%)
	PULMONAR	6373 (68.1%)	567 (76.8%)	6940 (68.7%)
CONDICION_INGRESO	NUEVO	8849 (94.5%)	588 (79.7%)	9437 (93.4%)
	OTROS PREVIAMENTE TRATADOS	269 (2.9%)	48 (6.5%)	317 (3.1%)
	REINGRESO TRAS FRACASO	19 (0.2%)	6 (0.8%)	25 (0.2%)
	REINGRESO TRAS PÉRDIDA EN EL SEGUIMIENTO	132 (1.4%)	82 (11.1%)	214 (2.1%)
	REINGRESO TRAS RECAIDA	94 (1.0%)	14 (1.9%)	108 (1.1%)
	REMITIDO	1 (0.0%)	0 (0%)	1 (0.0%)
CONDICION_VIH	DESCONOCIDO	507 (5.4%)	54 (7.3%)	561 (5.6%)
	NEGATIVO	7100 (75.8%)	445 (60.3%)	7545 (74.7%)
	POSITIVO	1757 (18.8%)	239 (32.4%)	1996 (19.8%)
RESULTADO_BK_RECOD	NEGATIVO	5073 (54.2%)	378 (51.2%)	5451 (54.0%)
	NO REALIZADO	1346 (14.4%)	64 (8.7%)	1410 (14.0%)
	POSITIVO	2819 (30.1%)	289 (39.2%)	3108 (30.8%)
	SD	126 (1.3%)	7 (0.9%)	133 (1.3%)
RESULTADO_CULTIVO_REC	NEGATIVO	2859	226 (30.6%)	3085 (30.5%)

Variable	Categoría	NO (N=9364)	SI (N=738)	Overall (N=10102)
OD		(30.5%)		
	NO REALIZADO	2134 (22.8%)	134 (18.2%)	2268 (22.5%)
	POSITIVO	3834 (40.9%)	338 (45.8%)	4172 (41.3%)
	SD	537 (5.7%)	40 (5.4%)	577 (5.7%)
RESULTADO_PRUEBA_MOL_RECOD	NEGATIVO	1433 (15.3%)	92 (12.5%)	1525 (15.1%)
	NO INTERPRETABLE	4 (0.0%)	0 (0%)	4 (0.0%)
	NO REALIZADO	42 (0.4%)	3 (0.4%)	45 (0.4%)
	POSITIVO	4192 (44.8%)	379 (51.4%)	4571 (45.2%)
	SD	3693 (39.4%)	264 (35.8%)	3957 (39.2%)
PRUEBA_SUSCEPTIBILIDAD_FARMACOS	BACTEC MGIT	258 (2.8%)	13 (1.8%)	271 (2.7%)
	LIPA	872 (9.3%)	99 (13.4%)	971 (9.6%)
	NO REALIZADA	4914 (52.5%)	321 (43.5%)	5235 (51.8%)
	PCR EN TIEMPO REAL	3320 (35.5%)	305 (41.3%)	3625 (35.9%)
FARMACORRESISTENCIA	Isoniacida	87 (0.9%)	1 (0.1%)	88 (0.9%)
	MDR	20 (0.2%)	0 (0%)	20 (0.2%)
	Monoresistencia	4 (0.0%)	1 (0.1%)	5 (0.0%)
	Ninguna	4477 (47.8%)	421 (57.0%)	4898 (48.5%)
	NO REALIZADA	4723 (50.4%)	313 (42.4%)	5036 (49.9%)
	Poliresistente	1 (0.0%)	0 (0%)	1 (0.0%)
	RR	52 (0.6%)	2 (0.3%)	54 (0.5%)
Alcoholismo	NO	9335 (99.7%)	734 (99.5%)	10069 (99.7%)
	SI	29 (0.3%)	4 (0.5%)	33 (0.3%)
Cancer	NO	8830 (94.3%)	710 (96.2%)	9540 (94.4%)
	SI	534 (5.7%)	28 (3.8%)	562 (5.6%)
Cardiovascular	NO	9342 (99.8%)	738 (100%)	10080 (99.8%)
	SI	22 (0.2%)	0 (0%)	22 (0.2%)
Consumidor_SPA	NO	9289 (99.2%)	689 (93.4%)	9978 (98.8%)
	SI	75 (0.8%)	49 (6.6%)	124 (1.2%)
Desnutricion	NO	7766 (82.9%)	576 (78.0%)	8342 (82.6%)
	SI	1598 (17.1%)	162 (22.0%)	1760 (17.4%)
Diabetes	NO	8498 (90.8%)	698 (94.6%)	9196 (91.0%)
	SI	866 (9.2%)	40 (5.4%)	906 (9.0%)
