MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DE LA MORTALIDAD ASOCIADA AL BAJO PESO AL NACER A TÉRMINO, EN MENORES DE UN AÑO EN EL VALLE DEL CAUCA.

Carlos Andrés Torres Ricaurte

Liz Mary Gutiérrez Rendón

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

DELIA ORTEGA LENIS
Directora

DIEGO LUIS LINARES OSPINA
Jurado

JULIÁN GIL GONZÁLEZ
Jurado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ciencia de Datos.

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Cali 8 de febrero del 2024

Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, 8 de febrero del 2024

Autor: Carlos Andrés Torres Ricaurte y Liz Mary Gutiérrez Rendón

Título del Trabajo de Grado: "MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DE LA MORTALIDAD ASOCIADA AL BAJO PESO AL NACER A TÉRMINO, EN MENORES DE UN AÑO EN EL VALLE DEL CAUCA".

Director: Delia Ortega Lenis

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

Firma del Director del Trabajo de Grado

Santiago de Cali, 7 de diciembre del 2023

Doctora
Gloría Inés Alvarez V.
Directora Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana de Cali

Asunto: Presentación para evaluación del proyecto aplicado

Cordial Saludo,

1. Con el fin de cumplir con los requisitos exigidos por la Universidad para optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto denominado "MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DEL RIESGO DE MORTALIDAD ASOCIADA AL BAJO PESO AL NACER A TÉRMINO, EN MENORES DE UN AÑO EN EL VALLE DEL CAUCA.", el cual fue realizado por los estudiantes Carlos Andrés Torres Ricaurte y Liz Mary Gutiérrez Rendón con códigos 0065195 y 8974207 pertenecientes a la Maestría en Ciencia de Datos, bajo la dirección de Delia Ortega Lenis.

El suscrito director del Proyecto Aplicado autoriza para que se proceda a hacer la evaluación de este proyecto, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado y sustentado oficialmente.

Atentamente,

Carlos Andrés Torres Ricaurte
C.C. 94.492.400 de Cali

Delia Ortega Lenis
C.C. 1.130.622.412 de Cali

Liz Mary Gutiérrez Rendón
C.C. 1.151.944.191 de Cali

Documentación anexa:
Resumen del Proyecto Aplicado en formato digital (máximo 1 página).
Una copia digital (PDF) del documento del proyecto aplicado

Valle Invencible

GOBERNACIÓN VALLE DEL CAUCA
Secretaria de Salud

FO-M9-P3-02- V01
1.220.01-18-2023166610

Santiago de Cali, 10 de abril de 2023

Señor
CARLOS ANDRÉS TORRES RICAURTE
Aspirante a grado de maestría en Ciencia de Datos
Universidad Javeriana

Asunto: Respuesta a solicitud información de la base de datos de estadísticas vitales de nacimientos y defunciones RUAF ND 2011-2021, insumo para el proyecto de investigación MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DEL RIESGO DE MORTALIDAD ASOCIADA AL BPNT, EN MENORES DE UN AÑO EN EL VALLE DEL CAUCA.

Cordial saludo,

En respuesta a su solicitud de entrega de la base de datos Nacimientos y defunciones de 2011 a 2021, se le informa que, en reunión del Comité de Investigaciones realizada el día 1 de marzo del 2023, se emitió concepto de viabilidad del proyecto de la referencia.

La entrega de la información solicitada está condicionada al envío del aval ético institucional y el formato de confidencialidad y manejo de datos, los cuales deben ser dirigido al Comité de Investigaciones de la Secretaría de Salud.

Para la Secretaría Departamental de Salud es importante el compromiso del equipo investigador con el envío del reporte de los resultados obtenidos y recibir los créditos institucionales en los productos que generen esta investigación. Reiteramos nuestra disposición a apoyar las publicaciones que se proyecten realizar como resultado de la investigación, para lo cual ponemos a disposición el Grupo de Investigación en Gestión y Estudios de Salud (GIGES).

Atentamente,

MARIA CRISTINA LESMES DUQUE
Secretaria de Salud del Valle del Cauca

Anexo: Resultados de la evaluación.
Revisó: Helmer Zapata, profesional Especializado de la Oficina Asesora de Planeación. Integrante del Comité de Investigaciones.
Proyectó: María Constanza Victoria García
Archivar en: 1.220.01-18

COMITÉ DE ÉTICA EN INVESTIGACIÓN - CEEI
ACTA DE APROBACIÓN DE PROYECTOS
Registro en acta de aprobación **No. 007-2023 del CEEI**

TÍTULO DEL PROYECTO: **"MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DEL RIESGO DE MORTALIDAD ASOCIADA AL BAJO PESO AL NACER A TÉRMINO, EN MENORES DE UN AÑO EN EL VALLE DEL CAUCA"**

Sometido por: Carlos Andrés Torres y Liz Mary Gutiérrez.
Tutor: Dra. Delia Ortega Lenis

El CONSEJO DE LA FACULTAD DE CIENCIAS DE LA SALUD (FCS), ha establecido el Comité de Ética en Investigación en Salud de la FCS (CEEI) de la Pontificia Universidad Javeriana Cali, el cual está regido por la Resolución 008430 del 4 de octubre de 1993 del Ministerio de Salud de Colombia por la cual se establecen las normas científicas, técnicas y administrativas para la investigación en salud; los principios de la Asamblea Médica Mundial expuestos en su Declaración de Helsinki de 1964, última revisión 2013 y las pautas éticas internacionales para la investigación biomédica en seres humanas enunciadas en 1982 por el Consejo de Organizaciones Internacionales de las Ciencias Médicas (CIOMS).

Este Comité certifica que:
1. Los miembros del comité han revisado y evaluado los siguientes documentos del proyecto
a) Protocolo de investigación
b) Carta de aval metodológico por parte del programa de Maestría en Salud Pública
c) Declaración de conflicto de intereses
d) Formato para la recolección de datos
e) Consentimiento informado

2. Los documentos fueron evaluados y aprobados por el Comité.
3. El Presidente del Comité informa que este estudio pertenece a la categoría sin riesgo
4. El Comité informa que el presente proyecto ha sido aprobado por un periodo de **6 meses** a partir de la fecha. En caso de extenderse, los investigadores deberán enviar carta donde se sustente la razón por la cual no se cumplieron los tiempos y la modificación del cronograma. Si es necesario el CEEI podría determinar que se debe de someter de nuevo el proyecto a evaluación. No obstante, el CEEI puede ser convocado a solicitud de alguno de sus miembros o de las directivas institucionales para revisar cualquier asunto relacionado con la realización del estudio.

6. El Comité informará inmediatamente a las directivas institucionales:
a. Todo desacato de los investigadores a las solicitudes del Comité.
b. Cualquier suspensión o terminación de la aprobación del protocolo por parte del Comité.
7. El investigador principal deberá informar al Comité
a. Cualquier cambio o enmienda que se proponga introducir en este proyecto. Estos cambios no podrán iniciarse sin la revisión y aprobación del Comité
b. Sucesos inesperados relacionados con la conducción del estudio, informando la subsiguiente respuesta por parte de los investigadores.
d. Informar cualquier decisión tomada por otros comités de ética.
e. La terminación prematura o suspensión del proyecto explicando la razón para esto.
f. Nueva información que pudiera afectar la proporción riesgo-beneficio del estudio.
g. El investigador principal deberá presentar un informe a los 6 meses de aprobación, para verificar el estado actual del protocolo y determinar si se requiere ampliación del aval.
h. Si el proyecto se extiende **se debe renovar el aval ético**.

Para constancia se firma el presente certificado,

Firma 
Nombre: Dr. EDUARDO CASTRILLON
Presidente
Comité de Ética en Investigación en Salud de la FCS (CEEI)

Firma 
Nombre: GLORIA S. LIZARRALDE G. PhD.
Secretaría Técnica
Comité de Ética en Investigación en Salud de la FCS (CEEI)

Fecha: CALI, SEPTIEMBRE DE 2023

Realizo aval	GLORIA S. LIZARRALDE G.
Reviso y avalo	EDUARDO CASTRILLÓN
Archivo y envío a investigadores	MAIRA ARIZA

FORMATO DE CONFIDENCIALIDAD Y MANEJO DE DATOS

TÍTULO DEL PROYECTO: MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DEL RIESGO DE MORTALIDAD ASOCIADA AL BAJO PESO AL NACER A TÉRMINO, EN MENORES DE UN AÑO EN EL VALLE DEL CAUCA.
NOMBRE DEL INVESTIGADOR PRINCIPAL: CARLOS ANDRES TORRES RICAURTE NÚMERO DE IDENTIFICACIÓN: 94492400
CLÁUSULAS PRIMERA. EL INVESTIGADOR Y SU GRUPO se obliga a no divulgar a terceras partes, la "Información confidencial", que reciba por parte de la SECRETARÍA DE SALUD DEPARTAMENTAL DEL VALLE, y a darle a dicha información el mismo tratamiento que le darían a la información confidencial de su propiedad. Para efectos de la presente acta, comprende toda la información entregada en forma oral, visual, escrita o en cualquier otra forma tangible y que se encuentre claramente marcada como tal al ser entregada a la parte receptora. SEGUNDA. El INVESTIGADOR Y SU GRUPO se obliga a utilizar la información entregada por la SECRETARÍA DE SALUD DEPARTAMENTAL DEL VALLE solo dentro de la investigación que fue autorizada por el comité de investigación y el comité de ética de dicha institución. TERCERA. Es obligación del INVESTIGADOR Y SU GRUPO destruir la información entregada posterior a su uso para el proyecto autorizado. CUARTA el INVESTIGADOR Y SU GRUPO se comprometen a entregarlos resultados de la investigación a la SECRETARIA DEPARTAMENTAL DE SALUD DEL VALLE. QUINTA. INVESTIGADOR Y SU GRUPO se compromete a efectuar una adecuada custodia y reserva de la información y gestión -es decir tratamiento- de los datos suministrados por SECRETARIA DEPARTAMENTAL DE SALUD DEL VALLE al interior de las redes y bases de datos (físicas y/o electrónicas) en donde se realice su recepción y tratamiento en general. SEXTA. Para el caso del manejo de información que incluya datos personales, el INVESTIGADOR Y SU GRUPO darán estricto cumplimiento a las disposiciones constitucionales y legales sobre la protección del derecho fundamental de habeas data, en particular a lo dispuesto en el artículo 15 de la Constitución Política y la ley 1581 de 2012.
Suscrita a los 13 días del mes de septiembre de 2023, en Santiago de Cali
Firma:

Carlos Andrés Torres Ricaurte
Estudiante de la Maestría en Ciencia de Datos
Facultad de ingeniería y Ciencias

Pontificia Universidad
JAVERIANA
Cali

MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DE LA MORTALIDAD ASOCIADA AL BAJO PESO AL NACER A TÉRMINO, EN MENORES DE UN AÑO EN EL VALLE DEL CAUCA.

Carlos Andrés Torres Ricaurte
Liz Mary Gutiérrez Rendón

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
SANTIAGO DE CALI,
DICIEMBRE 9 DE 2023

FICHA RESUMEN PROYECTO DE TRABAJO DE GRADO

TÍTULO: Modelo de aprendizaje automático para la predicción de la mortalidad asociada al bajo peso al nacer a término, en menores de un año en el Valle del Cauca.

1. ÁREA DE TRABAJO: Clasificación del riesgo de mortalidad infantil.
2. TIPO DE PROYECTO (Aplicado, Innovación, Investigación): Aplicado.
3. ESTUDIANTE(S): Carlos Andrés Torres Ricaurte, Liz Mary Gutiérrez Rendón.
4. CORREO ELECTRÓNICO: carlos.torres@javerianacali.edu.co,
5. DIRECCIÓN Y TELÉFONO: Calle 1B oeste # 4A oeste – 201, 3103834365.
6. DIRECTORA: Delia Ortega Lenis
7. VINCULACIÓN DEL DIRECTOR: Docente de la Pontificia Universidad Javeriana
8. CORREO ELECTRÓNICO DEL DIRECTOR: delia.ortega@javerianacali.edu.co
9. GRUPO O EMPRESA QUE LO AVALA: N/A
10. OTROS GRUPOS O EMPRESAS: N/A
11. PALABRAS CLAVE (al menos 5): Bajo Peso al Nacer, Mortalidad Infantil, Determinantes Sociales de la Salud, aprendizaje automático, clasificación.
12. FECHA DE INICIO: Enero 2023
13. DURACIÓN ESTIMADA (En meses): 12 meses
14. RESUMEN:

La detección del peso al nacer es un indicador importante del estado de salud del recién nacido; según la Organización Mundial de la Salud (OMS), el bajo peso al nacer, que se ha establecido como menor a 2.500 gramos, es un problema de salud pública a nivel mundial que debe atenderse para evitar consecuencias fatales como la muerte. Colombia no es ajena a esta problemática, la prevalencia del bajo peso al nacer en 2018 fue de 7.22 y en el 2020 aumentó a 9.20. Es por esto, que este proyecto desarrolló un modelo de aprendizaje automático para la predicción del riesgo de mortalidad asociada al bajo peso al nacer a término, en menores de un año en el Valle del Cauca, que, mediante el uso de las técnicas de predicción y clasificación, permitió analizar las dinámicas del comportamiento del bajo peso al nacer a término y el riesgo de la mortalidad infantil de esta manera generar acciones de carácter preventivo que procuren su reducción en el territorio.

TABLA DE CONTENIDO

LISTADO DE TABLAS

LISTADO DE GRÁFICOS

LISTADO DE ILUSTRACIONES

INTRODUCCIÓN

La Organización Mundial de la Salud (OMS) señala que cuando un recién nacido presenta un peso inferior a 2500 gramos, se considera que tiene “bajo peso al nacer” (BPN), esta es una condición latente a nivel mundial que puede tener consecuencias graves en el corto y largo plazo, en la salud del neonato; la padecen entre el 15% y el 20% de los recién nacidos anualmente en todo el mundo, lo que equivale a más de 20 millones de ellos. El nacimiento prematuro es una de las principales causas de muerte de los neonatos por situaciones asociadas a este hecho; el BPN también aumenta el riesgo de padecer enfermedades no transmisibles en la vida posterior. [1]

El Grupo Interinstitucional de las Naciones Unidas para la Estimación de la Mortalidad Infantil (UN-IGME), precisa que en el 2021, murieron 5 millones de niños y niñas, antes de cumplir cinco años, a causa de la privación a su derecho básico de una atención sanitaria, a vacunas, a una alimentación adecuada, a agua potable y saneamiento; condiciones que a la postre desencadenan enfermedades transmisibles e infecciones, pudiendo ser estas prevenibles; lo anterior sin conocerse aún, los efectos del periodo de pandemia en el cual muchos programas de salud sufrieron intermitencias. Durante ese mismo año, se estima que 2,3 millones de niños y niñas fallecieron durante sus primeros 28 días de vida, siendo este el periodo más vulnerable para su supervivencia; esta cifra representa el 47% de las defunciones de niños y niñas menores de cinco años en el mundo. Según esta entidad internacional, la tasa de mortalidad neonatal (0-28 días) es de 18 y en lactantes menores (1 - 12 meses) es de 11. [2]. Lo anterior, demuestra lo alarmante de este fenómeno de la mortalidad infantil MI en el mundo.

De otra parte, el Departamento Administrativo Nacional de Estadística DANE, en Colombia, muestra que el BPN ha venido en aumento desde del año 2020 cuando se obtuvo una prevalencia de 9,1%, en el 2021 fue del 9,5%; según los estudios preliminares en el 2022 la tasa fue de 10,5% y en lo que va corrido del 2023, la tasa se encuentra en el 11%. [3]. Con respecto a la MI, se conoció, de forma preliminar que para el año 2022 fallecieron en el país, 3.776 niños menores de un año y en el 2023, fallecieron 3.358. [4]

A nivel departamental, el DANE muestra que para el 2019, la tasa de mortalidad infantil en el Valle del Cauca se situaba en el 9,45%, mientras que el BPN era del 9%; sin embargo, las estimaciones de la tasa de mortalidad en función de sus determinantes, expone que el mayor determinante biológico es el BPN. [5]. Por ello, en esta investigación se desarrolló un “Modelo de aprendizaje automático para la predicción de la mortalidad asociada al bajo peso al nacer a término (BPNT), en menores de un año en el Valle del Cauca” mediante el cual fue posible generar predicciones que permitieron establecer el riesgo de muerte al momento del nacimiento, y de esta manera evitar la mortalidad infantil relacionada con el evento epidemiológico del BPN.

Los resultados del proyecto basados en los objetivos propuestos, fueron los siguientes 1) Construcción de una base de datos unificada a partir de los datos de nacimientos con bajo peso al nacer a término (BPNT) y defunciones de menores de un año, que permitió caracterizar los factores sociodemográficos de la madre en el Valle del Cauca en el periodo de tiempo 2011-2020. 2) Se probaron varios modelos de aprendizaje automático y se entrenaron, para la clasificación del BPN y la MI de acuerdo con las características sociodemográficas de la madre en el Valle del Cauca en el período de tiempo 2011-2020. 3) Se generaron los resultados a partir de la evaluación de la solución del modelo de aprendizaje automático, utilizando las métricas exhaustividad (Recall), exactitud (Accuracy), precisión y F1-score, para la predicción del riesgo de mortalidad infantil asociada al bajo peso al nacer a término, en menores de un año en el Valle del Cauca.

1. DEFINICIÓN DEL PROBLEMA

1.1. Planteamiento del problema

La Organización Mundial de la Salud (OMS) indica que un peso al nacer inferior a 2.500 gramos se considera bajo. El BPN es un problema de salud pública a nivel mundial que se asocia a consecuencias de corto y largo plazo, siendo el factor más importante para presentar un crecimiento y un desarrollo óptimos. Los números significativos de recién nacidos con BPN y las cifras elevadas de Mortalidad Infantil (MI) se plantean como un rasgo problemático en el proceso de desarrollo de un país. Está comprobado a través de varias investigaciones que la MI y el BPN son, indudablemente, dos indicadores que están relacionados directamente: “los países con las tasas de MI más elevadas tienden también a presentar las tasas más altas de BPN” [6].

De acuerdo con la publicación realizada en el año 2021 por la UNICEF sobre el “Estado Mundial de la Infancia”, el número de niños y niñas con BPN que se registra en un año a nivel mundial es de aproximadamente 20 millones [7]. Estos recién nacidos tienen un mayor riesgo de morir durante los primeros meses y años de vida, los que sobreviven están propensos a sufrir alteraciones del sistema inmunológico y en el futuro pueden presentar alguna de las enfermedades crónicas. En el mundo, para el año 2019 según la OMS, las muertes de recién nacidos o neonatos, constituyen el 47% de las muertes de niños menores de cinco años. La mayoría de las muertes de neonatos (75%) se produce durante la primera semana de vida, y de éstos, entre el 25% y el 45% se producen en las primeras 24 horas. Las causas principales de muerte de los recién nacidos son: el nacimiento prematuro, el BPN, las infecciones, la asfixia (falta de oxígeno al nacer) y los traumatismos en el parto. Estas causas explican casi el 80% de las muertes en este grupo de edad [8].

En Colombia, según el Instituto Nacional de Salud [9], para el 2021, la tasa de mortalidad infantil era de 10,93 por cada 1000 nacidos vivos, mientras que en el Valle del Cauca fue de 9,3; entre tanto y para el mismo año a nivel nacional, la prevalencia del BPN fue de 9,73 y a nivel departamental se encontró en 9,59 [10].

Un hecho aún más preocupante, es el del bajo peso al nacer a término BPNT que se define como el recién nacido de 37 o más semanas de gestación cuyo peso al nacer registrado sea ≤2.499 gr. [11]. En Colombia, para el 2021 la proporción del BPNT fue de 3,3% y para el 2022 aumentó a 3,4% [12]. En Santiago de Cali a partir del Boletín Epidemiológico BPNT de la Secretaria de Salud Pública para el periodo epidemiológico I (enero 01 al 28) del 2023, se presenta el comportamiento de este fenómeno para ese

mismo periodo entre el 2018 y 2023, encontrando que para el año 2021, la proporción del BPNT fue 3,20, manteniéndose igual para el 2022 [13].

Además, en Cali se observa que, del total de defunciones en menores de un año, 8 de cada 10 defunciones en nacidos vivos con BPNT ocurren antes de cumplir el primer mes de vida. De hecho, hay evidencia que el evento de interés epidemiológico BPN incrementa la morbilidad y mortalidad infantil, así como el tiempo y costo de hospitalización [14].

Pese a numerosas investigaciones sobre BPN y sus factores determinantes en el periodo gestacional, donde se consideran aspectos tanto biológicos como sociales [15], el BPN tiene una prevalencia del 9% a nivel distrital, siendo partícipe del 65% en las defunciones respecto al total de NV fallecidos y en el 20% de NV a término fallecidos [16].

Los factores de riesgo de BPN, de acuerdo a diferentes estudios son los siguientes: edad materna menor de 20 años y mayor de 35 años, peso del recién nacido menor a 2.500 gramos, estatura materna menor de 150 centímetros, antecedentes personales patológicos, edad a la menarca menor o igual a 12 años, primiparidad (un parto) y multiparidad (más de 4 partos), antecedentes de abortos, antecedentes obstétricos patológicos, nivel socioeconómico bajo, estado civil “no casada”, tabaquismo, alcoholismo, inicio de la atención prenatal a partir o después de la semana 20 de gestación, número de consultas prenatales menor de seis, género femenino [17]. Hasta ahora, la literatura científica muestra estudios relacionados con la identificación de factores de riesgo y algunos modelos de inteligencia artificial para el apoyo de la vigilancia epidemiológica, pero en Colombia no se han desarrollado este tipo de soluciones [7], [18].

Si bien el BPN es vigilado por el Instituto Nacional de Salud a través del Sistema Nacional de Vigilancia en Salud Pública - SIVIGILA, y que desde el Ministerio de Salud y la Protección Social se tienen protocolos para la atención clínica del neonato con esta condición, es evidente que las cifras BPN han venido en aumento, de modo que es necesario continuar desarrollando procesos de investigación, de la mano de la ciencia de datos para contribuir de manera preventiva con el mejoramiento de la salud de los neonatos.

En este sentido, este proyecto desarrolló un modelo de aprendizaje automático para la predicción del riesgo de mortalidad asociada al BPNT, en menores de un año en el Valle del Cauca; el cual permite generar alertas tempranas y precisas sobre el riesgo de mortalidad infantil en NV con BPNT, descartando a aquellos nacidos a

pretérmino/prematuros puesto que, debido a su condición, corren mayor riesgo de fallecer debido a complicaciones de salud más específicas.

1.2. Formulación del problema

A partir del planteamiento del problema en la sección anterior, se logran formular las siguientes preguntas: ¿Qué modelo de aprendizaje automático se puede desarrollar para predecir la mortalidad asociada al bajo peso al nacer a término, en menores de un año en el Valle del Cauca?, ¿Cuáles son las variables que se consideran como factores de riesgo que se deben priorizar en la recolección de los datos?, ¿Cuáles podrían las herramientas de aprendizaje automático más apropiadas para lograr el desarrollo del modelo predictivo? ¿Cuáles serán las métricas más adecuadas para evaluar la eficiencia del modelo y que permita realizar predicciones con la mayor precisión?.

2. OBJETIVOS DEL PROYECTO

2.1. Objetivo general

Desarrollar un modelo de aprendizaje automático para la predicción de la mortalidad infantil asociada al bajo peso al nacer a término, en menores de un año en el Valle del Cauca.

2.2. Objetivos específicos

a) Elaborar una base de datos unificada a partir de los datos de nacimientos con BPNT y defunciones de menores de un año, que permita caracterizar los factores sociodemográficos de la madre en el Valle del Cauca en el periodo de tiempo 2011-2021.

b) Entrenar el modelo de aprendizaje automático para la clasificación del BPNT y la MI de acuerdo con las características sociodemográficas de la madre en el Valle del Cauca en el período de tiempo 2011-2021.

c) Evaluar la solución del modelo de aprendizaje automático, utilizando las métricas exhaustividad (Recall), exactitud (Accuracy), precisión y F1-score, para la predicción de la mortalidad infantil asociada al BPNT, en menores de un año en el Valle del Cauca.

3. MARCO DE REFERENCIA

3.1. Marco teórico

3.1.1. Situación del bajo peso al nacer BPN y la mortalidad infantil MI, en el contexto mundial y regional

El BPN es una de las medidas que indican en qué condiciones transcurrió el embarazo y puede hablar sobre cómo será la salud del recién nacido a medida que va creciendo, e incluso podría dar pistas sobre cuánto tiempo vivirá. Un estado de salud adecuado y un desarrollo mental sano, pueden verse impactados para aquellos que nacen con un peso menor a 2.500 gramos. En este sentido, el BPN es considerado como uno de los principales indicadores de la salud y el bienestar infantil, de ahí que entender los factores que se asocian a este problema es un aspecto clave para establecer mecanismos de control que permitan prevenir este hecho, previniendo de paso la muerte durante los primeros años de vida o retrasos en el crecimiento de las niñas y los niños. El BPNT se cataloga como el recién nacido de 37 o más semanas de gestación con un peso al nacer inferior o igual a los 2.499 gr. [11].

De acuerdo con Instituto Nacional de Salud [19], cada año nacen en el mundo alrededor de 20 millones de niñas y niños con un peso inferior a los 2.500 gramos, lo que representa entre el 15% y el 17% de los nacimientos; añadiendo a esto que “cerca al 96% de los nacimientos con bajo peso ocurren en los países en desarrollo”. Para el caso de América Latina y el Caribe, el indicador se mantiene entre 9,3% y 9,6% para el periodo comprendido de los años 2000 al 2020, y para Colombia está también ha sido la cifra en los últimos años, mostrando un preocupante aumento [20]. Para el año 2020, nuestro país tenía una prevalencia de 9,0, para el 2021 fue del 9,7, de manera preliminar para el 2022 fue de 10,4% y en lo que va corrido del 2023, asciende a 10,9% lo que expone claramente dicho aumento [10].

Con respecto al BPNT en Colombia, para el 2021 la proporción de este fenómeno fue de 3,3% y se incrementó a 3,4% para el 2022 [12]. En Santiago de Cali acorde con [13] para el periodo epidemiológico I (enero 01 al 28) del 2023, se presenta el comportamiento de este fenómeno para ese mismo periodo entre el 2018 y 2023, encontrando que para el año 2021, la proporción del BPNT fue 3,20 y se mantuvo igual para el 2022. Esto señala la necesidad de realizar un análisis detallado de factores asociados a este fenómeno, ya que se observa un incremento progresivo en las cifras para este mismo periodo epidemiológico.

Por otro lado, la Organización de las Naciones Unidas a través de la formulación de Objetivos de Desarrollo, exhorta a los países del mundo a tomar acción para el

mejoramiento de las condiciones de vida de la población. Los ODM (Objetivos de Desarrollo del Milenio), incluyeron para el periodo 1990-2015, reducir en dos terceras partes la mortalidad en niños menores de 5 años. Para el periodo 2015-2030, los ODS (Objetivos de Desarrollo Sostenible) proponen reducir la mortalidad neonatal a 12 nacidos vivos por cada 1.000 y la mortalidad en menores de 5 años, a 25 por cada 1.000 [21].

Como lo explica Torres [22], la Encuesta Nacional de Demografía y Salud – ENDS, diseñada con el propósito de medir cambios en las variables demográficas, evaluar y hacer los ajustes necesarios en los programas de salud y obtener datos e información actualizados en población, salud, salud sexual y salud reproductiva y nutrición, planteaba que, si bien en 2015 se mostraba un descenso progresivo de la MI en el país, pasando de 27 en 1990-1995 a 14 muertes por 1.000 nacidos vivos para el 2010-2015, se presentaban algunos aspectos relevantes a tener en cuenta, como por ejemplo, diferencias en tasas de mortalidad infantil según zona rural o urbana, la enorme influencia que tiene la edad de la madre, el tiempo que transcurre entre un embarazo y otro y los niveles socioeconómicos y educativos de la madre, por mencionar algunos [23].

Ahora bien, múltiples investigaciones [6], [24], [25] han mostrado la asociación directa del problema del BPN con la mayor frecuencia durante los primeros años de vida de diversos trastornos, entre los que se destacan el retraso del desarrollo neurológico, la hemorragia cerebral, las alteraciones respiratorias y otras enfermedades que pueden estar provocando la muerte a temprana edad del recién nacido; esto se relaciona directamente con los problemas de mortalidad infantil, ya que hay datos que prueban la relación del BPN con una mayor probabilidad de muerte. Es debido a estas razones que el BPN es considerado como un indicador general del estado de la salud de una población y la frecuencia de ocurrencia de este hecho evidencia el desarrollo de la salud reproductiva de la misma.

Los nacidos con un peso entre 2.000 y 2.499 gramos, tienen cuatro veces mayor riesgo de muerte neonatal que aquellos que nacen pesando más de 2.500 gramos y este riesgo se incrementa entre 10 y 14 veces frente a quienes han nacido con un peso entre los 3.000 y los 3.499 gramos; y a esto se suman las cifras preocupantes del 72,2% de niñas y niños recién nacidos que mueren por una causa asociada al BPN, según la Organización Panamericana de la Salud [19].

Aunado a esto, el Departamento Administrativo Nacional de Estadística – DANE, define de manera clara [18] cuatro tipos de determinantes que intervienen en la mortalidad infantil: biológicos, socioeconómicos, servicios públicos y ambientales. Entre los primeros, incluye sexo, edad gestacional, las características reproductivas de la madre y la edad, entre otras; entre los factores que juegan un rol fundamental en los determinantes de tipo biológico, está el peso menor a 2.500 gramos al nacer, considerado

como un factor de riesgo. El DANE en esta investigación reconoce que, si bien inicialmente no es clara la relación entre este factor y la MI en los departamentos colombianos para el 2019, sí menciona explícitamente que el porcentaje de nacidos vivos con bajo peso al nacer, juega un papel determinante que no puede pasarse por alto.

El DANE muestra que para el 2021, la tasa de MI en el Valle del Cauca rondaba el 9,3%, mientras que el BPN era del 9%; sin embargo, las estimaciones de la tasa de mortalidad en función de sus determinantes, expone que el mayor determinante biológico es el BPN, aun cuando no arroja conclusiones definitivas o contundentes sobre la relación entre ambas variables, ya que advierte que al revisar por determinantes relacionados con el servicio de salud o con la seguridad alimentaria del hogar, esta relación parece perder fuerza en su efecto [5].

3.1.1.1. Situación de la mortalidad infantil y bajo peso al nacer en el contexto regional y departamental del Valle del Cauca 2011-2021

En el caso de Colombia, tanto la MI como la tasa de mortalidad en niños menores de 5 años han experimentado una reducción constante, disminuyendo de 12,25 en 2011 a 10,94 muertes por cada 1,000 nacidos vivos en 2021 (Tabla 1). Un análisis detallado de la MI en diversos departamentos del país para el período 2011-2021 revela disparidades significativas. Departamentos como Chocó, La Guajira, Atlántico y Magdalena presentan tasas que superan el promedio nacional en este lapso de tiempo. Aunque las tasas de MI han descendido en diferentes departamentos y regiones, persisten desigualdades notables en los contextos territoriales. Las marcadas diferencias entre regiones y departamentos subrayan la necesidad de abordar de manera más efectiva las disparidades persistentes.

Tabla 1. Tasa de mortalidad infantil para 10 departamentos 2011 - 2021

Departamento	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Total Nacional	12,25	12,14	11,56	11,34	10,96	11,15	10,73	11,29	11,35	10,12	10,94
Chocó	24,56	26,63	22,81	25,01	28,35	29,21	25,55	24,30	27,36	16,16	20,44
La Guajira	19,80	18,64	18,38	17,97	15,72	18,59	16,18	19,51	23,51	17,06	18,53
Nariño	14,99	13,50	12,05	12,14	11,92	9,83	9,61	10,70	11,26	10,33	9,86
Atlántico	14,63	12,14	14,16	13,90	13,98	12,99	12,55	14,35	15,12	14,67	15,02
Magdalena	13,95	13,99	15,14	13,84	13,08	11,88	12,83	12,84	12,42	10,73	13,02
Meta	12,23	12,28	11,71	12,55	9,84	11,01	8,56	9,44	11,40	9,51	9,07
Norte de Santander	12,17	10,98	10,05	9,84	10,41	11,04	11,48	11,02	12,02	9,70	11,78
Bogotá D.C	12,06	11,83	10,19	10,10	8,86	9,41	9,41	9,24	9,58	8,21	8,54
Antioquia	10,74	9,90	9,93	9,59	8,78	9,00	9,11	8,48	8,75	7,73	9,44
Valle del Cauca	10,68	10,73	9,65	9,45	10,24	9,85	10,05	10,04	9,45	9,58	9,30

Fuente: DANE, Estadística Vitales 2011 – 2021

Para el caso del del departamento del Valle del Cauca y los 10 primeros municipios con las tasa de mortalidad más altas (Bolívar, Versalles, El Dovio, Argelia, Buenaventura, Toro, Trujillo, La Victoria, Ansermanuevo, El Cairo), se observa un patrón similar al descrito para los departamentos en general. La disminución progresiva de la tasa de MI a lo largo del tiempo es evidente. No obstante, surgen diferencias entre algunos municipios del departamento. Por ejemplo, se observa que Cali registró una tasa de MI de 9,76 en 2011, Bolívar presentó una tasa considerablemente más alta, alcanzando 41,10 muertes por cada 1,000 nacidos vivos en el mismo año (Tabla 2). Estas diferencias intermunicipales sugieren la necesidad de examinar factores sociodemográficos para comprender estas brechas, que están estrechamente vinculadas a las condiciones ambientales, sociales y económicas particulares de cada territorio.

Tabla 2. Tasa de mortalidad infantil por municipios del Valle del Cauca 2011 – 2021.

Municipios	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Bolívar	41,10	16,67	64,52	8,93	24,59	32,52	26,32	21,58	26,49	7,41	0,00
Versalles	24,69	11,36	12,82	0,00	13,16	50,85	54,55	0,00	32,79	15,15	30,30
El Dovio	23,26	48,00	13,89	20,83	37,74	15,75	18,29	25,00	15,50	12,99	12,82
Argelia	25,97	0,00	26,67	16,67	0,00	33,33	39,22	19,23	14,08	0,00	46,88
Buenaventura	12,03	14,78	17,00	13,88	20,20	8,55	15,58	20,82	21,88	19,47	12,63
Toro	10,93	31,06	12,58	19,61	14,39	14,71	15,38	6,90	0,00	0,00	43,48
Trujillo	9,71	19,70	35,00	5,21	4,61	26,74	10,31	21,55	9,71	5,59	14,63
La Victoria	25,64	25,86	22,22	15,63	17,09	9,52	0,00	9,43	0,00	24,10	0,00
Ansermanuevo	11,45	24,39	12,71	13,70	9,39	18,60	23,70	0,00	5,71	10,36	17,05
El Cairo	17,86	13,89	0,00	24,69	0,00	0,00	35,09	15,87	0,00	19,61	14,08
La Unión	11,20	18,18	12,82	14,53	17,00	11,53	14,41	2,94	8,47	6,67	12,08
Zarzal	9,62	8,51	6,44	7,38	17,32	15,84	19,27	16,95	2,26	12,66	13,48
Yotoco	21,28	6,94	6,17	13,61	13,25	13,61	13,99	17,14	7,30	13,42	0,00
Tuluá	12,19	13,01	15,84	9,31	11,30	12,43	12,05	6,20	9,46	12,67	9,16
La Cumbre	8,70	18,35	8,47	10,00	8,70	7,30	9,26	11,49	9,71	27,52	0,00
Calima	0,00	6,13	6,54	13,16	13,99	11,17	12,42	5,52	26,74	10,64	11,36
San Pedro	12,05	5,78	11,63	11,76	5,95	10,26	6,29	13,16	13,89	19,23	6,13
Sevilla	22,94	15,02	9,17	14,53	7,08	7,50	7,33	4,90	10,67	5,42	10,72
Dagua	6,51	6,54	14,89	9,09	11,71	26,43	4,37	12,25	5,32	11,01	5,88
El Cerrito	9,80	10,93	10,36	16,33	7,46	12,89	12,54	5,48	12,20	8,98	6,04
Florida	21,99	17,81	1,47	8,44	17,27	7,49	10,07	4,73	6,42	10,01	7,18
Valle Del Cauca	**10,68**	**10,70**	**9,65**	**9,45**	**10,24**	**9,83**	**10,05**	**10,04**	**9,45**	**9,58**	**9,30**
Riofrío	12,42	0,00	5,65	5,78	21,05	29,94	6,45	5,62	14,08	0,00	7,19
Roldanillo	17,81	8,00	5,75	8,02	5,95	11,33	11,17	8,50	15,67	6,02	9,20
Pradera	4,76	14,57	8,11	14,08	14,06	9,58	10,89	9,65	9,71	0,00	9,42
Bugalagrande	8,89	12,35	4,42	21,39	0,00	8,89	4,27	0,00	4,81	5,26	29,41
Cali	9,76	10,11	8,95	8,54	8,86	9,09	9,52	9,08	8,26	7,92	9,13

Municipios	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Restrepo	10,47	16,13	15,31	4,81	5,92	10,15	5,81	5,62	0,00	22,73	0,00
Alcalá	9,71	17,24	10,15	0,00	0,00	5,68	17,14	12,74	0,00	5,62	17,86
Obando	6,71	13,99	29,41	7,87	29,41	0,00	0,00	8,70	0,00	0,00	0,00
Cartago	13,11	5,52	7,23	10,55	6,90	9,05	3,14	7,78	10,36	12,93	9,30
Jamundí	13,01	7,59	5,26	12,66	6,56	9,89	7,94	10,50	5,76	8,06	8,09
Guadalajara De Buga	7,94	10,45	7,36	7,02	8,26	9,77	6,61	7,89	9,29	7,61	11,97
El Águila	0,00	9,80	0,00	9,62	23,53	28,99	0,00	21,74	0,00	0,00	0,00
Caicedonia	5,67	3,12	3,50	19,48	7,58	10,95	3,25	18,80	8,81	7,46	3,95
Andalucía	4,46	4,48	14,71	8,93	4,37	19,42	0,00	0,00	14,93	11,98	6,45
Candelaria	14,34	8,73	8,46	8,70	6,05	8,36	7,82	5,12	3,04	10,54	8,37
Yumbo	7,57	10,00	8,61	1,79	4,48	8,02	10,21	11,83	4,60	12,78	7,92
Palmira	9,68	7,38	3,98	7,71	9,86	9,26	8,83	8,90	6,92	6,65	6,31
Vijes	10,31	9,52	7,69	0,00	0,00	20,00	8,77	0,00	9,35	8,55	9,62
Guacarí	12,12	0,00	8,96	12,25	10,61	2,80	5,52	8,31	3,19	10,84	0,00
Ulloa	32,79	0,00	0,00	0,00	0,00	22,73	0,00	0,00	0,00	0,00	0,00
Ginebra	10,05	0,00	0,00	0,00	9,85	4,95	0,00	0,00	5,92	5,85	0,00

Fuente: DANE, Estadística Vitales 2011 - 2021

La dinámica de la MI, tanto a nivel nacional como en diversos departamentos y municipios del país, subraya la necesidad de identificar y analizar minuciosamente los patrones, factores sociodemográficos, así como las causas y determinantes sociales de la salud que más inciden en este fenómeno. Este análisis debe tener en cuenta los contextos territoriales internos a nivel departamental. Al examinar detenidamente el indicador, se corrobora la semejanza de los resultados con diversos estudios vinculados a este fenómeno. Las relaciones establecidas entre variables como el sexo, la zona geográfica y los niveles socioeconómicos y educativos de la madre indican que en condiciones socioeconómicas y demográficas menos favorables, las tasas de MI tienden a ser más elevadas [24]. Este enfoque proporciona una comprensión más integral de las disparidades en la MI y destaca la importancia de abordar no solo las consecuencias inmediatas, sino también las causas subyacentes de estas desigualdades.

Situación del bajo peso al nacer

Durante el periodo considerado en esta investigación y en línea con los Objetivos de Desarrollo Sostenible para América Latina y el Caribe en 2018, aproximadamente el 10,0% de los nacidos vivos en Colombia fueron registrados con bajo peso al nacer. Al presentar estos datos por zonas geográficas, se observa que la proporción de nacimientos con bajo peso en México fue del 7.9% para ese mismo período. En los países como Honduras y Guatemala, se evidenció una prevalencia de bajo peso al

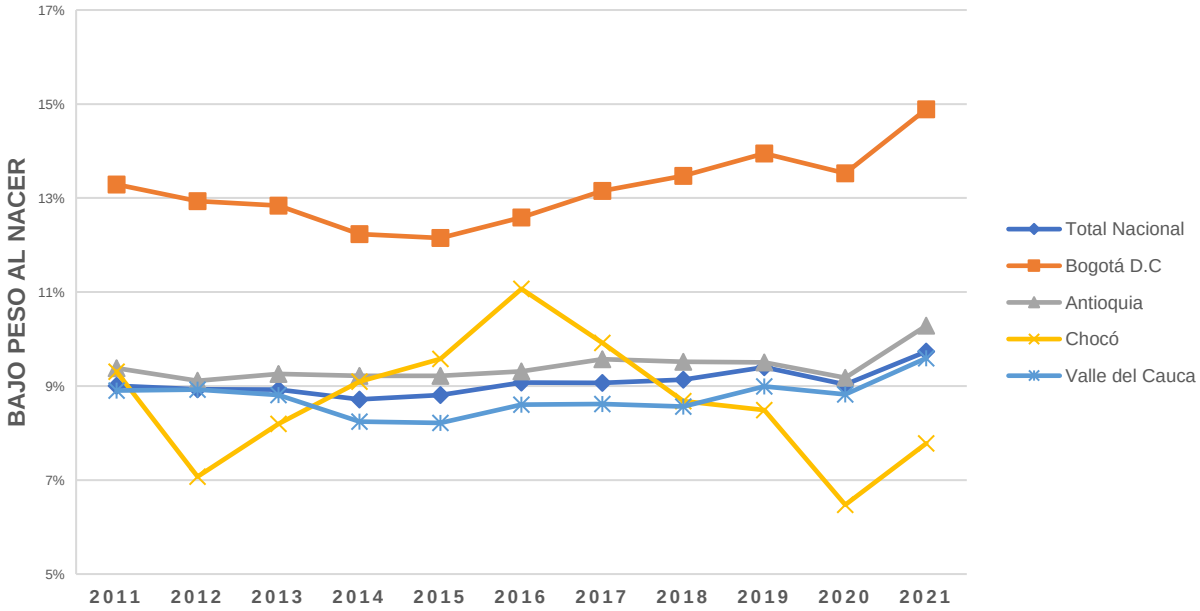
nacer del 10,9% y 11,0%, respectivamente. En los países del Caribe, como Jamaica, esta proporción alcanzó el 14,6%, mientras que en la región andina de América, se registró una proporción del 9,8% de nacimientos con bajo peso. Por otro lado, Brasil presentó un indicador por debajo del promedio regional, con un 8.4% de nacimientos con bajo peso durante este periodo [26].