Enf_Mental	NO	9357 (99.9%)	734 (99.5%)	10091 (99.9%)
	SI	7 (0.1%)	4 (0.5%)	11 (0.1%)
Enf_Autoinmune	NO	9124 (97.4%)	723 (98.0%)	9847 (97.5%)
	SI	240 (2.6%)	15 (2.0%)	255 (2.5%)
Enf_Hepatica	NO	9212 (98.4%)	728 (98.6%)	9940 (98.4%)

Variable	Categoría	NO (N=9364)	SI (N=738)	Overall (N=10102)
	SI	152 (1.6%)	10 (1.4%)	162 (1.6%)
Enf_Renal	NO	8579 (91.6%)	707 (95.8%)	9286 (91.9%)
	SI	785 (8.4%)	31 (4.2%)	816 (8.1%)
EPOC	NO	8328 (88.9%)	692 (93.8%)	9020 (89.3%)
	SI	1036 (11.1%)	46 (6.2%)	1082 (10.7%)
Silicosis	NO	9234 (98.6%)	731 (99.1%)	9965 (98.6%)
	SI	130 (1.4%)	7 (0.9%)	137 (1.4%)
Tabaquismo	NO	9297 (99.3%)	726 (98.4%)	10023 (99.2%)
	SI	67 (0.7%)	12 (1.6%)	79 (0.8%)
Hipotiroidismo	NO	8348 (89.2%)	667 (90.4%)	9015 (89.2%)
	SI	1016 (10.9%)	71 (9.6%)	1087 (10.8%)
Otra_Enf	NO	8370 (89.4%)	658 (89.2%)	9028 (89.4%)
	SI	994 (10.6%)	80 (10.8%)	1074 (10.6%)
MODALIDAD_TDO	No evaluado	5176 (55.3%)	396 (53.7%)	5572 (55.2%)
	TDO comunitario	61 (0.7%)	6 (0.8%)	67 (0.7%)
	TDO domiciliario	53 (0.6%)	3 (0.4%)	56 (0.6%)
	TDO en IPS	3142 (33.6%)	271 (36.7%)	3413 (33.8%)
	TDO EN IPS	65 (0.7%)	3 (0.4%)	68 (0.7%)
	TDO hospitalario	740 (7.9%)	51 (6.9%)	791 (7.8%)
	TDO virtual	127 (1.4%)	8 (1.1%)	135 (1.3%)
PROGRAMAS_PROTECC_SOCIAL	Cuenta con varios subsidios de apoyo	42 (0.4%)	9 (1.2%)	51 (0.5%)
	No aplica a subsidios	1965 (21.0%)	145 (19.6%)	2110 (20.9%)
	No evaluado	5152 (55.0%)	391 (53.0%)	5543 (54.9%)
	No recibe ninguno	2065 (22.1%)	173 (23.4%)	2238 (22.2%)
	Subsidio alimentario	79 (0.8%)	14 (1.9%)	93 (0.9%)
	Subsidio de desempleo	3 (0.0%)	0 (0%)	3 (0.0%)
	Subsidio de transporte	2 (0.0%)	0 (0%)	2 (0.0%)
	Subsidio de vivienda	26 (0.3%)	4 (0.5%)	30 (0.3%)
	Subsidio educativo	7 (0.1%)	0 (0%)	7 (0.1%)
	Subsidio monetario	23 (0.2%)	2 (0.3%)	25 (0.2%)
REACCIONES_ADVERSAS_TTO	Grave	16 (0.2%)	3 (0.4%)	19 (0.2%)
	Leve	11 (0.1%)	1 (0.1%)	12 (0.1%)
	Moderada	14 (0.1%)	4 (0.5%)	18 (0.2%)
	Ninguna	4179 (44.6%)	341 (46.2%)	4520 (44.7%)
	SD	5144 (54.9%)	389 (52.7%)	5533 (54.8%)
METODOLOGIA_CAPTACION	BAI	4166 (44.5%)	346 (46.9%)	4512 (44.7%)
	Busqueda trabajador salud.	41 (0.4%)	1 (0.1%)	42 (0.4%)
	Durante estudio de	4 (0.0%)	0 (0%)	4 (0.0%)

Variable	Categoría	NO (N=9364)	SI (N=738)	Overall (N=10102)
	contactos			
	Remitido por CNE	1 (0.0%)	0 (0%)	1 (0.0%)
	SD	5152 (55.0%)	391 (53.0%)	5543 (54.9%)
PERTENENCIA_ETNICA	INDIGENA	131 (1.4%)	22 (3.0%)	153 (1.5%)