Tabla 3. Bajo peso al nacer por Departamentos 2011 - 2021

Departamento	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Total Nacional	9,01%	8,93%	8,92%	8,72%	8,81%	9,07%	9,07%	9,14%	9,40%	9,04%	9,73%
Bogotá D.C	13,29%	12,94%	12,84%	12,23%	12,15%	12,59%	13,15%	13,48%	13,95%	13,53%	14,89%
Antioquia	9,38%	9,11%	9,26%	9,22%	9,22%	9,32%	9,57%	9,52%	9,51%	9,18%	10,28%
Chocó	9,31%	7,08%	8,20%	9,10%	9,58%	11,07%	9,92%	8,68%	8,49%	6,47%	7,78%
Valle del Cauca	8,90%	8,92%	8,81%	8,24%	8,22%	8,61%	8,62%	8,56%	8,99%	8,82%	9,59%
Atlántico	9,02%	8,59%	9,04%	9,21%	8,83%	9,01%	8,50%	8,74%	9,22%	8,69%	9,35%
Nariño	8,49%	8,90%	8,69%	8,99%	9,21%	9,41%	9,18%	9,63%	10,24%	9,52%	10,58%
La Guajira	8,09%	8,34%	9,32%	9,06%	8,96%	9,38%	9,13%	9,48%	9,39%	8,82%	9,69%
Magdalena	7,66%	7,79%	8,20%	7,93%	7,77%	7,71%	8,33%	8,43%	8,53%	7,57%	8,29%
Meta	6,57%	5,71%	5,95%	5,94%	5,75%	5,99%	6,05%	6,09%	6,38%	5,91%	6,07%
Norte de Santander	6,19%	6,49%	6,20%	5,92%	5,85%	6,37%	6,35%	6,99%	7,23%	7,05%	7,40%

Fuente: DANE, Estadística Vitales 2011 - 2021

Gráfica 1. Bajo peso al nacer por Departamentos 2011 - 2021



Fuente: DANE, Estadística Vitales 2011 - 2021

Estas cifras revelan que la evolución del BPN, durante el periodo de 2011 a 2021, ha experimentado un aumento continuo, con una disminución moderada en los años 2014 y 2015 a nivel nacional. Al igual que con la MI, persisten disparidades en el comportamiento de este indicador a nivel territorial (Tabla 1). Es importante resaltar que los departamentos con tasas elevadas de MI en comparación con el promedio nacional, muestran proporciones similares de casos de BPN. Por ejemplo, en el año 2011, el Chocó presentó una proporción de 9,31 recién nacidos con bajo peso por cada 100 nacidos vivos, cifra no significativamente superior a la de departamentos como Antioquia y Valle del Cauca (ver Tabla 3). La información presentada permite identificar las disparidades en el comportamiento de la MI y el BPN entre los diversos departamentos y municipios del Valle del Cauca, razón por la cual, corresponde explorar y comprender ambas problemáticas a nivel de las entidades territoriales municipales, teniendo en cuenta incluso los factores sociodemográficos asociados al comportamiento de estos indicadores.

La importancia de reconocer eventos de esta índole en unidades territoriales, en los municipios del Valle del Cauca, se encuentra en la capacidad de estos análisis para influir positivamente en el diseño y fortalecimiento de políticas y estrategias específicas. Este enfoque busca no solo abordar las disparidades identificadas en indicadores como la MI, sino también sentar las bases para iniciativas que promuevan un cambio en la calidad de vida de la población.

A pesar de que los indicadores de TMI y BPN presentan aspectos destacados al analizarlos a nivel departamental, al enfocarnos en los municipios del Valle del Cauca, existe el riesgo de ocultar una realidad más compleja debido a las variaciones que pueden existir entre distintos municipios dentro del mismo departamento. Aquí, las tasas de mortalidad infantil podrían parecer más elevadas en comparación con la proporción de casos de BPN en áreas con una mayor concentración de población de bajos ingresos. No obstante, es esencial abordar estos casos con detalle para identificar posibles tendencias, si las hubiera, evitando así llegar a conclusiones poco acertadas.

Tabla 4. Bajo peso al nacer por municipios del Valle del Cauca 2011 – 2021.

Municipios	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Promedio
Bolívar	8,2%	5,8%	8,6%	20,5%	12,3%	13,0%	6,6%	11,5%	9,9%	3,7%	8,8%	9,91%
El Cairo	11,6%	15,3%	4,5%	9,9%	16,7%	9,2%	8,8%	9,5%	6,9%	5,9%	8,5%	9,69%
Jamundí	8,8%	9,8%	9,1%	9,2%	7,2%	10,1%	8,5%	10,6%	10,7%	10,3%	11,1%	9,57%
Pradera	6,5%	8,7%	9,5%	9,3%	9,2%	10,2%	9,3%	6,8%	12,4%	9,7%	13,7%	9,57%
El Cerrito	9,0%	7,8%	11,9%	8,2%	9,0%	8,7%	10,8%	8,2%	11,0%	10,1%	10,5%	9,54%
Bugalagrande	7,6%	13,6%	4,9%	10,2%	8,8%	9,8%	11,1%	10,8%	12,0%	7,4%	8,3%	9,48%
Versalles	8,6%	5,7%	12,8%	9,1%	5,3%	6,8%	10,9%	5,4%	19,7%	7,6%	10,6%	9,31%
Riofrío	9,9%	8,6%	11,9%	11,0%	7,9%	10,2%	7,1%	8,4%	9,9%	9,0%	6,5%	9,12%

Municipios	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Promedio
Cali	9,4%	9,5%	9,1%	8,6%	8,6%	8,8%	8,7%	8,9%	8,8%	9,3%	9,9%	9,06%
Buenaventura	9,1%	9,0%	8,2%	7,5%	9,5%	9,7%	9,1%	8,5%	10,4%	9,2%	8,6%	8,99%
Yotoco	10,6%	9,7%	6,2%	8,8%	8,6%	8,2%	8,4%	12,0%	6,6%	8,7%	10,8%	8,97%
El Dovio	6,2%	10,4%	12,5%	9,0%	5,7%	15,7%	8,5%	12,5%	4,7%	4,5%	8,3%	8,92%
Calima	6,8%	5,5%	6,5%	13,8%	9,8%	7,8%	8,1%	6,6%	12,8%	8,0%	11,9%	8,88%
Guadalajara De Buga	8,8%	7,7%	9,6%	8,3%	8,9%	8,0%	8,7%	8,6%	9,1%	8,5%	10,7%	8,80%
Florida	7,5%	9,5%	8,2%	8,2%	7,7%	7,9%	8,2%	9,9%	9,5%	9,2%	10,6%	8,76%
Valle Del Cauca	**8,9%**	**8,9%**	**8,8%**	**8,2%**	**8,2%**	**8,6%**	**8,6%**	**8,6%**	**9,0%**	**8,8%**	**9,6%**	**8,75%**
Andalucía	8,5%	9,9%	7,8%	6,3%	6,1%	10,2%	10,5%	8,1%	8,5%	6,0%	12,9%	8,60%
Yumbo	7,8%	9,3%	8,3%	7,7%	8,4%	9,2%	7,9%	8,8%	9,4%	8,2%	9,4%	8,57%
Candelaria	9,2%	8,7%	9,9%	8,9%	7,1%	8,3%	7,5%	7,2%	8,4%	9,4%	9,0%	8,51%
Restrepo	7,9%	7,5%	6,6%	6,7%	8,3%	7,6%	8,7%	12,4%	7,5%	10,2%	8,3%	8,34%
Guacarí	11,5%	6,3%	9,0%	8,8%	6,4%	7,0%	8,6%	8,9%	7,0%	7,9%	9,3%	8,24%
Dagua	11,3%	8,1%	8,7%	8,0%	8,7%	7,3%	4,8%	5,9%	10,1%	7,9%	9,8%	8,23%
La Cumbre	6,1%	6,4%	11,9%	6,0%	10,4%	8,8%	12,0%	3,4%	7,8%	12,8%	4,5%	8,20%
Tuluá	8,2%	8,5%	7,9%	8,6%	7,0%	7,0%	9,2%	7,7%	8,6%	7,6%	9,8%	8,19%
Alcalá	10,2%	10,9%	6,6%	6,3%	5,6%	7,4%	9,1%	4,5%	6,0%	8,4%	14,9%	8,17%
Ginebra	8,5%	8,2%	9,8%	7,0%	4,9%	5,4%	8,9%	9,0%	7,7%	7,0%	13,3%	8,17%
La Victoria	4,3%	12,9%	8,9%	8,6%	9,4%	5,7%	13,4%	2,8%	8,5%	9,6%	4,8%	8,09%
Trujillo	8,3%	5,4%	12,5%	6,8%	2,3%	8,0%	6,7%	10,8%	7,8%	7,8%	10,7%	7,91%
Palmira	7,8%	7,7%	8,4%	6,8%	7,4%	7,8%	8,9%	7,9%	8,2%	7,7%	8,3%	7,90%
Vijes	10,3%	6,7%	11,5%	5,0%	2,4%	8,0%	8,8%	10,5%	9,3%	5,1%	8,7%	7,84%
Toro	5,5%	8,1%	8,8%	13,1%	5,0%	10,3%	6,9%	6,2%	5,7%	4,4%	12,2%	7,83%
San Pedro	4,8%	6,9%	8,1%	8,8%	6,0%	6,2%	6,3%	7,9%	9,7%	11,5%	9,2%	7,77%
Roldanillo	9,4%	8,8%	8,0%	7,0%	6,0%	5,9%	9,5%	6,2%	9,4%	9,3%	5,5%	7,74%
Caicedonia	7,9%	6,5%	6,3%	7,1%	6,4%	6,6%	7,8%	7,9%	12,3%	6,7%	8,3%	7,63%
Argelia	10,4%	4,8%	8,0%	3,3%	4,7%	6,7%	9,8%	3,8%	5,6%	4,7%	20,3%	7,47%
Sevilla	9,4%	8,6%	9,2%	5,6%	6,4%	5,8%	8,3%	6,1%	7,5%	7,0%	7,8%	7,42%
Zarzal	8,7%	7,9%	7,9%	8,5%	6,7%	6,3%	6,6%	6,5%	8,8%	5,3%	7,0%	7,30%
Cartago	6,5%	5,5%	7,5%	7,5%	7,0%	8,2%	6,1%	8,0%	8,5%	7,6%	7,4%	7,24%
El Águila	4,4%	3,9%	5,6%	6,7%	8,2%	7,2%	8,9%	5,4%	8,7%	4,8%	11,8%	6,89%
La Unión	6,2%	7,9%	8,3%	6,4%	5,4%	8,1%	5,8%	4,1%	6,5%	5,3%	10,0%	6,72%
Ansermanuevo	5,0%	5,6%	6,8%	7,8%	3,8%	7,4%	5,7%	4,7%	5,7%	6,2%	10,8%	6,31%
Obando	7,4%	4,2%	7,4%	3,1%	3,9%	7,4%	5,5%	5,2%	4,9%	4,8%	4,7%	5,32%
Ulloa	1,6%	7,1%	1,9%	7,1%	4,3%	11,4%	0,0%	8,1%	0,0%	5,6%	3,1%	4,58%

Fuente: DANE, Estadística Vitales 2011 – 2021

El enfoque de este proyecto de investigación se dirigió específicamente hacia las cifras del Valle del Cauca. El objetivo principal fue analizar el comportamiento del BPNT y la MI en el periodo de 2011 a 2021, a través de la implementación de un modelo de aprendizaje automático diseñado para predecir el riesgo de mortalidad asociado al bajo peso al nacer a término en menores de un año en esa región. Esta tarea resultó desafiante, ya que cada municipio del departamento presenta sus propias características.

3.1.2 Modelos de aprendizaje automático

El uso de técnicas de aprendizaje de máquina, como un área de la Inteligencia Artificial (IA) es de gran ayuda ya que, a través de la definición de algoritmos, modelos y rutinas de análisis de datos para generar aprendizajes automatizados, permite identificar patrones y a partir de ahí, tomar decisiones. Y esto ha venido cobrando mayor relevancia en el área de la salud, donde la declaración de pandemia debido al Covid-19, dejó claras evidencias de la necesidad de acoger y aplicar activamente la ciencia de datos; pero para ello, y acorde con lo expresado por la Organización Mundial de la Salud OMS y la Organización Panamericana de la Salud OPS en su Consejo Directivo No. 59 del 2021 [27], es importante que se establezcan políticas públicas que permitan aprovechar la gran cantidad de datos que hoy se produce, mediante la aplicación de estrategias de ciencia de datos, que puedan ser incluso interoperables, para impulsar la transformación digital en el ámbito de la salud.

El machine learning (aprendizaje de máquina) se centra en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender patrones a partir de datos y tomar decisiones sin ser explícitamente programadas; actualmente existen tres tipo de sistemas de machine learning, agrupados por criterios [28] que son:

Tabla 5. Tipos de sistemas de aprendizaje automático (Machine Learning)

Criterio	Tipos de sistema
Si se entrenan o no bajo supervisión humana	Aprendizaje supervisado
	Aprendizaje no supervisado
	Aprendizaje semi-supervisado
	Aprendizaje por refuerzo
Si se pueden aprender o no de forma gradual sobre la marcha	Aprendizaje online frente a aprendizaje por lotes
Si funcionan comparando simplemente puntos de datos nuevos con puntos de datos conocidos o si detectan patrones en los datos de entrenamiento y crean un modelo predictivo, como hacen los científicos.	Aprendizaje usado en instancias frente a aprendizaje basado en modelos

Fuente: Adaptación a partir de [28]

En el campo de la medicina se ha recurrido al uso de modelos entrenados bajo supervisión humana, por lo que a continuación se describen cada uno de estos:

- Aprendizaje supervisado: en este tipo de modelo, se utilizan conjuntos de datos etiquetados, es decir con entradas y salidas conocidas, para hacer predicciones o clasificaciones.

- Aprendizaje no supervisado: en este caso, el modelo no utiliza conjuntos de datos etiquetados, con lo cual es posible establecer patrones, estructuras y relaciones intrínsecas dentro de los datos.
- Aprendizaje semisuperivsado: este modelo utiliza conjuntos de datos etiquetados y no etiquetados, por ello requieren la combinación de algoritmos no supervisados y supervisados.
- Aprendizaje por refuerzo: en este caso, el modelo utiliza un agente (un programa informático o robot) que observa un entorno, selecciona y realiza acciones para recibir recompensas a cambio.

En el marco de los objetivos del presente proyecto, en adelante se describen los algoritmos más utilizados en la construcción de modelos de aprendizaje supervisado.

3.1.2.1. Regresión logística

La regresión logística es un algoritmo utilizado con frecuencia para determinar la probabilidad de una instancia pertenece a una clase concreta; si esta probabilidad es mayor del 50%, el modelo predice que la instancia pertenece a esa clase (positiva, "1") y en caso contrario predice que no (negativa, "0"), por ello es un clasificador binario. Este algoritmo utiliza la función de probabilidad logística que se denota a continuación [28], pp. 166:

$$\hat{p} = \ h_0(x) = \sigma(x^T\theta)$$

$\hat{p}$ es la probabilidad estimada de que una instancia pertenezca a la clase positiva ($y = 1$).
$h_0(x)$ es la función de hipótesis, que toma las variables de entrada (x) y produce la salida $\hat{p}$.
$\sigma(.)$ es la función sigmoide (gráficamente conforma de "S") que transforma la salida de la función lineal entre 0 y 1.
$x^T\theta$ representa el producto escalar de los vectores x (características de entrada) y θ (parámetros del modelo)

En detalle su ecuación es la siguiente:

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

Una vez calculada la $\hat{p} = \ h_0(x)$, se puede predecir $\hat{y}$: cuando $\hat{p} \geq 0.5$ entonces $\hat{y} = 1$, pero cuando $\hat{p} < 0.5$ entonces $\hat{y} = 0$.

Ahora bien, para entrenar el modelo, esto requiere inicializar los pesos θ y el sesgo b_0; definir una función de costo como la entropía cruzada para evaluar la discrepancia en las predicciones del modelo y etiquetas reales del conjunto de entrenamiento; posteriormente, ajustar los parámetros utilizando un algoritmo de optimización como el del Descenso de Gradiente que ajusta iterativamente los valores de θ hasta alcanzar la convergencia a los valores que minimizan la función de costo.

3.1.2.2. Máquinas de vectores soporte MSV

Este algoritmo MSV tiene la capacidad de realizar clasificaciones lineales o no lineales, regresiones y detección de valores atípicos; funciona apropiadamente con la clasificación de conjunto de datos complejos de mediano o menor tamaño [28], pp. 178.

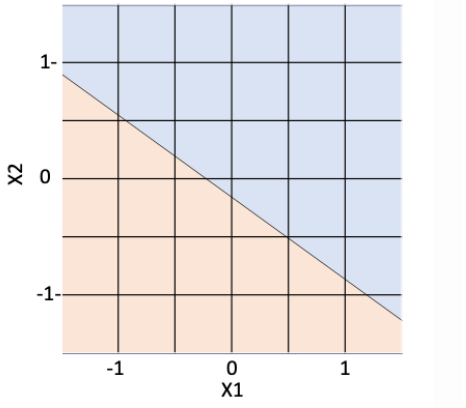
Además de ello, las MSV pueden eficientemente realizar clasificación no lineal mediante la técnica kernel trick. Este método implica un mapeo implícito de las entradas a espacios de características de alta dimensión en la máquina [29].

La MSV se basa en la noción de clasificador de margen máximo, un concepto asociado a un hiperplano [30]. Un hiperplano es un subespacio plano que separa en dos mitades las clases del dominio original. En el caso de dos dimensiones un hiperplano corresponde con la ecuación de la recta:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0$$

Un punto x que satisface la ecuación pertenece al hiperplano. A su vez, un punto que no satisface la ecuación se ubica a los lados del hiperplano. El lado donde se ubica un punto con respecto al hiperplano se determina mediante el signo de la ecuación. Véase la siguiente ilustración.

Ilustración 1. Hiperplano de separación



Fuente: Adaptación de [30]

El algoritmo consiste en mapear los datos del dominio y dividir los datos mediante el uso de hiperplanos. Un hiperplano permite definir un clasificador para predecir a que grupo pertenece un dato en función de sus predictores. Es necesario establecer un clasificador optimo cuando la distribución de los datos es separable linealmente mediante un número infinito de hiperplanos. El hiperplano óptimo de separación es aquel que presenta un mayor margen, es decir, que la mayor distancia posible entre el hiperplano de separación y las instancias a los lados. Esta condición reduce el límite superior del error de generalización esperado [31].

3.1.2.3. Árboles de decisión

Los árboles de decisión son algoritmos usados para tareas de clasificación y regresión, representando las decisiones mediante una forma de árbol con nodos y hojas, cada nodo representa una prueba o elección sobre una característica de entrada, mientras que las hojas representan una etiqueta o valor de salida [32].

La construcción de un árbol de decisión requiere inicialmente la selección de atributos, lo que implica reconocer cuales características del conjunto de datos ofrecen la mayor información para realizar la división de los nodos; posteriormente se hace la división de los nodos, es decir la división de un nodo en dos hijos (binario) o más (múltiple). Y, finalmente, se realiza la poda del árbol, mediante la cual se eliminan las ramas que no se requieren; esto puede realizarse antes (pre-pruning) o después de construir el árbol (post-pruning) con el fin de evitar el desarrollo excesivo del árbol, evitar el sobreajuste, como también para obtener una mayor precisión según el tipo de árbol de decisión (clasificación o regresión). Para elaborar árboles de decisión tanto de clasificación y de regresión, se utiliza el algoritmo de su clase CART (Classification and Regression Tree).

En la creación de árboles de decisión de clasificación, una vez realizada la selección de atributos y las divisiones o nodos posibles, se calcula el índice Gini para cada nodo, buscando la partición más homogénea. Este proceso se repite iterativamente para los nodos resultantes hasta obtener un árbol entrenado; para realizar una predicción para una nueva entrada, se sigue el camino del árbol basándose en los valores de sus variables predictoras hasta alcanzar uno de los nodos finales.

Por otro lado, en árboles de decisión de regresión, se busca predecir valores continuos. Se determinan nodos basados en las características y se evalúan las particiones resultantes, calculando los errores cuadráticos medios. Luego, se selecciona el nodo con el menor costo, y este proceso se repite iterativamente hasta obtener un modelo de

regresión eficiente. Al realizar una nueva predicción se recorren los nodos del árbol hasta llegar a un nodo terminal; en este último, se utiliza el valor promedio de las observaciones en ese nodo, como la predicción de la nueva entrada [33].

3.1.2.4. Bosque aleatorio

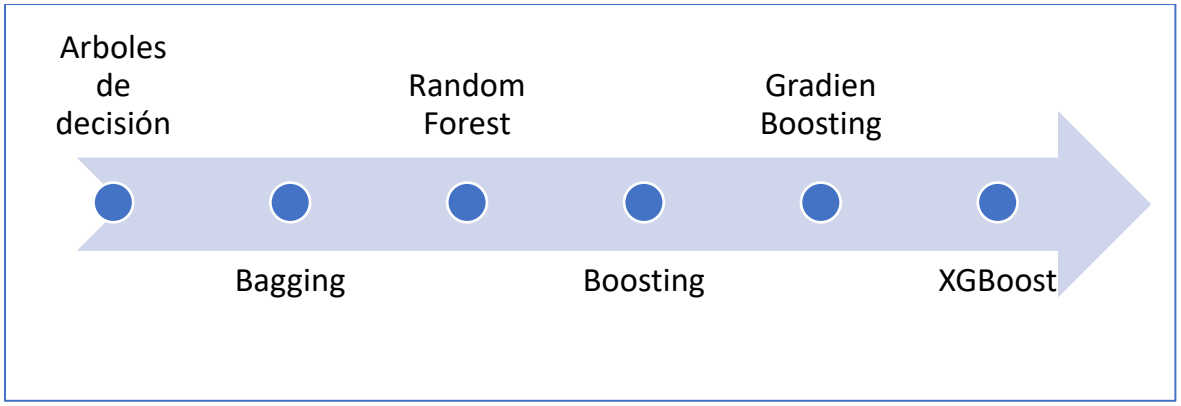
Este algoritmo consiste en una combinación o "ensamble de árboles de decisión, entrenado, por lo general, mediante el bagging [28] pp. 216-217 o bootstrap aggregating. Mediante esta técnica se seleccionan subconjuntos de datos de manera aleatoria, así cada árbol se entrena con un conjunto de características diferentes; el propósito consiste en crear una amplia variedad de árboles que equilibre un mayor sesgo con una menor variabilidad, resultando típicamente en una mejora general del rendimiento del modelo.

En el caso específico de la mortalidad infantil por BPN, esta metodología permite capturar la complejidad de las relaciones entre las diversas variables asociadas, como las características del recién nacido y las condiciones sociodemográficas de la madre, contribuyendo a un modelo más robusto y adaptado a las particularidades de la problemática local.

3.1.2.5. XGBoost

Este es un algoritmo de aprendizaje automático escalable para impulsar árboles creado por Chen, disponible además en un paquete código abierto y ha sido ampliamente reconocido por su capacidad para manejar grandes cantidades de datos, con alta precisión [34]. Está basado en arboles de decisión y se considera que es la evolución de estos, de la siguiente manera [35]:

Ilustración 2. Evolución de los algoritmos basados en árboles de decisión



Fuente: Adaptación propia de la Figura No.2 tomada de [35]

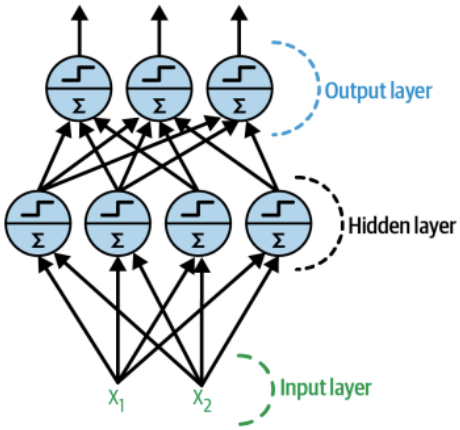
Acorde con Zúñiga [35], este algoritmo se conforma de un equipo o ensamble de varios árboles de decisión que trabajan juntos. Cada árbol se construye uno después del otro y se ayuda a corregir los errores cometidos por los árboles anteriores. Esto mejora continuamente el rendimiento del ensamble. A diferencia de los Bosques aleatorios, en XGBoost, el usuario puede decidir cuánto deben crecer los árboles, mientras que en el primero, los árboles crecen tanto como sea posible. En este algoritmo, se utilizan técnicas como la regularización (optimización que penaliza la complejidad del modelo) para evitar que el modelo se sobreajuste demasiado a los datos de entrenamiento y para prevenir que tenga un sesgo demasiado alto. El objetivo es lograr un equilibrio para que el modelo pueda hacer predicciones precisas tanto en los datos de entrenamiento como en nuevos datos que no ha visto antes.

3.1.2.6. Redes Neuronales Artificiales RNA

Las redes neuronales artificiales RNA existen desde 1943 cuando el neuropsicólogo Warren MacCulloch y el matemático Walter Pitts “desarrollaron un modelo computacional sencillo de cómo podrían trabajar juntas las neuronas bilógicas en los cerebro animales para realizar computaciones complejas usando lógica posicional” [28], pp. 292.

Posteriormente, en 1957 el psicólogo Frank Rosenblant introdujo el perceptrón, como la unidad básica de procesamiento en las RNA y forma la base de modelos más complejos de aprendizaje profundo. El perceptrón toma múltiples entradas binarias, realiza operaciones ponderadas en ellas, y produce una única salida binaria. Esta unidad básica de procesamiento cuenta con entradas y cada conexión de entrada, está asociada a un peso que se suma de manera ponderada y finalmente mediante una función escalonada produce un resultado único. En RN profundas, se cuenta con más de un perceptrón generándose capas de entrada (varias entradas), capas ocultas (varias neuronas que procesan información) y capas de salida (neuronas que generan los resultados o salidas); lo anterior, se conoce como el perceptrón multicapa PMC. Véase la siguiente ilustración.

Ilustración 3. Arquitectura del Perceptrón Multicapa PMC



Fuente: Figura 10.7. Arquitectura de un perceptrón multicapa con dos entradas, una capa oculta de cuatro neuronas y tres neuronas de salida. [23, pp. 299].

El proceso de entrenamiento de una RNA implica en estos casos el uso de los algoritmos de Retropropagación (Backpropagation) y de Descenso de Gradiente, que en conjunto permiten ajustar los pesos de la red, de forma que la salida se acerque lo mejor posible a la salida deseada, es decir, reduciendo el error.

En conclusión, las RNA, entrenadas mediante la Retropropagación y el Descenso del Gradiente, son una alternativa conveniente para construir modelos de aprendizaje automático pues tienen la capacidad de aprender patrones complejos y representar relaciones no lineales, aumentando la precisión en las predicciones asociadas a la problemática de la clasificación de la muerte del recién nacido frente al BPNT.

3.1.2.7. Naive Bayes

Según Mahesh [29] este es un algoritmo de clasificación que se basa en el teorema de Bayes, asumiendo independencia entre características; esto implica que la presencia de una característica no está relacionada con las demás. Este enfoque, especialmente útil en clasificación de texto, se emplea para agrupar y clasificar basado en probabilidades condicionales.

El entrenamiento del modelo, consiste en calcular las probabilidades condicionales de las características de un conjunto de datos; las métricas de precisión, recall, F1-score y curva ROC, son apropiadas para evaluar el rendimiento del modelo y una vez entrenado, el modelo está en capacidad de hacer predicciones con nuevos datos.

Este método, reconocido por asumir la independencia entre características, resulta particularmente valioso al clasificar en base a probabilidades condicionales, siendo especialmente aplicable en problemas complejos como la mortalidad infantil y su relación con el BPNT.

3.1.2.8. K-vecinos más cercanos

Este algoritmo se usa en problemas de regresión y clasificación [36]; su funcionamiento consiste en identificar los k vecinos más cercanos para un punto determinado. Se fundamenta en que los puntos más cercanos entre si son considerados similares, y la medida de cercanía se determina mediante la distancia Euclidiana entre dos puntos. El objetivo es encontrar la clase más grande de elementos que se encuentran cerca de los datos de prueba, lo que permite concluir que los datos pertenecen a esa clase.

El procedimiento del algoritmo consiste en seleccionar un valor de k. Se recomienda un valor alto de k para prevenir que la clasificación se vea afectada por datos etiquetados erróneamente; posteriormente, se identifican los k vecinos más cercanos y, finalmente se realiza una predicción del valor para un nuevo dato, basándose en las etiquetas de los k vecinos más cercanos, usando algún criterio específico.

En problemas de regresión se utiliza como criterio el promedio de los valores de etiqueta de los k vecinos más cercanos. En cambio, en problemas de clasificación, se puede usar la mayoría.

3.1.3. Métricas de evaluación de Modelos de Aprendizaje Automático

En la evaluación de un modelo de aprendizaje automático, la selección de métricas apropiadas desempeña un papel crucial en la comprensión de su rendimiento. La elección de métricas debe ser estratégica, alineándose no sólo con los objetivos generales del modelo sino también con las particularidades y requisitos específicos del problema en estudio.

En este contexto, la **Matriz de confusión** se constituye como una herramienta técnica de gran utilidad. Al desglosar las predicciones del modelo en cuatro categorías fundamentales: verdaderos positivos (VP), falsos positivos (FP), verdaderos negativos (VN) y falsos negativos (FN), la matriz de confusión proporciona una visión detallada y cuantitativa del rendimiento del modelo. La matriz se establece como lo muestra la siguiente ilustración [37].

Ilustración 4. Matriz de Confusión

Fuente: Adaptación Figura 5.3 de [37]

Para este proyecto de investigación, la interpretación de la matriz de confusión es fundamental para evaluar el rendimiento del modelo de aprendizaje automático, la cual se interpreta de la siguiente manera:

- Los Verdaderos Positivos (VP) muestra la cantidad de casos en los que el modelo acertó al identificar a recién nacidos con BPNT que murieron.
- Los Falsos Positivos (FP) indican la cantidad de casos que el modelo clasificó erróneamente a recién nacidos con BPNT, como si hubieran muerto.
- Los Verdaderos Negativos (VN) representan la cantidad de casos identificados correctamente de nacimientos con BPNT, que no murieron.
- Los Falsos Negativos (FN) revelan la cantidad de casos en los que el modelo no logró clasificar adecuadamente a recién nacidos con BPNT, que si murieron.

Desde este punto de vista, la matriz de confusión se integra en el cálculo de otras métricas importante que determinan el rendimiento de los modelos de aprendizaje que a continuación se describen [28]:

- **Exactitud (Accuracy)**, que se determina como la proporción de predicciones correctas (VP y VN) sobre el total de predicciones (VP+VN+FP+FN); de este modo proporciona una visión general de la precisión global del modelo y se expresa mediante la fórmula.

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN}$$

- **Precisión**: se calcula como la proporción de verdaderos positivos (VP) entre la suma de verdaderos positivos y falsos positivos (FP). La fórmula de precisión es la siguiente:

$$Presición = \frac{VP}{VP + FP}$$

- **Sensibilidad (recall):** conocida como tasa de verdaderos positivos (VP), mide la capacidad del modelo para identificar correctamente instancias positivas. Se calcula de la siguiente manera:

$$Sensibilidad = \frac{VP}{VP + FN}$$

- **Especificidad (specificity):** mide la capacidad del modelo para identificar correctamente los casos negativos en relación con el total de casos negativos reales. La fórmula es:

$$Especificidad = \frac{VP}{VN + FP}$$

- **F1 Score**: esta es una métrica que combina la Precisión y el Sensibilidad (Recall) en una sola medida. Se calcula utilizando la siguiente fórmula:

$$F1 - Score = \frac{VP}{VP + \frac{FN + FP}{2}}$$

- Área bajo la Curva (AUC – ROC): esta métrica proporciona un valor único que cuantifica la capacidad general del modelo para clasificar correctamente los casos positivos y negativos; un AUC-ROC más cercano a 1 indica un modelo más efectivo en la clasificación, mientras que un valor cercano a 0,5 sugiere un rendimiento similar al azar.

  Considerando el objetivo de este proyecto de investigación sobre los nacimientos con Bajo Peso al Nacer a Término (BPNT) en el Valle del Cauca, la AUC-ROC resulta útil para medir la habilidad del modelo en distinguir entre recién nacidos que fallecieron y los que sobrevivieron.

3.1.4. Métodos de balanceo de clases para conjuntos de datos desbalanceados

La aplicación de modelos de aprendizaje automático en el ámbito de la salud ha emergido como una herramienta valiosa para abordar diversas problemáticas, sin embargo, enfrenta un obstáculo común y significativo: el desbalanceo de clases en los conjuntos de datos [38]. Este problema, ampliamente documentado en el campo de la ciencia de datos, consiste en una distribución desproporcionada de instancias entre las clases de interés, y es un desafío que se torna especialmente crítico en la predicción de eventos como la MI asociada al BPNT.

La prevalencia de datos desbalanceados, como se evidencia en estudios del sector salud, introduce complejidades sustanciales en el entrenamiento y evaluación de modelos de clasificación. Para el caso de este proyecto de investigación, la tarea de identificar con precisión a los recién nacidos con BPNT en riesgo de MI se ve afectada por este tipo de problema. Esto se debe a que los algoritmos de clasificación se entrenan mayoritariamente con un mayor número de muestras relacionadas con casos normales, lo cual resulta en un sesgo hacia esta clase. En consecuencia, la escasez de ejemplos positivos en comparación con los casos normales, se traduce en una dificultad adicional para los modelos al intentar clasificar la clase minoritaria asociada a los recién nacidos con BPNT que, desafortunadamente, mueren.

Este desbalanceo del conjunto de datos afecta adversamente el rendimiento de los modelos, ya que al ser entrenados sobre ese conjunto de datos, existe el riesgo de que se favorezca la clase mayoritaria, resultando en clasificaciones sesgadas y subóptimas [38].

La literatura también resalta que el uso de métricas de rendimiento global, como la precisión (accuracy), puede ser engañoso en este escenario. Aunque un modelo puede alcanzar una alta precisión general al clasificar la mayoría de las instancias, su capacidad para prever eventos relacionados con la clase minoritaria puede ser insuficiente [39]. En consecuencia, la atención médica prenatal y otros campos de la salud, demandan un enfoque más cauteloso para evitar la subrepresentación de eventos críticos y garantizar la fiabilidad de los modelos en situaciones de datos desbalanceados.

Para afrontar este problema de desbalanceo de las clases en este trabajo de investigación, nos enfocamos específicamente en abordar el problema a través de las siguientes propuestas:

- **Sobre-muestreo:** Esta técnica también conocida como "Oversampling" es un enfoque utilizado para contrarrestar el desbalance en la distribución de clases en conjuntos de datos. Su principal objetivo es equilibrar la proporción entre las clases, incrementando el número de muestras de la clase minoritaria, mientras mantiene constante la cantidad de la clase mayoritaria.

  Esta técnica puede ser efectiva para equilibrar el conjunto de datos desbalanceados para la predicción de la MI asociada al bajo BPNT; también implica el incremento aleatorio de casos de la clase minoritaria que para este proyecto, serian específicamente los recién nacidos con BPNT que lamentablemente fallecieron. El incremento se lleva a cabo hasta alcanzar la proporción deseada entre las clases, lo que facilita la creación de un conjunto de datos más equilibrado.

Es importante destacar que el sobre-muestreo puede conllevar el riesgo de sobreajuste del clasificador, si se aplica de manera excesiva; en otras palabras, al incrementar la clase minoritaria, puede conducir a que el modelo memorice específicamente los datos de entrenamiento, perdiendo así su capacidad de clasificar adecuadamente nuevos datos [40].

Dentro del contexto de las técnicas de sobre-muestreo, diversas investigaciones [40] [41] [42] destacan el Synthetic Minority Over Sampling Technique (SMOTE) como uno de los métodos más representativos. A diferencia del simple incremento de instancias existentes en la clase minoritaria, SMOTE adopta un enfoque más avanzado, generando instancias sintéticas mediante la interpolación entre observaciones de esta clase. La metodología de SMOTE se basa en la creación de nuevas muestras a lo largo de las líneas que conectan instancias vecinas en el espacio de características. Este proceso tiene como objetivo principal introducir una variabilidad adicional al conjunto de datos de la clase minoritaria, mejorando así la capacidad del modelo para generalizar patrones y mitigar el riesgo de sobreajuste.

La principal ventaja de SMOTE radica en su capacidad para abordar el desafío del sobreajuste, al proporcionar datos sintéticos que reflejan la diversidad del conjunto de datos original. Al introducir la variabilidad adicional, esta técnica mejora la robustez del modelo, permitiéndole aprender patrones más representativos y generalizables asociados con la MI en casos de BPNT.

- **Submuestreo:** Esta técnica también llamada como “Undersampling” en inglés, se focaliza en contrarrestar el desbalance de clases reduciendo aleatoriamente el número de muestras de la clase mayoritaria, manteniendo constante la cantidad de la clase minoritaria. Es decir, se eliminan al azar muestras seleccionadas de la clase mayoritaria con el objetivo de igualar la proporción entre las clases en un conjunto de datos [40]. Aunque su implementación es sencilla y puede resultar en una reducción del tiempo de procesamiento de datos, existe el riesgo inherente de eliminar muestras potencialmente significativas durante el proceso de clasificación.

  Para el caso de esta investigación la aplicación del submuestreo aleatorio podría realizarse disminuyendo aleatoriamente instancias de recién nacidos con BPNT que no experimentaron mortalidad. Esta estrategia busca equilibrar las clases y focalizar el aprendizaje del modelo en la clase minoritaria (BPNT asociada a la MI); no obstante, se corre el riesgo de eliminar muestras potencialmente importantes.

Para mitigar este riesgo, se han desarrollado algoritmos que realizan una selección inteligente sobre las muestras de la clase mayoritaria, permitiendo una reducción más precisa y enfocada. Estos algoritmos buscan preservar la representatividad y la información crítica de la clase mayoritaria, mitigando así la posibilidad de pérdida sustancial de conocimiento durante el proceso de submuestreo.

La técnica "Tomek links", conceptualizada por [43], se fundamenta en la proximidad entre vecinos más cercanos. También, se define como un par de muestras de clases distintas que comparten una similitud significativa

Este método, en su implementación técnica, se enfoca en la identificación y eliminación de muestras redundantes o cercanas en la clase mayoritaria, con respecto a las muestras de la clase minoritaria; en otras palabras, este algoritmo busca activamente estos pares en el espacio de características y procede a la eliminación de la instancia perteneciente a la clase mayoritaria.

En nuestro caso, la clase mayoritaria representaría a los recién nacidos con BPNT que no experimentaron mortalidad, mientras que la clase minoritaria corresponde a aquellos con BPNT que lamentablemente fallecieron. La eliminación estratégica de instancias de la clase mayoritaria en estos pares, contribuirá a establecer grupos más definidos en el conjunto de datos, lo que, en términos de la MI, implica identificar de manera más precisa los casos críticos asociados al BPNT.

3.2. Antecedentes

Para este ejercicio de investigación se revisaron un conjunto de investigaciones relacionadas con el uso de modelos de aprendizaje automático aplicados al BPN y a la mortalidad infantil, que, mediante el uso de las técnicas de predicción y clasificación, analizaron las dinámicas de ambos fenómenos. A continuación, se presenta de manera general la metodología y resultados generados en cada uno de estos estudios y cómo se usó la ciencia de datos para dar respuesta al problema identificado. Finalmente, se exponen las diferencias encontradas frente a la presente investigación y los beneficios que ofrecen para el desarrollo de este.

En 2018 [44] el grupo de investigación de la Facultad de Matemáticas y Ciencias Naturales de la Universidad de Brawijaya en Indonesia llevó a cabo un estudio utilizando datos de la Encuesta Demografía y Salud de Indonesia (IDHS) del 2012. Esta investigación se centró en la aplicación de técnicas de aprendizaje automático para desarrollar un modelo capaz de clasificar y predecir si una madre experimentaría un nacimiento con bajo peso al nacer (BPN), considerando variables sociodemográficas como nivel educativo, edad, número de hijos y lugar de residencia. Se definieron dos

modelos, que luego fueron evaluados: la regresión logística binaria y los bosques aleatorios. Aunque la regresión logística demostró un rendimiento aceptable en la predicción de datos de BPN, mostró limitaciones en la clasificación, evidenciadas por la curva ROC y el AUC (área bajo la curva) de 0,505, considerablemente lejos del ideal 1, indicando que no es un buen predictor. En contraste, los bosques aleatorios exhibieron un mejor rendimiento tanto en predicción como en clasificación, con un error del 7% en datos de entrenamiento, indicando una precisión del 93%. Esta técnica se destaca como la más recomendable para la clasificación y predicción del parto con BPN, proporcionando valiosas pautas para nuestro proyecto y respaldando la selección de variables críticas en el desarrollo del modelo.

Para el año 2019 [45] este mismo grupo de investigación, de la Facultad de Matemáticas y Ciencias Naturales de la Universidad de Brawijaya en Indonesia, y de acuerdo con las sugerencias de la anterior investigación [44], realizó con los mismo datos y conservando los mismos objetivos, la aplicación del algoritmo de Máquinas de vectores soporte (SVM), con el cual se buscaba clasificar los nacimientos con BPN y a su vez comparar su desempeño con la Regresión Logística Binaria (RLB). Al desarrollar las etapas de un proyecto de ciencia de datos, la investigación concluye que la SVM con sus distintas funciones kernel, funciona adecuadamente para predecir la clasificación binaria de los recién nacidos con BPN; además resultó ser un buen predictor ya que su error promedio fue inferior al 10%. Es relevante destacar que, la función kernel lineal tuvo un mejor rendimiento que las SVM con otras funciones de kernel propuestas y demostró mejores resultados en comparación con la Regresión Logística Binaria. A diferencia de nuestra investigación, este estudio tuvo como objetivo primordial probar la eficacia de la predicción del BNP utilizando una herramienta de aprendizaje automático SVM y comparar sus resultados con la herramienta tradicional estadística RLB. Aunque en ésta investigación no se revisan aspectos de la mortalidad infantil, sino únicamente del BNP asociados a las condiciones sociodemográficas de la madre, el proyecto aporta un conocimiento importante a tener en cuenta frente a las prácticas de ciencias de datos utilizadas y los rendimientos de cada modelo.

Por su parte, en el 2022 [46], se usaron datos de 7.472 registros de nacimientos de la Red Neonatal Coreana. En este caso, el estudio se concentró en neonatos de sexo masculino nacidos antes de término, es decir con una edad gestacional entre las 24 y las 36 semanas de gestación, y que catalogaron de Muy Bajo Peso al Nacer MBPN. Se excluyeron los nacimientos con problemas genéticos, aquellos con más de 37 semanas de gestación y aquellos sin datos completos.

Para el análisis de los datos, se tuvieron en cuenta 11 variables predictoras:

- Factores neonatales: sexo masculino, edad gestacional, puntajes de Apgar a los 5 min, temperatura corporal y reanimación al nacer.

- Factores maternos: diabetes mellitus, hipertensión arterial, corioamnionitis, ruptura prematura de membranas, esteroides prenatales y parto por cesárea.