	NEGRO, MULATO, AFROCOLOMBIANO	70 (0.7%)	12 (1.6%)	82 (0.8%)
	OTRO	9144 (97.7%)	702 (95.1%)	9846 (97.5%)
	PALENQUERO	1 (0.0%)	0 (0%)	1 (0.0%)
	RAIZAL	6 (0.1%)	1 (0.1%)	7 (0.1%)
	ROOM (GITANO)	12 (0.1%)	1 (0.1%)	13 (0.1%)
gp_discapa	NO	9234 (98.6%)	733 (99.3%)	9967 (98.7%)
	SI	130 (1.4%)	5 (0.7%)	135 (1.3%)
gp_desplaz	NO	9301 (99.3%)	722 (97.8%)	10023 (99.2%)
	SI	63 (0.7%)	16 (2.2%)	79 (0.8%)
gp_migrant	NO	8916 (95.2%)	663 (89.8%)	9579 (94.8%)
	SI	448 (4.8%)	75 (10.2%)	523 (5.2%)
gp_carcela	NO	9057 (96.7%)	706 (95.7%)	9763 (96.6%)
	SI	307 (3.3%)	32 (4.3%)	339 (3.4%)
gp_gestan	NO	9348 (99.8%)	734 (99.5%)	10082 (99.8%)
	SI	16 (0.2%)	4 (0.5%)	20 (0.2%)
gp_indigen	NO	9027 (96.4%)	534 (72.4%)	9561 (94.6%)
	SI	337 (3.6%)	204 (27.6%)	541 (5.4%)
gp_pobicbf	NO	9352 (99.9%)	736 (99.7%)	10088 (99.9%)
	SI	12 (0.1%)	2 (0.3%)	14 (0.1%)
gp_psiquia	NO	9357 (99.9%)	734 (99.5%)	10091 (99.9%)
	SI	7 (0.1%)	4 (0.5%)	11 (0.1%)
gp_vic_vio	NO	9352 (99.9%)	734 (99.5%)	10086 (99.8%)
	SI	12 (0.1%)	4 (0.5%)	16 (0.2%)
trabajador_salud	NO	9168 (97.9%)	730 (98.9%)	9898 (98.0%)
	SI	196 (2.1%)	8 (1.1%)	204 (2.0%)
gp_otros	NO	964 (10.3%)	247 (33.5%)	1211 (12.0%)
	SI	8400 (89.7%)	491 (66.5%)	8891 (88.0%)
LOC_RES	CO	1701 (18.2%)	181 (24.5%)	1882 (18.6%)
	FDB	1860 (19.9%)	82 (11.1%)	1942 (19.2%)
	NORTE	2408 (25.7%)	156 (21.1%)	2564 (25.4%)
	Sin Dato	267 (2.9%)	75 (10.2%)	342 (3.4%)
	SO	2029 (21.7%)	133 (18.0%)	2162 (21.4%)
	SUR	1099 (11.7%)	111 (15.0%)	1210 (12.0%)

Variable	Categoría	NO (N=9364)	SI (N=738)	Overall (N=10102)
LOC_DX	CO	3233 (34.5%)	325 (44.0%)	3558 (35.2%)
	NORTE	4444 (47.5%)	239 (32.4%)	4683 (46.4%)
	Sin Dato	17 (0.2%)	2 (0.3%)	19 (0.2%)
	SO	1127 (12.0%)	102 (13.8%)	1229 (12.2%)
	SUR	543 (5.8%)	70 (9.5%)	613 (6.1%)
LOCALIZACION_EXTRA	Cutanea	46 (0.5%)	3 (0.4%)	49 (0.5%)
	Ganglionar	332 (3.5%)	34 (4.6%)	366 (3.6%)
	Genitourinaria	82 (0.9%)	0 (0%)	82 (0.8%)
	Intestinal	80 (0.9%)	2 (0.3%)	82 (0.8%)
	Laringea	1 (0.0%)	0 (0%)	1 (0.0%)
	Meningea	796 (8.5%)	52 (7.0%)	848 (8.4%)
	NA	6373 (68.1%)	567 (76.8%)	6940 (68.7%)
	Osteoarticular	196 (2.1%)	13 (1.8%)	209 (2.1%)
	Otro	215 (2.3%)	8 (1.1%)	223 (2.2%)
	Pericardica	105 (1.1%)	6 (0.8%)	111 (1.1%)
	Peritoneal	143 (1.5%)	5 (0.7%)	148 (1.5%)
	Pleural	977 (10.4%)	46 (6.2%)	1023 (10.1%)
	Renal	17 (0.2%)	2 (0.3%)	19 (0.2%)
	Vejiga	1 (0.0%)	0 (0%)	1 (0.0%)