Para predecir la mortalidad de los recién nacidos con MBPN, se aplicaron las siguientes técnicas: Red neuronal artificial (ANN), Bosque aleatorio (RF), Máquina de Vectores Soporte (SVM), Regresión logística (LR). En cada caso, se encontró que el rendimiento del AUC de la ROC, se comportó de forma similar en los siguientes modelos así: ANN 0,845 (0,815–0,875), RF 0,826 (0,795–0,858), LR 0,841 (0,811–0,872). Sin embargo, para la SVM fue 0,631 (0,578–0,683). Frente a ello, los autores señalaron que la predicción de la mortalidad de neonatos con MBPN de sexo masculino con las técnicas de ANN y RF, arrojaron la misma tasa de predicción que el modelo estadístico tradicional LR; pero no es concluyente en establecer cuál de todos sería el más apropiado y por ello sugiere continuar realizando otras investigaciones.

En conclusión, dicha investigación comparte con este proyecto, abordar la mortalidad en neonatos; sin embargo, se encontraron algunas diferencias: el artículo indica que se utilizaron datos de recién nacidos de sexo masculino y se consideraron únicamente los nacimientos entre la semana 24 y antes de la 36 de gestación. En cambio, nuestro proyecto de investigación

Por otro lado, un artículo científico en los Emiratos Árabes [47] tuvo como objetivo evaluar el rendimiento de 30 algoritmos de Aprendizaje Automático tanto para la estimación del peso corporal infantil como para la clasificación del BPN. Los experimentos se realizaron en un conjunto de datos de creación propia con 88 características incluyendo el peso corporal infantil como etiqueta objetivo. Para este caso combinaron múltiples subconjuntos de características para realizar predicciones con y sin técnicas de selección de características, además de emplear la técnica de sobremuestreo de minorías sintéticas para sobremuestrear la clase minoritaria.

Se encontró que los datos tenían un desequilibrio significativo razón por lo cual se utilizó la técnica Synthetic Minority Over-sampling Technique (SMOTE) para equilibrar los datos con múltiples proporciones de sobremuestreo en la clasificación de BPN. También se encontró un desequilibrio de clases que afectó significativamente el rendimiento de los algoritmos de Aprendizaje Automático, generando resultados sesgados hacia la clase mayoritaria o frente a la clasificación errónea de todas las instancias minoritarias, por lo que se también se utilizó esta misma técnica para sobreestimar la clase minoritaria.

Así mismo utilizaron diferentes métricas de rendimiento para evaluar los resultados de cada algoritmo y su desempeño como lo son exactitud, precisión, recuperación, F1 score y la matriz de confusión. Para la estimación del peso corporal se empleó la técnica de validación cruzada de diez veces para las predicciones y para la clasificación de bajo

peso al nacer se analizó el rendimiento de clasificación de múltiples clasificadores. La mejor estimación de peso se obtuvo mediante el Modelo de Bosques Aleatorios, y el mejor rendimiento se obtuvo utilizando las técnicas de sobremuestreo LR y SMOTE, obteniendo para este caso valores de exactitud, precisión, recuperación y F1 score de 90,24 %, 87,6 %, 90,2 % y 0,89, respectivamente.

A partir de esto, se concluyó que la diabetes, la edad gestacional y la hipertensión son características importantes para la estimación del peso corporal y la clasificación del BPN. Según lo anterior, la investigación expuesta en este artículo aportó a nuestra investigación la selección de métodos de clasificación y regresión más efectivos y eficientes, aunque únicamente enfocados al BPN.

Otro artículo publicado en el 2021 [48], cuyo objetivo fue determinar los factores relacionados con la mortalidad infantil, utilizando algoritmos de minería de datos. Este estudio de casos y controles se realizó teniendo como población objetivo 2.386 madres (1.076 casos y 1.310 controles) de ocho provincias de Irán. En esta investigación se utilizaron los siguiente algoritmos: Clasificador AdaBoost, Máquina de Vectores de Soporte, Redes Neuronales Artificiales, Bosques Aleatorios, K-Vecinos más cercano y Naive Bayes.

Se evaluaron la precisión y discrepancias de los datos usando el recall (sensibilidad), la especificidad, la precisión, el accuracy (exactitud) y el F1 Score para medir y comparar el rendimiento de los métodos de clasificación. Luego utilizaron seis algoritmos para indicar las variables más relevantes para estudiar la mortalidad infantil, encontrando que factores como la edad de la madre en el momento del embarazo, el lugar de residencia, la alfabetización, el trabajo de la madre, el matrimonio consanguíneo, la brecha del embarazo, el peor evento de la vida, el tabaquismo durante el embarazo, el sexo del niño, los gemelos, los trastornos dentales, el síndrome psicológico, la diabetes gestacional, la presión arterial alta y la anemia durante el embarazo, fueron seleccionados como predictores. Luego aplicaron el modelo de regresión logística binaria para definir el papel de cada predictor que fue seleccionado.

Para comparar la predicción de los resultados, se utilizaron siete modelos estadísticos: la Regresión Logística (tradicional) y seis algoritmos de machine learning Clasificador Adaboost, Maquinas de vectores soporte, Redes Neuronales, Bosques Aleatorios, K-vecino más cercano, y Naive Bayes. Los resultados de eficacia indicaron que Naive Bayes y Random Forest, presentaron mejores resultados en las métricas AUC, el F1 Score, la precisión y el recall (sensibilidad) a diferencia de los otros algoritmos, demostrando tener las mejores métricas para la predicción.

La selección de variables predictoras y el análisis sobre el comportamiento de los modelos, su precisión y sensibilidad, fueron aspectos tenidos en cuenta en la presente investigación, aun cuando este estudio reseñado haga alusión únicamente a la mortalidad infantil.

En la investigación [49], se comparó la regresión logística con otras técnicas de aprendizaje automático para identificar las variables predictoras más influyentes y el desarrollo de un sistema para ayudar a los médicos a tomar mejores decisiones en el parto con bajo peso. Los algoritmos utilizados en este trabajo fueron Máquinas de Vectores de Soporte, Regresión Logística, Redes Neuronales, Naive Bayes, Random Forest y Árbol de Decisión. Las variables que más influencia tuvieron en la predicción del parto de bajo peso fueron: la edad de la madre, el número de visitas al médico durante el primer trimestre y el número de partos prematuros anteriores.

En la validación de los datos se encontró que el árbol de decisión presenta una mayor precisión de predicción en general Accuracy (89.95%), specificity (72.88%) y AUC (93.80%) F-value (93.04%) y Precisión (88.81%) en comparación con los otros métodos de clasificación. Así mismo, el área bajo la curva del árbol de decisión, fue la más alta de los demás métodos, lo que significa que este método clasifica muy bien a los recién nacidos de BPN. Las conclusiones arrojan que una de las principales causas de la mortalidad infantil es el bajo peso al nacer y el parto prematuro, siendo los factores que más influyen el peso de la madre antes de quedar embarazada, su edad, el número de visitas al médico durante el primer trimestre y el número de partos prematuros.

La coincidencia en las variables predictoras entre este y otros estudios referenciados, facilitó la selección de las variables en el actual trabajo, para el estudio de la mortalidad infantil y el BPN.

Otro artículo científico [50] analizó la mortalidad infantil aplicando métodos avanzados de aprendizaje automático para predecirla con base en la información de la Encuesta de Demografía y Salud de Ruanda 2014-2015. Las variables independientes seleccionadas fueron: lugar de residencia, estado civil, educación materna, ocupación materna, índice de riqueza, edad de la madre al primer parto, sexo del hijo, orden de nacimiento, intervalo entre nacimientos, hijos nacidos vivos, lactancia materna, acceso al agua potable, tipo de instalaciones sanitarias y tipo de combustible para cocinar.

En esta investigación fueron usados los siguientes algoritmos la regresión logística, los bosques aleatorios, el árbol de decisión y clasificadores de máquina de vectores soporte, para construir el modelo predictivo de la mortalidad infantil. Al revisar la métricas de los modelos trabajados, la capacidad de predicción de estos oscilo entre el 68,6% y 61,5%. El modelo seleccionado por los autores fue bosques aleatorios aunque su predicción fue

las más baja con respecto a los otros modelos, es el que presenta las mejores métricas con un accuracy (84,3%), recall (91,3%), precisión (80,3%), F1 score (85,5%) y AUROC (84,2%). En general todos los modelos indicaron que las 4 variables predictoras más importantes en la predicción de la mortalidad infantil fueron el estado civil de la madre, los hijos nacidos vivos, el orden de nacimiento y nivel socioeconómico de la madre.

Cabe decir que, si bien este estudio es generoso en cuanto a la identificación de factores de riesgo asociados a la mortalidad infantil usando métodos de aprendizaje automático, no es posible evidenciar la relación entre ésta y el BPN, ya que el peso al nacer no fue una variable tenida en cuenta.

Por otro lado, Jepkorir Sawe [51] elaboró un modelo de aprendizaje automático para anticipar el bajo peso al nacer, haciendo uso de los factores de riesgo maternos asociados con esta condición (BPN). Este estudio se realizó con los datos recopilados a través de la Encuesta Demografía y Salud de Kenia del 2014, y su población objetivo fueron las madres con edades comprendidas entre 15 y 49 años. Los algoritmos de aprendizaje automático implementados para esta investigación fueron la regresión logística, árboles de decisión, bosque aleatorio, máquinas de vectores soporte, aumento de gradiente y aumento de gradiente extremo. Para la evaluación del rendimiento, se emplearon métricas tales como Accuracy, precisión, recall, F1 score y ROC-AUC. Dentro de los algoritmos usados el bosque aleatorio fue el que obtuvo el mejor rendimiento al presentar resultados superiores, con una Accuracy de 0,956679, precisión de 0,956831, recall de 0,956679, F1 score de 0,95666 y un AUC de 0,988. Además, se llevó a cabo un análisis de la importancia de las variables mediante el enfoque de este mismo algoritmo, con el propósito de identificar los factores de riesgo materno más significativos para la predicción del BPN. Los resultados más significativos fueron que el peso de la madre ostentaba la mayor importancia en la predicción de esta condición. Entre otras variables cruciales se encontraban la altura de la madre, la edad materna y el número de visitas prenatales durante el embarazo.

El análisis de las anteriores investigaciones sobre la BPN y MI, aportó diferentes perspectivas para la implementación de los modelos de aprendizaje automático seleccionados en la presente investigación. Resultados como la eficacia de los bosques aleatorios en la clasificación de BPN, en [44], fueron un referente clave para la selección de algoritmos y variables sociodemográficas para la caracterización de factores maternos. Así mismo, la incorporación de SVM lineales, como se evidencia en la investigación [45], demostró la capacidad de este algoritmo para realizar predicciones eficaces, contribuyendo a la elección de modelos de clasificación en el contexto específico del BPNT en el Valle del Cauca.

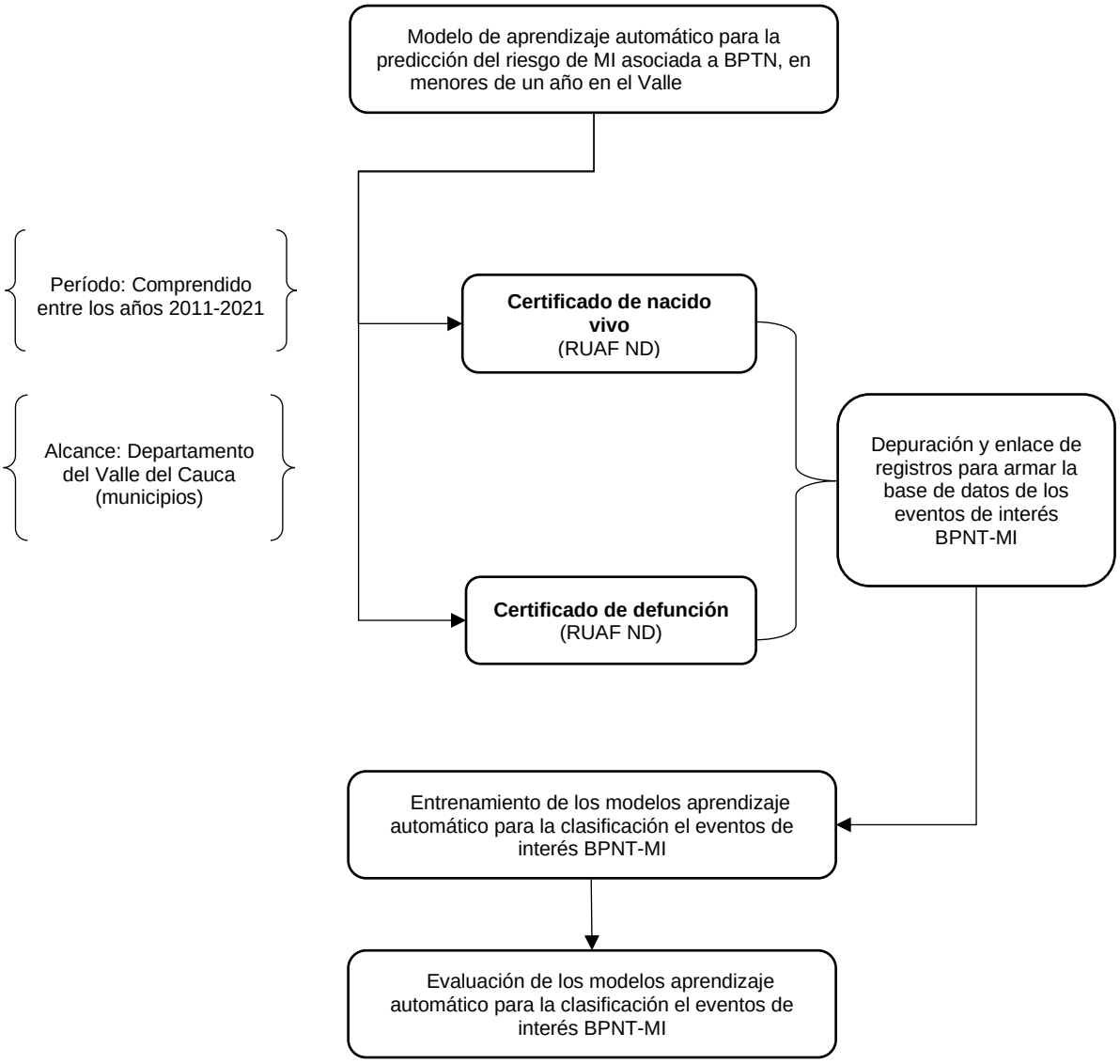
La consideración de desafíos fundamentales, como el desequilibrio en las clases y la elección de variables, indicadas en [47], proporcionó una mejor comprensión en la definición de los modelos más apropiados para enfrentar las complejidades asociadas a la predicción del BPN y la mortalidad infantil. En síntesis, la comparación revela la diversidad de enfoques, destacando variables específicas, evaluación de algoritmos y abordaje de desafíos clave como aspectos cruciales para el desarrollo de un modelo efectivo en el Valle del Cauca.

Lo anterior, destaca la diversidad de enfoques y metodologías de la ciencia de datos aplicadas en investigaciones relacionadas con estos dos fenómenos, que sirvieron de base para el diseño del modelo de aprendizaje automático en el Valle del Cauca.

4. ENTENDIMIENTO Y PREPARACIÓN DE LOS DATOS

Para la elaboración del modelo de aprendizaje automático para la predicción del riesgo de MI asociada al BPTN, en menores de un año en el Valle del Cauca, se surtieron las etapas claves que se indican en la siguiente ilustración.

Ilustración 5. Operacionalización de la investigación



Fuente: Elaboración propia

Como se observa, este ejercicio incluyó el proceso de solicitud de los datos de nacimientos y defunciones en el Valle del Cauca para un periodo de investigación

comprendido entre el 2011 y 2021; posteriormente se llevó a cabo la depuración de los datos, se realizó el enlace de bases de datos, hasta obtener un conjunto de datos óptimo y se seleccionaron las variables a incluir en el modelo; en el presente capítulo se describen los resultados de estas primeras etapas.

Finalmente, se realizó el entrenamiento y evaluación de modelos de aprendizaje automático para clasificar los eventos de salud de interés, enfocándose en la precisión de las predicciones del riesgo de MI asociada al BPNT en el Valle del Cauca. Estas actividades se describen con detalle en el capítulo 5.

4.1. Fuentes de datos disponibles

La elaboración del modelo de aprendizaje automático para la predicción del riesgo de MI asociada al BPTN, en menores de un año en el Valle del Cauca, fue necesario recolectar los datos de los nacidos vivos y las defunciones que evidencian estos hechos vitales en el territorio. Estos datos son acopiados de forma permanente por el Ministerio de Salud y de la Protección Social, mediante el aplicativo Registro Único de Afiliados Nacimientos y Defunciones RUAF-ND, con el propósito de recopilar la información de los nacimientos y defunciones ocurridos en el todo el territorio nacional [52]. Estos mismos datos son utilizados por el Departamento Administrativo Nacional de Estadística DANE para la publicación de los informes de Estadísticas Vitales.

La funcionalidad del aplicativo permite certificar el nacido vivo, cuando este tiene autonomía para respirar, presenta latidos del corazón, pulsaciones, movimientos musculares involuntarios, entre otros que determinan esta condición. Este certificado es necesario para la inscripción del nacimiento en los lugares designados para el registro civil. Certificar la defunción, implica la evidencia de la desaparición de cualquier signo de vida, posterior al nacimiento; este documento es requerido para la elaboración de la licencia de inhumación y para la inscripción del hecho en el registro civil del individuo por la entidad competente.

Para acceder a estas bases de datos consolidadas en el territorio, fue necesario acudir a la Secretaria de Salud Departamental del Valle del Cauca, como también a la Secretaria de Salud Pública Distrital de Santiago de Cali, con el fin de lograr la completitud de los datos, dada la competencia gubernamental de cada entidad pública; en ambos casos, las entidades aportaron las bases de datos relativas a los registros de Nacimientos y Defunciones durante el periodo comprendido entre 2011 y 2021, de manera que este proyecto de investigación se ajustó a dicho período.

A través del certificado de nacido vivo y el de defunción se lograron extraer las variables que tienen relación directa con el objeto de estudio de la presente investigación: BPNT y MI. De esta manera, con estas variables se pudo caracterizar el comportamiento del BPNT y la MI en el tiempo, fue posible identificar el lugar de ocurrencia del hecho vital y el lugar de residencia de la madre en este caso, los municipios del Valle del Cauca.

Si bien ambas bases de datos suministradas, se derivan del mismo aplicativo RUAF-ND, se detectaron diferencias en la estructura de las tablas de los datos existente en cada una de ellas; para superar este obstáculo, se realizó un ejercicio exhaustivo de contraste y validación de cada una de las variables, siguiendo como derrotero la estructura de la base de datos entregada por la Secretaría de Salud de Cali, encontrándose está mucho mejor organizada de acuerdo con la guía documental de los formularios de certificados de nacimientos y defunciones del DANE.

Finamente, cabe anotar que para la obtención de las bases de datos, especialmente de la Secretaria de Salud Departamental del Valle del Cauca, se exigió dos requisitos: 1) el Formato de Confidencialidad y Manejo de datos de la SSD y 2) el documento Aval Ético emitido por el Comité de Ética en Investigación – CEEI de la Pontificia Universidad Javeriana No. 007-2023.

4.2 Descripción de los datos

Los cuatro conjuntos de datos recibidos por parte de la Secretaria de Salud Departamental del Valle del Cauca y de la Secretaria de Salud Pública Distrital de Santiago de Cali se listan a continuación.

Tabla 6. Fuentes y conjuntos de datos de nacimientos y defunciones 2011- 2021

Fuente nivel departamental y distrital	Registros (filas)	Variables (columnas)
Nacimientos en el Valle del Cauca	265.838	44
Defunciones en el Valle del cauca	144.067	73
Nacimientos Santiago de Cali	402.278	51
Defunciones Santiago de Cali	231.643	98

Fuente: Elaboración propia

Al respecto es importante señalar que comparativamente, entre las bases de datos correspondientes a nacimientos, se presentaron 51 variables (columnas) en la de Santiago de Cali, superior a las 44 detectadas en la base de datos del Valle del Cauca;

con respecto a las defunciones, se presentaron 98 variables (columnas) en la de Santiago de Cali, superior a las 73 detectadas en la base de datos del Valle del Cauca.

En la base de datos de nacimientos del Valle del Cauca de 44 variables, se encontraron sólo 3 variables con un porcentaje datos faltantes así: 99,1% para país de nacimiento de la madre; 25,5% para Fecha de nacimiento de la madre y 22,4% para el barrio de residencia de la madre. 41 variables, presentaron un porcentaje inferior al 22%. En la base de datos de defunciones del departamento con 73 variables, la situación fue la siguiente: 34 variables con un porcentaje superior al 30% de datos faltantes y las demás 39 variables presentaron un porcentaje inferior a 30% de datos faltantes.

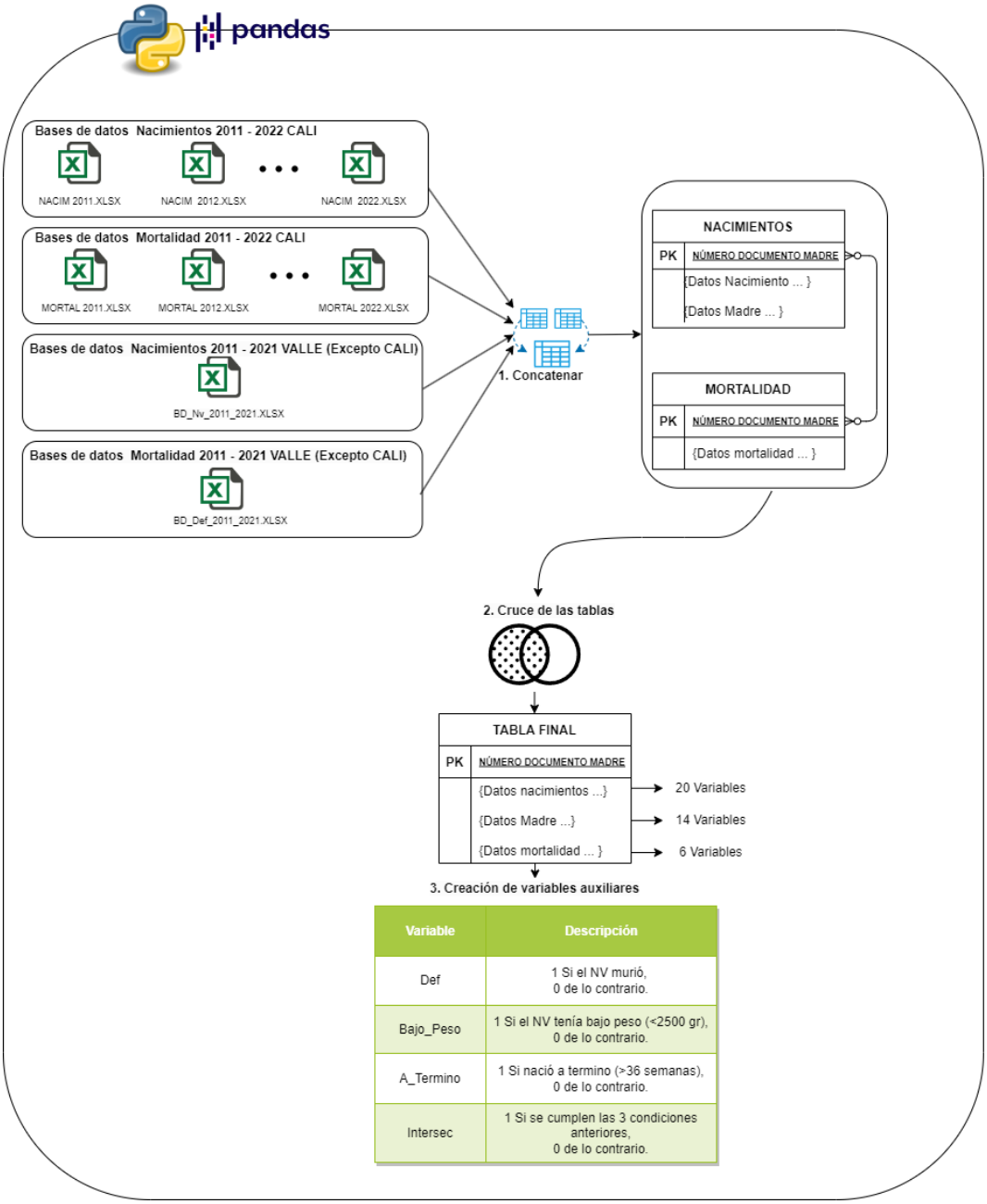
Si bien en las bases de datos de nacimientos y defunciones de Santiago de Cali, se detectó un número mayor de variables respectivamente, se encontró que para la base de datos de nacimientos con 51 variables, 42 de estas presentaron un porcentaje de datos faltantes menor al 20%, y las 9 restantes presentaron un porcentaje de datos faltantes superior al 2 %. En la base de datos de defunciones con 98 variables, se encontró que 46 de estas presentaron un porcentaje de datos faltantes menor al 30%, y 52 restantes resultaron con un porcentaje de datos faltantes superior al 30%.

Además de lo anterior, se detectó la digitación incorrecta en algunos campos de las bases de datos; por lo anterior, se debió realizar un proceso de depuración de los datos con el fin de obtener las bases de datos optimas con las variables requeridas y con la completitud de los datos que permitieran continuar con la etapa de modelación de conjuntos de datos.

4.3. Depuración de los datos

Todo el proceso de depuración de los datos se llevó a cabo en Python mediante la librería Pandas. En la ilustración 5 se relaciona un diagrama de flujo que resume las actividades llevadas a cabo.

Ilustración 6. Diagrama de flujo – proceso de depuración de datos



Fuente: Elaboración propia.

Acorde con la ilustración anterior, a continuación se describe el proceso de depuración de los datos realizados inicialmente a las bases de datos de nacimientos, luego a las base de datos de defunciones hasta la cruce de bases de datos y la obtención de un solo conjunto de datos de nacimiento y defunciones en el Valle del Cauca. Cabe anotar que los datos suministrados por las entidades públicas de salud, comprenden estos hechos vitales ocurridos entre el año 2011 al 2021.

4.3.1. Depuración de los datos de Nacimientos 2011 - 2021

En esta etapa, se llevó a cabo la carga inicial y manipulación de datos relacionados con los nacimientos ocurridos la ciudad de Cali y en el departamento del Valle del Cauca.

Primero, se preparó una lista llamada `dfs` destinada a almacenar los dataframe o conjuntos de datos de Excel que se obtendrían; a través de un bucle que iteraba sobre los archivos en el directorio designado (`directorio`), se cargaron solo aquellos archivos de la Secretaria de Salud de Cali cuyos nombres comenzaban con "NACIM" y tenían extensiones xls o xlsx. Cada dataframe resultante se enriqueció con una nueva columna, 'anyo', indicando el año correspondiente. Estos dataframes individuales se visualizaron para seguimiento.

Después de cargar todos los archivos, se visualizaron las columnas de cada dataframe en un nuevo dataframe, proporcionando así una visión general de la estructura de los datos. Posteriormente, se eliminó una columna específica, 'Fecha Última Modificación', de todos los dataframes en la lista. Finalmente, todos los dataframes se concatenaron en uno solo denominado `df_final`. Este dataframe consolidado representó los datos de nacimientos de todos los años de Santiago de Cali, listos para su análisis.

Adicionalmente, se cargó la base de datos de nacimientos del Valle del Cauca `nac_valle`; este dataframe contenía información adicional o específica para la región del Valle del Cauca.

Después de la carga y consolidación de los datos de nacimientos en la ciudad de Cali, se procedió a una etapa crucial de filtrado; se seleccionaron únicamente los nacimientos que pertenecieran al departamento del Valle del Cauca para enfocarse en la región de interés y desechar datos que no eran pertinentes.

En el caso de Cali, se evaluó la condición de residencia en el dataframe consolidado df_final`, contando el número de registros donde el 'Departamento Residencia' era igual a 'VALLE DEL CAUCA'. De este modo se identificaron 377,416 registros que cumplían con esta condición, y 24,862 registros que no.

Finalmente, se creó un filtro booleano, `f_valle_dptres`, basado en la condición de residencia en el departamento del Valle del Cauca en el dataframe consolidado. Este filtro se utilizó para seleccionar solo los registros relevantes y se aplicó mediante la función `copy()` para evitar modificar la base de datos original. Con esta acción, el `df_final` contenía exclusivamente los datos de nacimientos con residencia en el Valle del Cauca.

Después de la carga y filtrado de los datos de nacimientos para Cali y el Valle del Cauca, se procedió a aplicar filtros específicos para llegar a la población objetivo. Este paso se centró en la caracterización de nacimientos de interés mediante la evaluación de las medidas iniciales de 'Tiempo Gestación' (>= 37 semanas) y 'Bajo Peso al Nacer a Término' (< 2500) en ambas regiones.

Mediante este filtro, se obtuvo que para el conjunto de datos de Cali, coincidieron con los criterios 11.776 de 365.640 (3,12%) registros, mientras que para Valle del Cauca, resultaron acordes 7.721 de 258.117 (2.05%).

Así, se crearon estos subconjuntos `nac_cali` y `nac_valle`, que contenían la población objetivo de nacimientos de interés en Cali y el Valle del Cauca.

Tabla 7. Dataframe de Nacimientos de Cali y Valle del Cauca

Dataframe	Registros (filas)	Variables (columnas)
nac_cali	11776	51
nac_valle	7721	45

Fuente: Elaboración propia

Como se puede observar, la extensión de variables en cada conjunto es distinta por lo que cual se realizó una recategorización de las variables y la selección de Columnas o variables de interés. Se recategorizaron las variables del Valle del Cauca utilizando un archivo de mapeo (`df_mapeo`) que relacionaba números y nombres, según el formulario de nacimientos RUAF. Además, se seleccionaron parcialmente las variables de interés, manteniendo los nombres de las columnas del conjunto de datos de Cali. Esto permitió mantener la consistencia entre los conjuntos de datos de Cali y el Valle del Cauca, facilitando su posterior unión. Se verificó que las columnas seleccionadas fueran consistentes entre ambas regiones.

Hecho esto, se realizaron los siguientes cambios en los dataframes.

Tabla 8. Ajuste de variables Dataframe Nacimientos de Cali y Valle del Cauca

Dataframe	Registros (filas)	Variables (columnas)
nac_cali	11776	33
nac_valle	7721	45

Fuente: Elaboración propia

En el siguiente proceso de depuración, se llevó a cabo una comparación mediante las llaves de integración (Número de identificación de la madre y fecha de nacimiento) para identificar y eliminar registros duplicados en los conjuntos de datos.

Mediante esta operación, se eliminaron registros con número de identificación nulo o faltante y se contaron los valores resultantes. Para Cali, 51 registros fueron eliminados y para el Valle del Cauca, 39 registros.

De otra parte, se llevó a cabo la validación de la integridad temporal del conjunto de datos, garantizando la coherencia de las fechas de nacimientos, la detección y eliminación de posibles datos redundantes, así como la normalización del formato de fecha para asegurar su uniformidad entre ambos conjuntos de datos. No obstante, los resultados de esta evaluación demostraron que todas las fechas fueron convertidas de manera precisa y no se identificaron valores nulos durante dicho proceso de verificación.

Se realizó una comparación entre los conjuntos de datos de Cali y Valle para encontrar duplicados basados en el número de identificación de la madre y la fecha de nacimiento, para ello se convierten los números de identificación a un formato común antes de la comparación. Esta prueba dio como resultado que no se encontraron nacimientos duplicados.

Finalmente, se unieron los conjuntos de datos de Cali y Valle en un conjunto de datos combinado y limpio solo utilizando `pd.concat`. El conjunto de datos de nacimientos del Valle del Cauca (incluido Santiago de Cali) se conformó de 19.407 registros y 32 variables.

4.3.2. Depuración de los datos de Defunciones 2011 - 2021

Se cargaron los conjunto de datos de defunciones de Cali y Valle del Cauca y fueron almacenados en la lista dfs2. Se procedió a concatenar los Dataframes de defunciones y seleccionar solo las columnas relevantes para el análisis; se seleccionaron las columnas

clave relacionadas con la defunción, como la fecha, el sexo, la edad, la causa probable de la muerte, entre otras.

Tabla 9. Dataframe de Defunciones de Cali y Valle del Cauca con ajuste de variables

Dataframe	Registros (filas)	Variables (columnas)
def_cali	231.643	12
def_valle	144.067	12

Fuente: Elaboración propia

En este caso, para llegar a la población objetivo, se realizó inicialmente un filtro de muertes infantiles, para identificar y seleccionar muertes infantiles menores de un año. Se utilizó la variable 'Tipo Edad Fallecido' para realizar el primer filtro. Se aplicó el filtro y se verificó la calidad del mismo para asegurarse de que no se estuvieran excluyendo erróneamente registros de menores a un año. Este análisis confirmó que las defunciones de menores a un año se identificaban correctamente en los conjuntos de datos.

Como resultado de este ejercicio, se obtuvo que las defunciones de menores a un año en Cali fueron de 40.238 registros y para el Valle del Cauca, 2.948. Cabe anotar que es posible que en el Valle del Cauca exista un subregistro debido a que la comparación con Cali, aunque debería ser menor, es extremadamente diferente.

Por otra parte, al realizar este filtro de los nacidos con bajo peso al nacer a término BPNT, el conjunto de datos se redujo para Cali, a 457 registros y para el Valle del Cauca a 206 registros.

Posteriormente, se compararon las cédulas duplicadas entre Cali y el Valle para asegurar la integridad de los datos. Esto proporcionó información sobre las cédulas duplicadas encontradas entre Cali y se procedió a eliminarlas en dicho conjunto de datos. También se identificaron cédulas nulas o anómalas cuyos registros también fueron eliminados.

Tabla 10. Dataframe de Defunciones de Cali y Valle del Cauca con ajuste de cédulas

Dataframe	Registros (filas)	Variables (columnas)
def_cali	352	13
def_valle	196	13

Fuente: Elaboración propia

Las 12 variables seleccionadas de la fuente de mortalidad se utilizaron como base para el proceso de cruce entre diferentes fuentes de datos, como el municipio de residencia, la fecha de fallecimiento, la causa de la muerte y el documento de la madre del fallecido. Además, estas variables se emplearon para construir la variable de respuesta 'def' (que incluye peso, talla y tiempo de gestación). Es importante destacar que estas variables no fueron incorporadas en el modelo, ya que la predicción se realiza antes del evento de interés.

Finalmente, se llevó a cabo la unión de los conjuntos de defunciones de Cali y el Valle en un solo conjunto de datos final (df_final_mort) compuesto por 548 registros y 13 variables.

4.3.3. Cruce de Nacimientos vs Defunciones 2011 - 2021

Se cargaron los datos finales obtenidos anteriormente para nacimientos y defunciones en Cali y el Valle del Cauca. Los conjuntos de datos finales se describen en la siguiente tabla.

Tabla 11. Dataframe finales de Nacimientos y Defunciones para el Valle del Cauca

Dataframe	Registros (filas)	Variables (columnas)
df_valle_nac	19.407	32
df_valle_mort	548	13

Fuente: Elaboración propia

Se ajustaron los tipos de datos de las columnas relacionadas con fechas y cédulas para facilitar el análisis; se compararon las cédulas entre los conjuntos de datos de nacimientos y defunciones para identificar aquellas que coincidían. Se verificó la cantidad de cédulas compartidas entre ambos conjuntos: para el conjunto de datos de Nacimientos coincidieron 564 registros y para Defunciones, 238 registros.

No obstante lo anterior, se realizó un análisis de la variable 'Multiplicidad Embarazo' en el conjunto de nacimientos, indicando que se registraron 17.794 embarazos simples, 1.612 dobles, y 1 triple; lo anterior para evaluar cada hecho vital de manera independiente en la siguiente tarea. Se identificaron en el conjunto de nacimientos las filas donde las cédulas coincidían con las de defunciones. Se creó una variable indicadora llamada 'apareció', que tomó el valor 1 si la cédula de la madre estaba presente en ambos conjuntos y 0 en caso contrario. Bajo este criterio, solo 279 registros se encontraron en ambos conjuntos mientras que 19.128 no.

Finalmente, se renombró la columna 'apareció' como 'def'. Se guardó el conjunto de datos resultante en un archivo CSV llamado 'df_previo_modelo.csv'.

4.3.4. Selección de variables

La selección de variables se basó en el criterio teórico y empírico reportado en la literatura científica dado que la mortalidad infantil es un problema bastante estudiado en epidemiología, algunos de estos estudios fueron ya reportados en el estado del arte. Entre los artículos se destaca el del grupo de investigación de la Facultad de Matemáticas y Ciencias Naturales de la Universidad de Brawijaya en Indonesia que llevó a cabo un estudio utilizando datos de la Encuesta Demografía y Salud de Indonesia (IDHS) del 2012 [44]; otro estudio del 2022 [46], en donde se usaron datos de 7.472 registros de nacimientos de la Red Neonatal Coreana; un artículo científico en los Emiratos Árabes [47] que tuvo como objetivo evaluar el rendimiento de 30 algoritmos de Aprendizaje Automático tanto para la estimación del peso corporal infantil como para la clasificación del BPN; Otro artículo publicado en el 2021 [48], cuyo objetivo fue determinar los factores relacionados con la mortalidad infantil, utilizando algoritmos de minería de datos.

Además de la variable indicadora de defunción se crearon otras 3 variables auxiliares con el fin de vincular en cada nacimiento si se presentó el evento de interés MI – BPNT.

- Def: Permite identificar si un recién nacido murió (1) o no (0).
- Bajo_Peso: Permite identificar si un bebé nació con bajo peso (1: Peso<2500 gr) o no (0).
- A_Termino: Permite identificar si un nacimiento fue prematuro (0) o a término (1: Tiempo de gestación > 36 semanas).
- Intersec: Es la variable de interés que marca los nacimientos que cumplen las 3 condiciones anteriores.

El total de variables del conjunto de datos fueron las siguientes:

Tabla 12. Variables del conjunto de datos

Nombre	Tipo	Escala	Descripción	Categorías/Rango
Municipio Nacimiento	Cualitativa	Nominal	Municipio donde ocurrió el nacimiento	42 categorías siendo 'CALI' el más frecuente.
Área Nacimiento	Cualitativa	Nominal	Área (urbana o rural) donde ocurrió el nacimiento	CABECERA MUNICIPAL RURAL DISPERSO CENTRO POBLADO
Sitio Parto	Cualitativa	Nominal	Lugar específico donde ocurrió el parto	INSTITUTO DE SALUD DOMICILIO OTRO SITIO
IPS	Cualitativa	Nominal	Institución Prestadora de Servicios de Salud que atendió el parto	189 categorías, siendo el HUV el más frecuente.

Sexo	Cualitativa	Nominal	Sexo del recién nacido (masculino o femenino)	FEMENINO MASCULINO
Peso	Cuantitativa	Continua	Peso del recién nacido al nacer en kilogramos	[300 – 2.499]
Talla	Cuantitativa	Continua	Talla del recién nacido al nacer en centímetros	[21 – 99]
Fecha Nacimiento	Fecha	N/A	Fecha en la que ocurrió el nacimiento	[2011-01-01 – 2022-12-31]
Parto Atendido Por	Cualitativa	Nominal	Profesional de la salud o entidad que atendió el parto	MÉDICO OTRA PERSONA PARTERA PROMOTOR DE SALUD ENFERMERO AUXILIAR DE ENFERMERIA
Tiempo Gestación	Cuantitativa	Continua	Duración del embarazo en semanas	[37 – 99]
Número Consultas Prenatales	Cuantitativa	Discreta	Número de consultas médicas que tuvo la madre durante el embarazo	[0 – 99]
Tipo Parto	Cualitativa	Nominal	Tipo de parto (normal, cesárea, etc.)	ESPONTANEO CESÁREA INSTRUMENTADO IGNORADO
Multiplicidad Embarazo	Cualitativa	Nominal	Número de fetos en el embarazo (gemelos, mellizos, etc.)	SIMPLE DOBLE TRIPLE
APGAR1	Cuantitativa	Discreta	Puntuación de Apgar del recién nacido al minuto de vida	[0-10]
APGAR5	Cuantitativa	Discreta	Puntuación de Apgar del recién nacido a los 5 minutos de vida	[0-10]
Grupo Sanguineo	Cualitativa	Nominal	Grupo sanguíneo del recién nacido	‘O’ ‘A’ ‘B’ ‘AB’ ‘SIN INFO’
Factor RH	Cualitativa	Nominal	Factor RH del recién nacido	‘POSITIVO’ ‘NEGATIVO’
Pertenencia Étnica	Cualitativa	Nominal	Pertenencia étnica de la madre	‘NEGRO, MULATO…’ ‘INDIGENA’ ‘PALENQUERO’ ‘ROM’ ‘RAIZAL’
Pueblo Indígena	Cualitativa	Nominal	Nombre del pueblo indígena	Nombre del pueblo indígena.
Tipo Documento Madre	Cualitativa	Nominal	Tipo de documento de identidad de la madre	‘TARJETA DE ID’ ‘CÉDULA DE CIUDADANÍA’ ‘CÉDULA DE EXRANJERÍA’ ‘PASAPORTE’ ‘SIN INFORMACIÓN’
Numero Documento Madre	Cualitativa	Nominal	Número de documento de identidad de la madre	N/A
Edad Madre	Cuantitativa	Continua	Edad de la madre en el momento del parto	[11 – 54]
Estado Conyugal Madre	Cualitativa	Nominal	Estado conyugal de la madre (soltera, casada, divorciada, etc.)	‘NO CASADO(A) DOS AÑOS O MÁS CON PAREJA’ ‘NO CASADO(A) Y MENOS DE DOS AÑOS CON PAREJA’ ‘SOLTERO(A)’ ‘CASADO(A)’ ‘SIN INFORMACIÓN’ ‘ESTABA SEPARADO(A)’ ‘ESTABA VIUDO(A)’
Último Año Estudios Madre	Cualitativa	Ordinal	Último año de estudios de la madre	MEDIA ACADÉMICA BÁSICA SECUNDARIA BÁSICA PRIMARIA PROFESIONAL TÉCNICA PROFESIONAL TECNOLÓGICA

				SIN INFORMACIÓN MEDIA TÉCNICA NINGUNO ESPECIALIZACIÓN PREESCOLAR MAESTRÍA NORMALISTA DOCTORADO
Municipio Residencia	Cualitativa	Nominal	Municipio de residencia de la madre	53 categorías, siendo 'CALI' el más frecuente.
Área Residencia	Cualitativa	Nominal	Área de residencia de la madre (urbana o rural)	'CABECERA MUNICIPAL' 'RURAL DISPERSO' 'CENTRO POBLADO'
Número Hijos Nacidos Vivos	Cuantitativa	Discreta	Número total de hijos nacidos vivos de la madre hasta el momento del parto	[1 – 13]
Fecha Nacimiento Anterior Hijo	Fecha	N/A	Fecha de nacimiento del hijo anterior (en caso de haberlo)	N/A
Numero Embarazos	Cuantitativa	Discreta	Número total de embarazos de la madre hasta el momento del parto	[1 – 18]
Régimen Seguridad Social	Cualitativa	Nominal	Tipo de régimen de seguridad social al que está afiliada la madre	'ASEGURADA' 'SUBSIDIADA' 'ESPECIAL'
EPS	Cualitativa	Nominal	Entidad Promotora de Salud a la que está afiliada la madre (en el sistema de salud colombiano)	139 categorías siendo EMSANAR ESS, la más frecuente.
Edad Padre	Cuantitativa	Continua	Edad del padre en el momento del parto	[14 – 76]
Último Año Estudios Padre	Cualitativa	Ordinal	Último año de estudios del padre	MEDIA ACADÉMICA BÁSICA SECUNDARIA BÁSICA PRIMARIA PROFESIONAL TÉCNICA PROFESIONAL TECNOLÓGICA SIN INFORMACIÓN MEDIA TÉCNICA NINGUNO ESPECIALIZACIÓN PREESCOLAR MAESTRÍA NORMALISTA DOCTORADO
Año	Cualitativa	Ordinal	Año en el que ocurrió el evento relacionado con el nacimiento	2011- 2021
def	Cualitativa	Nominal	Indicación de si el recién nacido falleció (sí o no)	SI : 1 NO : 1
Tipo_edad_def	Cualitativa	Ordinal	Unidad de tiempo que indica la edad del fallecido	MENOR DE 1 MES (EN DÍAS) MENOR DE 1 AÑO (EN MESES) MENOR DE 1 DÍA (EN HORAS) MENOR DE 1 HORA (EN MINUTOS)
Edad_fallecido	Cuantitativa	Continua	Edad del fallecido de acuerdo con la unidad de tiempo en la que se registró el acontecimiento	[1 – 40]
Tipo Identificación_def	Cualitativa	Nominal	Tipo de documento de identidad de la madre del fallecido	'Tarjeta de Identidad' 'Registro Civil' 'Cédula de ciudadanía' 'Cédula de extranjería' 'Pasaporte' 'Sin info'
Número de Identificación_def	Cualitativa	Nominal	Número de documento de identidad de la madre del fallecido	N/A
peso_def	Cuantitativa	Continua	Peso del recién nacido al momento del fallecimiento, en caso de que haya ocurrido	[2 – 2.495]
t_gest_def	Cuantitativa	Continua	Duración del embarazo en semanas al momento del fallecimiento, en caso de que haya ocurrido	[37– 99]

sexo_def	Cualitativa	Nominal	Sexo del recién nacido al momento del fallecimiento, en caso de que haya ocurrido	'FEMENINO' 'MASCULINO'
manera_def	Cualitativa	Nominal	Manera en que ocurrió el fallecimiento del recién nacido, en caso de que haya ocurrido	'VIOLENTO' 'NATURAL'
fecha_def	Fecha	N/A	Fecha en la que ocurrió el fallecimiento del recién nacido, en caso de que haya ocurrido	Fecha
Municipio_residencia_def	Cualitativa	Nominal	Municipio de residencia del fallecido.	93 categorías
Municipio_ocurrencia_def	Cualitativa	Nominal	Municipio de residencia donde ocurre el hecho.	14 categorías

Descripción del procesamiento de las variables categóricas:

Recategorización:

- Se han simplificado las categorías de variables como último año de escolaridad, área de nacimiento, pertenencia étnica y estado conyugal.
- La variable edad ha sido agrupada en rangos de 5 años en lugar de ser tratada como variable numérica.
- La variable atención al parto ahora tiene solo 2 categorías: Médico u Otra persona, en lugar de las 5 anteriores.
- Se ha reducido la variable municipio de residencia a 3 categorías: Cali, Buenaventura y Otros.

• Codificación de variables ordinales:

- Dada la naturaleza ordenada de las variables de Multiplicidad del embarazo y el periodo intergenésico, se han recodificado a números del 0 al máximo número de categorías presentes.

• Dumificación:

- Después de la recodificación de las variables categóricas, se lleva a cabo la dumificación, proceso mediante el cual cada columna que representa una variable se divide en tantas columnas como categorías existan. Cada valor de estas nuevas columnas se presenta como 0 o 1, indicando la ausencia o presencia de la categoría en ese registro respectivamente.

Finalmente, las variables seleccionadas fueron las siguientes:

1. Municipio Nacimiento
2. Área Nacimiento
3. Sitio Parto
4. IPS
5. Sexo
6. Peso

7. Talla
8. Fecha Nacimiento
9. Parto Atendido Por
10. Tiempo Gestación
11. Número Consultas Prenatales
12. Tipo Parto
13. Multiplicidad Embarazo
14. APGAR1
15. APGAR5
16. Grupo Sanguíneo
17. Factor RH
18. Pertenencia Étnica
19. Pueblo Indígena
20. Tipo Documento Madre
21. Numero Documento Madre
22. Edad Madre
23. Estado Conyugal Madre
24. Último Año Estudios Madre
25. Municipio Residencia
26. Área Residencia
27. Número Hijos Nacidos Vivos
28. Fecha Nacimiento Anterior Hijo
29. Numero Embarazos
30. Régimen Seguridad Social
31. EPS
32. Año

En conclusión, una vez depurados los datos contenidos en los 4 conjuntos de datos, se obtuvo un conjunto de datos final de nacimientos con BPNT y defunciones en el Valle del Cauca (incluido Cali), con 19.128 nacimientos BPNT y 279 defunciones en menores de un año; esto, mediante la creación de una variable indicadora ('def') que representó la ocurrencia de defunciones relacionadas con nacimientos en función de las cédulas de las madres. Esto demuestra que para este conjunto de datos o dataframe existe un desbalanceo de los datos que fue tratado mediante las técnicas correspondientes y que se detalla en el capítulo 5.

4.4. Imputación de datos a partir del conjunto de datos de nacimientos con bajo peso al nacer a término y mortalidad infantil

Para este proceso, se usó Python y la librería Pandas, se identificaron y cuantificaron las variables con datos atípicos; en este contexto, se consideraron como atípicos los valores que superaban el umbral de 90. Como resultado de esta operación se encontraron valores atípicos en las siguientes variables y cantidades, como se muestra en la siguiente tabla.

Tabla 13. Variables y cantidad de datos faltantes en el Dataframe final

Variable	Cantidad de atípicos
Talla	1
Tiempo Gestación	58
No. de Consultas prenatales.	38
APGAR1	190
APGAR5	190
No. de Embarazos	1

Fuente: Elaboración propia

Al respecto, se priorizaron los registros de nacidos vivos NV que murieron y que tenían en alguna característica, entre los datos atípicos identificados. Se encontró que 4 de los 7 registros (NV) tenían datos faltantes ya que habían sido atendidos por parteras, y 2 por otro tipo de persona diferente del médico. Los primeros 4 fueron NV con pertenencia étnica indígena y los datos faltantes correspondían a las mediciones de APGAR1 y APGAR5. Otro de los registros restante, pertenece a una nacida viva con pertenencia étnica negra y el dato que faltaba era el número de Consultas prenatales.

Para subsanar lo anterior como propuesta de imputación en los datos numéricos, se generó un data set que guarda la mediana de la población que murió y no murió; se calculó su respectiva mediana en la variable donde se encontraron datos faltantes, con base en los datos disponibles. Sin embargo para los registros con datos faltantes de NV no muertos, no se generó una imputación en el APGAR ya que son datos sensibles y subjetivos al momento en que se mide y por lo tanto se descartaron. Los cálculos resultantes se muestran a continuación.

Tabla 14. Valor de la mediana en variables con datos faltantes según la defunción

Variable	Defunción	Mediana
Tiempo Gestación	No	38.0
Tiempo Gestación	Si	38.0
Talla	No	47.0
Talla	Si	46.0
Consultas Prenatales	No	7.0
Consultas Prenatales	Si	6.0
APGAR1	Si	7.0
APGAR5	Si	8.0

Variable	Defunción	Mediana
Numero Embarazos	No	1.0

Fuente: Elaboración propia

4.5. Cálculo del periodo intergenésico

El periodo intergenésico se define como el tiempo transcurrido entre el parto del anterior producto y el momento de concepción del actual; se considera que un periodo intergenésico adecuado oscila entre 18 y 27 meses, y no superior a 60. Esto resultó importante para la presente investigación por cuanto, dependiendo de la duración de este periodo intergenésico, se convierte en factor desencadenante de eventos adversos tanto para la madre y el neonato, incluida etapa perinatal. [53].

En especial, estos riesgos se asocian mayormente a un Periodo Intergenésico Corto PIC, inferior a los 18 meses, como se muestra en la siguiente tabla.

Tabla 15. Riesgos asociados al PIC inferior a 18 meses

Intervalo del PIC	Riesgos
12 < PIC < 18 Meses	Prematuridad Ruptura Uterina Bajo Peso al Nacer
3 < PIC < 9 Meses	Parto pretérmino Bajo Peso al Nacer Aborto / Óbito Malformaciones neonatales Ruptura uterina Muerte Neonatal

Fuente: Adaptación a partir de [53]

A continuación se presenta los pasos realizados para construir dicho indicador:

i. En primer lugar, se verificó si el número de hijos nacidos vivos es igual a 1 o si es igual a 2, excluyendo la multiplicidad del embarazo es "DOBLE". En estos casos, se consideró que la madre es "primeriza".

ii. Luego, se verificó si la fecha de nacimiento anterior del hijo está perdida (es nula) o si el año de nacimiento anterior es anterior a 1950. En estos casos, se categoriza como "perdido".

iii. Si las fechas de nacimiento son iguales (lo que no debería ocurrir, ya que deberían ser diferentes para calcular el período intergenésico), se considera como "perdido".

iv. Se agregó una condición adicional para verificar si la diferencia en días entre la fecha de nacimiento actual y la fecha de nacimiento anterior es menor a 365.25 días (un año). Si es así, se clasifica como "menos de un año".

v. Si ninguna de las condiciones anteriores se cumplía, se procedió a calcular el período intergenésico. Se restó la duración del embarazo (en semanas) al año de la fecha de nacimiento actual para obtener la fecha de inicio del período intergenésico.

vi. Luego, se calculó el período intergenésico dividiendo la diferencia en días entre la fecha de inicio y la fecha de nacimiento anterior del hijo entre 365.25 días (un año bisiesto).

vii. Finalmente, se categorizó el período intergenésico en tres grupos: "menos de un año" si es menor a 1, "un año a dos años" si está entre 1 y 2, y "más de dos años" para cualquier período mayor a 2.

4.6. Descriptivas del conjunto de datos de nacimientos con bajo peso al nacer a término y mortalidad infantil

A continuación, se presenta el análisis realizado a cada una de las variables cuantitativas, derivadas del conjunto de datos definitivos que abordan los nacimientos con BPNT y la consiguiente MI.

Tabla 16. Medidas de tendencia de las variables cuantitativas.

	Peso	Talla	Tiempo Gestación	Consultas Prenatales	APGAR1	APGAR5	Edad Madre	Hijos Nacidos Vivos	Numero Embarazos
Promedio	2292.28	46.76	37.82	6.35	8.47	9.57	25.19	1.69	1.84
Desv. Sd.	192.78	2.45	0.99	2.5	1.04	0.81	6.75	1.06	1.22
Min	300.0	21.0	37.0	0.0	1.0	1.0	12.0	1.0	1.0
Mediana	2345.0	47.0	38.0	7.0	9.0	10.0	24.0	1.0	1.0
Max.	2499.0	58.0	43.0	20.0	10.0	10.0	54.0	16.0	18.0

Fuente: Elaboración propia

En cuanto al peso, se observó un promedio de 2.292,28 gramos, con una desviación estándar de 192,78 gramos. El peso mínimo registrado es de 300 gramos, mientras que el máximo es de 2.499 gramos.

En cuanto a la talla, el promedio es de 46,76 centímetros, con una desviación estándar de 2,45 centímetros. La mediana de la talla es de 47 centímetros. La talla mínima registrada es de 21 centímetros, mientras que la máxima es de 58 centímetros. Esto indica que la mayoría de los bebés tienen una talla cercana a 47 centímetros.

El tiempo de gestación tiene una mediana de 38 semanas indicando que el 50% de los NV con BPNT nacieron alrededor de la semana 38, con una desviación estándar de 0.99 semanas. El mínimo tiempo de gestación registrado es de 37 semanas, y el máximo es de 43 semanas,

En cuanto al número de consultas prenatales, la mediana fue de 7 consultas, con una desviación estándar de 2,5 consultas. El mínimo número de consultas registrado es de 1, mientras que el máximo es de 20. Esto indica que la cantidad de consultas prenatales varía ampliamente entre los casos.

El puntaje APGAR1 tiene un promedio de 8,47, con una desviación estándar de 1,04. El puntaje mínimo es de 1, y el máximo es de 10, lo que sugiere que la mayoría de los recién nacidos tienen puntajes APGAR1 bastante altos.

El puntaje APGAR5 tiene un promedio de 9,57, con una desviación estándar de 0,81. La mediana del puntaje APGAR5 es de 10. El puntaje mínimo es de 1, y el máximo es de 10. Al igual que con APGAR1, la mayoría de los bebés tienen puntajes APGAR5 altos.

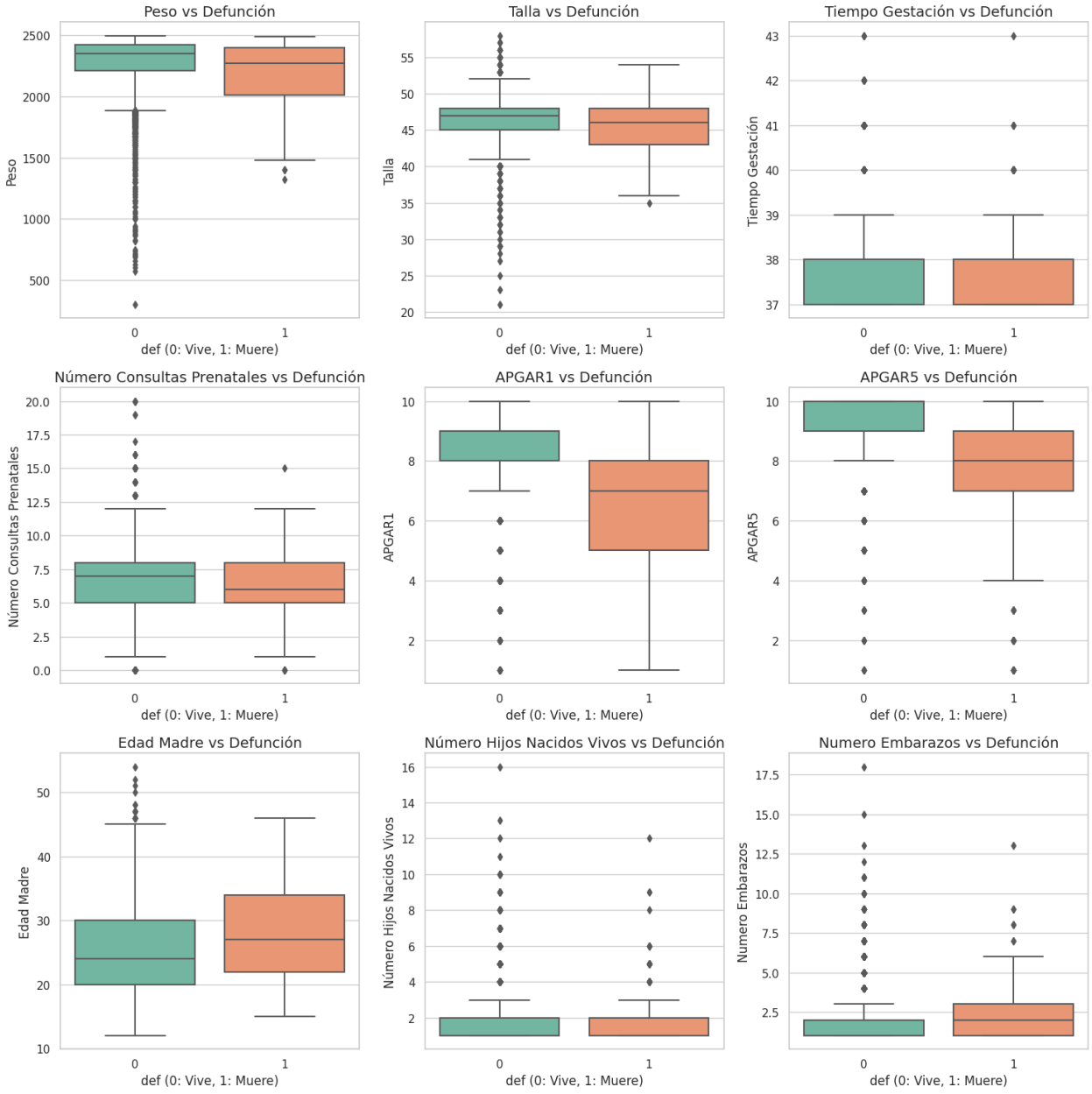
La edad promedio de la madre es de 25,19 años, con una desviación estándar de 6,75 años. La mediana de la edad de la madre es de 24 años. La madre más joven tenía 12 años, mientras que la más mayor tenía 54 años. Esto indica una amplia variación en las edades de las madres.

En cuanto al número de hijos vivos, el promedio es de 1,69 hijos, con una desviación estándar de 1,06. La mediana del número de hijos vivos es de 1. El mínimo número de hijos vivos es 1, y el máximo es de 16. Esto muestra una variabilidad en el número de hijos vivos entre las madres.

Finalmente, el número de embarazos tiene un promedio de 1,84, con una desviación estándar de 1,22. La mediana del número de embarazos es de 1. El mínimo número de embarazos registrado es de 1, mientras que el máximo es de 18. Esto indica que las madres tienen diferentes historias reproductivas, con algunas teniendo múltiples embarazos y otras teniendo solo uno.

Los resultados descritos anteriormente se presentan en la siguiente ilustración.

Gráfica 2. Diagramas BoxPlot - Distribución de valores por variable numérica en función de la defunción



Fuente: Resultados en Phyton desde https://github.com/Estocastic/Hello-world/blob/master/COD_BPN_V2.ipynb

En el diagrama de caja del peso, observamos que la mediana del peso para los individuos que viven (def=0) es superior a la de aquellos que mueren (def=1). Además, la distribución del peso para los que viven parece ser más amplia y con valores atípicos más extremos, indicando una mayor variabilidad en el peso en comparación con los que mueren.

El diagrama de caja de la talla muestra diferencias menos marcadas entre los dos grupos en términos de mediana. Sin embargo, al igual que con el peso, la variabilidad de la talla parece ser mayor en los individuos que viven, con un rango intercuartílico más amplio y varios valores atípicos bajos.

Para el tiempo de gestación, las medianas son muy similares entre los dos grupos. No obstante, el grupo de los que mueren tiene un rango intercuartílico ligeramente menor y menos valores atípicos, lo que sugiere una menor variabilidad en comparación con los que viven.

El número de consultas prenatales muestra una mediana más alta para los que viven en comparación con los que mueren. La distribución para los que viven es más dispersa con un rango intercuartílico más amplio y valores atípicos, lo que implica que tienden a tener un número más variado de consultas prenatales.

En cuanto al APGAR1, hay una diferencia notable entre las medianas de los dos grupos. Los individuos que viven tienen una mediana de APGAR1 significativamente más alta, indicando mejores condiciones de salud al nacer en comparación con los que mueren, que muestran una mediana más baja y un rango intercuartílico más estrecho. Similar al APGAR1, el APGAR5 presenta una mediana más alta y una distribución más amplia para los individuos que viven, lo cual es consistente con una mejor condición de salud al nacer. Los que mueren tienen una mediana más baja y un rango intercuartílico más concentrado.

La edad de la madre muestra medianas similares entre los dos grupos, aunque con un rango intercuartílico ligeramente más amplio para las madres de individuos que viven. Esto podría sugerir una mayor variedad en las edades de las madres cuyos hijos viven.

El número de hijos nacidos vivos tiene medianas comparables para ambos grupos, pero con una distribución más amplia para el grupo de los que viven, indicando una mayor variabilidad en el número de hijos previamente nacidos vivos entre este grupo.

Finalmente, para el número de embarazos, no se observa una gran diferencia en las medianas entre los dos grupos. La distribución es ligeramente más amplia para los que

viven, lo cual es coherente con las observaciones en las otras variables relacionadas con la reproducción.

Ahora, en la comparación entre las dos poblaciones, aquellas con defunción (def=1) muestran algunas diferencias significativas en algunas de las variables clave, con respecto a las que no tienen defunción (def=0). Véase la siguiente tabla.

Tabla 17. Análisis comparativo entre def=1 vs. def=00

Característica	Falleció	Min	Max	Mediana	Desv Est.	P-value
Peso	No	300	2499	2350	191,01	< 0,0001
	Si	1320	2490	2275	266,88	
Talla	No	21	58	47	2,44	< 0,0001
	Si	35	54	46	3,2	
Tiempo Gestación	No	37	43	38	0,99	0,3165
	Si	37	43	38	1,04	
Número Consultas Prenatales	No	0	20	7	2,5	0,2267
	Si	0	15	6	2,46	
APGAR1	No	1	10	9	0,97	< 0,0001
	Si	1	10	7	2,49	
APGAR5	No	1	10	10	0,72	< 0,0001
	Si	1	10	8	2,52	
Edad Madre	No	12	54	24	6,73	< 0,0001
	Si	15	46	27	7,62	
Numero Embarazos	No	1	18	1	1,21	< 0,0001
	Si	1	13	2	1,73	
Número Hijos Nacidos Vivos	No	1	16	1	1,05	< 0,0001
	Si	1	12	2	1,48	

Fuente: Elaboración propia

La característica "Peso" mostró una diferencia notable en la mediana entre los que no fallecen (2350 gramos) con respecto a los que si fallecen (2275 gramos), aunque ambos grupos presentan un rango de pesos similares.

En cuanto a la "Talla", los recién nacidos que fallecen tienden a tener una mediana ligeramente menor (46 cm) en comparación con los que sobrevivieron (47 cm).

El "Tiempo Gestación" es uniforme entre ambos grupos, con una mediana de 38 semanas.

Para el "Número de Consultas Prenatales", hay una mediana ligeramente menor en el grupo de no supervivientes (6 consultas) en comparación con los supervivientes (7 consultas), sugiriendo que una menor cantidad de atención prenatal podría estar relacionada con resultados negativos.

El APGAR a 1 minuto ("APGAR1") y a 5 minutos ("APGAR5") muestra diferencias más pronunciadas. Los supervivientes tienen medianas más altas (APGAR1: 9, APGAR5: 10) en comparación con los fallecidos (APGAR1: 7, APGAR5: 8), indicando que puntuaciones más bajas en estas escalas están asociadas también con un mayor riesgo de defunción.

La "Edad Madre" también difiere entre los grupos, con una mediana mayor en las madres de no supervivientes (27 años) en comparación con las madres de supervivientes (24 años), lo que podría reflejar factores de riesgo asociados con la edad materna.

En la característica "Número Hijos Nacidos Vivos", la mediana para los nacidos vivos con bajo peso que sobreviven es de 1, mientras que para aquellos que no sobreviven es de 2. Esta comparación podría sugerir que un mayor número de hijos nacidos vivos previos podría estar asociado con un aumento en el riesgo de defunción neonatal.

Finalmente, el "Número de Embarazos" es mayor en el grupo de no supervivientes (mediana de 2) que en el grupo de supervivientes (mediana de 1), sugiriendo que embarazos múltiples podrían estar relacionados con un incremento en el riesgo de defunción neonatal.

La desviación estándar (Desv.) en todas las variables sugiere la variabilidad dentro de cada grupo, siendo más alta en variables como el peso y la edad de la madre en el grupo de defunciones. Estas estadísticas proporcionan una visión general de cómo ciertas características clínicas se relacionan con los resultados de supervivencia o defunción en nacidos con bajo peso al nacer.

Finalmente, para las anteriores variables se muestra los resultados de la Prueba U de Mann-Whitney para comparar las medianas de los dos grupos de NV que fallecieron o sobrevivieron. Esta prueba arrojó resultados de forma que, con una significancia del 5% se puede afirmar que existen diferencias significativas entre las medianas del peso, la talla, el APGAR 1 y el APGAR 5, la edad madre, el número de embarazos y número de

hijos nacidos vivos. En contraste, el tiempo de Gestación (p-valor: 0,31) y el número de consultas prenatales (p-valor: 0,22) arrojaron p-valores mayores a la significancia, indicando que no se encontraron diferencias significativas entre sus medianas.

La descripción previa de las variables cuantitativas vinculadas al BPNT y la MI revelan tendencias relacionadas con el problema central que estamos abordando. Destacamos que el peso al nacer, factor fundamental, muestra diferencias notables entre los recién nacidos que sobreviven y los que lamentablemente fallecen. La variabilidad en la talla, tiempo de gestación, número de consultas prenatales y otros indicadores converge para comprender la complejidad de este fenómeno. Las disparidades en las medianas y amplitud de las distribuciones sugieren que aspectos como la atención prenatal y condiciones de salud al nacer desempeñan un papel crucial en la MI asociada al BPNT. Este primer análisis exploratorio sentó las bases para el desarrollo de modelos de aprendizaje automáticos, implementados y detallados en el próximo capítulo de esta investigación. Estos modelos, basados en estas variables, se evaluaron con el propósito de brindar una comprensión más profunda y precisa de los factores que contribuyen a la MI en nacimientos con BPNT.

5. ENTRENAMIENTO DE LOS MODELOS DE APRENDIZAJE AUTOMÁTICO

Continuando con las etapas del desarrollo del proyecto, según el flujo de tareas indicadas en la Ilustración No. 4, en este capítulo se detallan las actividades realizadas para el entrenamiento de los modelos de aprendizaje automático que permitieron predecir el riesgo de MI, asociada al BPNT, en menores de un año en el Valle del Cauca.

De antemano, vale la pena mencionar que se entrenaron siete modelos, Regresión Logística (RL), Support Vectorial Machine (SVM), Árbol de Decisión (DT), Random Forest – (RF), XGBoost – (XGB), K-Nearest Neighbors – (KNN) y el Naive Bayes – (NB). En todo este proceso, la partición de los datos se realizó en una proporción de 75% para entrenamiento y 25% para prueba. Además, indicar que todos los análisis se llevaron a cabo en el lenguaje de procesamiento Python 3, mediante el entorno de programación de Google Colab.

Inicialmente, se realizó el entrenamiento de los siete modelos con los datos arrojados en la etapa de depuración (Capitulo 4), que resultaron desbalanceados ya que el conjunto de datos final se estableció con 19.028 nacimientos y 279 defunciones. Si bien el uso de datos desbalanceados conduce al riesgo de obtener aprendizajes sesgados que perjudican la predicción de la clase minoritaria [54], es importante conocer el comportamiento del modelo bajo esta circunstancia, es decir cómo se puede afectar el

rendimiento del modelo [55] y determinar el tratamiento más conveniente para balancear los datos. Por lo anterior, para evaluar el rendimiento de estos modelos, se utilizaron las siguientes métricas: Exactitud, Precisión, Sensibilidad (Recall) y F1-Score.

Luego, se realizó el balanceo de los datos mediante las técnicas de Sobremuestreo, Submuestreo y SMOTE (Synthetic Minority Over-sampling Technique), se entraron nuevamente los 7 modelos; también, se evaluaron estos modelos bajo las métricas de Exactitud, Precisión, Sensibilidad (Recall) y F1-Score.

Finalmente, se realizó un proceso de optimización a través de hiperparámetros, con el propósito de mejorar el rendimiento de los modelos [56]. Una vez finalizado el ciclo de optimización para todos los modelos, se analizaron y se compararon las métricas de rendimiento logrando evidenciar el mejor rendimiento según la variable de interés.

5.1. Entrenamiento y evaluación de modelos con datos desbalanceados

Una vez realizada la partición del conjunto de datos con el 75% para entrenamiento y el 25% para validación, se entrenaron los 7 modelos con el conjunto de datos correspondiente

Tabla 18. Distribución de la defunción en los datos particionados

	Entrenamiento (75%)	**Prueba (25%)**
Def: 0	14.271	4.757
Def: 1	209	70

Fuente: Elaboración propia

Para la evaluación de los modelos se priorizaron las métricas de Exactitud, Precisión, Sensibilidad (Recall) y F1-Score de la clase de interés (Def: 1). Véase la siguiente Tabla.

Tabla 19. Comparación del rendimiento de modelos con datos desbalanceados

	Total	**Def: 1**		
Modelo	**Exactitud**	**Precisión**	**Sensibilidad (Recall)**	**F1-score**
LR	0,98	0,71	0,17	0,28
SVM	0,98	0,00	0,00	0,00
DT	0,97	0,26	0,34	0,30
RF	0,99	**0,95**	**0,30**	**0,46**
XGB	0,99	**0,87**	**0,29**	**0,43**

KNN	0,98	0,00	0,00	0,00
NB	0,97	0,14	0,13	0,13

Fuente: Elaboración propia

Como se observa, en la Tabla 17, todos los modelos lograron una Exactitud general promedio >97%, sobre la clase de interés. Sin embargo esta precisión, como bien es cierto, puede ser engañosa, a razón del desequilibrio de los datos. Al entrar en detalle, los resultados muestran las siguientes situaciones:

- Random Forest (RF) demostró una exactitud del 99%, indicando una alta tasa de predicciones correctas en general. Sin embargo, se observa una disminución en la precisión (95%), sensibilidad (30%), y F1-score (46%).
- XGBoost (XGB) mostró igualmente una exactitud del 99%, pero se observa una disminución en la precisión (87%), sensibilidad (29%), y F1-score (43%).
- Regresión Logística (LR) presentó una exactitud del 98%, pero muestra valores bajos en precisión (71%), sensibilidad (17%), y F1-score (28%)..
- Support Vector Machine (SVM) obtuvo una alta exactitud del 98%, pero mostro valores nulos en precisión, sensibilidad y F1-score.
- K-Nearest Neighbors (KNN) al igual que el anterior también presentó valores nulos en precisión, sensibilidad y F1-score, a pesar de tener una alta exactitud del 98%.
- Decision Tree (DT) obtuvo una exactitud del 97%. Sin embargo, la precisión es baja (26%), y aunque la sensibilidad es relativamente alta (34%), el F1-score es moderado (30%).
- Naive Bayes (NB) presentó una exactitud del 97%, pero con valores bajos en precisión (14%), sensibilidad (13%), y F1-score (13%).

En general, a pesar de altos niveles de exactitud, se observan disminuciones sustanciales en las demás métricas, lo que expone la dificultad de los modelos en su capacidad para clasificar correctamente la variable de interés.

5.2. Aplicación de técnicas de balanceo de los datos

Para reducir el impacto del desbalanceo de los datos en el rendimiento de los modelos, se aplicaron las siguientes técnicas: Sobremuestreo, Submuestreo y SMOTE (Synthetic Minority Over-sampling Technique).

Se utilizaron las bibliotecas ***over_sampling.SMOTE*** y ***under_sampling, RandomUnderSampler*** de la biblioteca ***implaren*** la cual proporciona una implementación de SMOTE y otras técnicas de manejo de desequilibrio de clases. La técnica SMOTE permitió generar muestras sintéticas para la clase minoritaria,

aumentando su representación en el conjunto de datos. Al mismo tiempo, ***RandomUnderSampler*** seleccionó aleatoriamente un subconjunto de muestras de la clase mayoritaria, reduciendo su cantidad de registros hasta igualarla con la clase de defunciones. Y finalmente, la función ***resample*** de ***scikit-learn*** se aplicó para equilibrar las proporciones de clases, mediante el oversampling tradicional.

Con ayuda de las librerías de Python, se puede decir que el SMOTE y el Sobremuestreo son técnicas similares, en tanto que se enfocan en la generación de registros aumentando la clase minoritaria, pero la diferencia radica en que el SMOTE genera para nuevos registros, también cambios en las variables que toma, por lo que pueden ser llamados datos sintéticos. Esto por supuesto, aumenta el riesgo de generar registros no representativos de la población original.

En cuanto al submuestreo, se utilizó para llevar a cabo el ejercicio comparativo, sin embargo, dado que el conjunto de datos, en especial la clase desbalanceada tiene relativamente pocos datos (n = 209) es una desventaja, puesto que se reduce significativamente el total de datos para modelar.

A continuación, en la Tabla 18 se muestran los resultados de las técnicas de balanceo de los datos y el ajuste de los algoritmos anteriormente aplicados, con el fin de tener una comparación de las métricas resultantes bajo los mismos parámetros que por defecto arrojan los módulos y paquetes utilizados.

Tabla 20. Resultados de la aplicación de técnicas de balanceo al conjunto de datos.

Método	Def: 0	Def:1
Original	**14271**	**209**
Submuestreo (Under-Sampling)	209	209
Sobremuestreo (Over-Sampling)	14271	14271
SMOTE (Synthetic Minority Over-sampling)	14271	14271

Fuente: Elaboración propia

En el Submuestreo, se redujo la clase mayoritaria hasta tener el mismo tamaño de la clase minoritaria eliminando aleatoriamente muestras de la clase mayoritaria. Por su parte, en el Sobremuestreo, la clase minoritaria se igualó a la mayoritaria replicando aleatoriamente muestras de la clase minoritaria. Y bajo el SMOTE, estas también se igualaron mediante la generación de datos sintéticos de la clase minoritaria para aumentar su presencia en el conjunto de datos.

5.3. Entrenamiento y evaluación de modelos con datos balanceados

Posterior a la aplicación de técnicas de balanceo, se entrenaron nuevamente los modelos y se evaluaron bajo las mismas métricas utilizadas anteriormente. Los resultados, incluido los obtenidos con el conjunto de datos desbalanceado, se presentan en la siguiente tabla.

Tabla 21. Comparación del rendimiento de los modelo con los datos balanceados

Modelo	Método de Balanceo	Exactitud	Precisión	Sensibilidad	F1-Score
RL	Desbalanceado	0.98	0.71	0.17	0.28
	Sobremuestreo	0.84	0.06	0.71	0.12
	SMOTE	0.98	0.27	0.37	0.31
	Submuestreo	0.81	0.05	0.71	0.09
SVM	Desbalanceado	0.98	0.00	0.00	0.00
	Sobremuestreo	0.78	0.03	0.41	0.05
	SMOTE	0.76	0.03	0.41	0.05
	Submuestreo	0.86	0.04	0.37	0.07
DT	Desbalanceado	0.97	0.26	0.34	0.30
	Sobremuestreo	0.98	0.36	0.37	0.36
	SMOTE	0.97	0.27	0.40	0.32
	Submuestreo	0.67	0.03	0.76	0.06
RF	Desbalanceado	0.99	0.95	0.30	0.46
	Sobremuestreo	0.99	0.90	0.27	0.42
	SMOTE	0.99	0.86	0.27	0.41
	Submuestreo	0.80	0.05	0.76	0.10
XGB	Desbalanceado	0.99	0.87	0.29	0.43
	Sobremuestreo	0.99	0.54	0.34	0.42
	SMOTE	0.99	0.76	0.33	0.46
	Submuestreo	0.74	0.04	0.74	0.08
KNN	Desbalanceado	0.98	0.00	0.00	0.00
	Sobremuestreo	0.97	0.16	0.30	0.21
	SMOTE	0.91	0.07	0.41	0.11
	Submuestreo	0.69	0.02	0.59	0.05
NB	Desbalanceado	0.97	0.14	0.13	0.13
	Sobremuestreo	0.18	0.02	0.94	0.03
	SMOTE	0.34	0.02	0.80	0.03
	Submuestreo	0.94	0.12	0.49	0.19

Fuente: Elaboración propia

Como se observa, la aplicación de las técnicas de balanceo de datos generó una alta exactitud en la mayoría de los modelos por encima de 70%. Sin embargo, dado el problema de identificar defunciones infantiles, la exactitud del modelo por sí sola no garantiza una métrica suficiente para evaluar la capacidad del modelo. En detalle, los modelos arrojaron los siguientes resultados:

- Regresión Logística (RL): en este modelo, el Sobremuestreo y el Submuestreo mejoraron la Sensibilidad, pero a expensas de una disminución significativa en la Precisión. Mediante SMOTE presentó mejoras en Sensibilidad y F1-Score, manteniendo una precisión relativamente estable.

- Support Vector Machine (SVM): para este modelo, a pesar de la aplicación de las técnicas de balanceo, la Exactitud se mantuvo en un promedio superior al 84% y se mejoró la Sensibilidad; sin embargo no logró mejorar ninguna de sus otras métricas Precisión y F1-Score; esto confirma su incapacidad para identificar correctamente casos críticos de defunción infantil..

- Decision Tree (DT): en este caso, el Sobremuestreo y SMOTE mejoraron todas las métricas; con el Submuestreo, se evidencio una fuerte disminución en todas excepto en la Sensibilidad que mejoro sustancialmente.

- Random Forest (RF): en este modelo, el Sobremuestreo y SMOTE mantuvieron la exactitud y la precisión con resultados superiores al 86%; sin embargo lograron disminuir las sensibilidad y el F1-Score, pero hay una disminución en la precisión. El Submuestreo presentó una disminución en todas las métricas excepto por la mejora notable en la Sensibilidad.

- XGBoost (XGB): en este caso, el Sobremuestreo y SMOTE mantuvieron la Exactitud, disminuyeron la Precisión; el SMOTE mejoró el F1-Score. Aunque el Submuestreo mejora la Sensibilidad, se afectan considerablemente todas las demás métricas incluida la Exactitud.

- K-Nearest Neighbors (KNN): aquí, las técnicas balanceo disminuyeron la Exactitud; si bien hubo aumentos en todas las demás métricas, estos no fueron significativos. El F1-Score mejora con el Sobremuestreo y SMOTE

- Naive Bayes (NB): en este modelo el Sobremuestreo y SMOTE mejoran únicamente la sensibilidad, pero con una disminución significativa en la precisión. El Submuestreo muestra una Exactitud favorable, pero disminuyendo la Precisión. F1-Score aumento levemente, aunque sigue siendo bajo.

En general, la Sensibilidad y el F1-Score mostraron una mejora consistente después de la aplicación de técnicas de balanceo de Sobremuestreo y SMOTE en los modelos, mejorando la capacidad de estos para identificar la clase minoritaria, que es la muerte infantil.

5.4. Optimización de modelos

En el contexto de un modelo de predicción que se enfoca en la mortalidad de recién nacidos a término con bajo peso al nacer durante su primer año de vida, es fundamental considerar la métrica de Sensibilidad; esta métrica cobra relevancia debido al desbalanceo de datos inherente a este tipo de problemas. La Sensibilidad mide la capacidad del modelo para identificar correctamente los casos positivos, en este caso, los recién nacidos con bajo peso que lamentablemente experimentarán una muerte temprana.

Dado que la vida de estos recién nacidos es de suma importancia, maximizar la Sensibilidad es esencial; un alto valor de Sensibilidad asegura que el modelo sea efectivo en la detección temprana de riesgos de mortalidad, lo que puede llevar a intervenciones médicas oportunas y, en última instancia, a la salvación de vidas. En este contexto, la pérdida de un caso positivo (falso negativo) podría tener consecuencias graves, y la Sensibilidad ayuda a minimizar esta posibilidad.

Sin embargo, es importante equilibrar la Sensibilidad con otras métricas, como la Precisión y el F1-score ya que un alto énfasis en la Sensibilidad puede llevar a un aumento en los falsos positivos. Es por esto que los resultados encontrados en esta sección se esmeran por ilustrar los resultados de los modelos teniendo en cuenta como métricas de optimización el F1-score y la Sensibilidad.

Dicho lo anterior, el proceso de optimización de los modelos se llevó a cabo usando la biblioteca de Aprendizaje automático PyCaret; esta biblioteca de aprendizaje automático permite agilizar el flujo de trabajo para la automatización de ciertas tareas rutinarias.

Aprovechando la funcionalidad para el ajuste de hiperparámetros “tune_model()” de PyCaret en los 7 modelos ajustados a los datos de los 4 subconjuntos de datos (Desbalanceados, Submuestreo, Sobremuestreo y SMOTE) a continuación, se detalla la metodología utilizada para realizar este proceso:

1) Preparación del entorno en PyCaret

   Se preparó el entorno en PyCaret configurando la sesión con **setup()**, donde se especificaron las características del conjunto de datos, y los argumentos para indicar el balanceo de los datos mediante las 3 técnicas aplicadas.

2) Lista de modelos para afinar

Se creó la lista de modelos a optimizar donde se incluyó regresión logística (**'lr'**), máquinas de vectores de soporte ('**svm**'), árboles de decisión ('**dt'**), bosques aleatorios ('**rf**'), XGBoost ('**xgboost**'), k-vecinos más cercanos ('**knn**'), y Naive Bayes ('**nb**'). Estos modelos fueron ingresados al proceso de optimización que utiliza la validación cruzada de K-folds con k=10.

3) Ciclo de optimización para cada modelo

Para cada modelo en la lista, se realizó un ciclo de optimización:

i. Creación del modelo: utilizando create_model(m), se creó una instancia del modelo especificado.

ii. Afinación de hiperparámetros: con tune_model(), se realizó la afinación de hiperparámetros. Los parámetros clave incluyeron:
   a. n_iter: número de iteraciones en la búsqueda de hiperparámetros. En este caso se utilizaron 20 interacciones, para un total de (K=10 * n=20 = 200 modelos).
   b. optimize: la métrica de rendimiento a optimiza; en este caso se probaron 2 escenarios, 'Recall' y 'F1' haciendo referencia a la sensibilidad y al F1-score.
   c. search_algorithm: Algoritmo de búsqueda, aquí 'random', que indica una búsqueda aleatoria.
   d. early_stopping: el cual es útil para optimizar el proceso de optimización puesto que detiene la búsqueda tempranamente si no se observan mejoras.
   e. choose_better: selecciona el modelo afinado solo si supera al original.

iii. Extracción de Métricas y Hiperparámetros:
   a. Métricas: Las métricas medias del modelo afinado se extrajeron con pull().loc['Mean'] y se agregaron ordenadamente a un dataframe.
   b. Hiperparámetros: Los hiperparámetros del modelo afinado se obtuvieron con tuned.get_params() y se almacenaron en el diccionario model_params_unb.

Una vez completado el ciclo de optimización para todos los modelos, se analizaron y compararon las métricas de rendimiento y los hiperparámetros. Esto ayudó a identificar qué modelos y configuraciones de hiperparámetros ofrecieron el mejor rendimiento según la métrica de interés. A continuación, se presentan los resultados obtenidos en el proceso

de búsqueda aleatoria de hiperparámetros, optimizando la métrica de Sensibilidad (Recall) y el F1-Score respectivamente.

Tabla 22. Métricas promedio de modelos optimizados con Sensibilidad

Método y Modelo	Promedio Exactitud	Promedio AUC	Promedio Sensibilidad	Promedio Precisión	Promedio F1
Sin balanceo					
lr	0.828	0.851	0.728	0.059	0.109
svm	0.975	0.000	0.201	0.177	0.187
dt	0.986	0.761	0.149	0.685	0.233
rf	0.893	0.830	0.651	0.084	0.149
xgboost	0.914	0.845	0.618	0.100	0.172
knn	0.985	0.656	0.283	0.493	0.353
nb	0.618	0.815	0.814	0.032	0.062
Submuestreo					
lr	0.847	0.841	0.746	0.068	0.124
svm	0.975	0.000	0.239	0.201	0.216
dt	0.985	0.705	0.167	0.420	0.232
rf	0.918	0.845	0.645	0.108	0.186
xgboost	0.958	0.829	0.511	0.179	0.265
knn	0.989	0.648	0.220	1.000	0.359
nb	0.552	0.827	**0.842**	0.029	0.055
Sobremuestreo					
lr	0.825	0.823	0.713	0.057	0.105
svm	0.974	0.000	0.182	0.144	0.159
dt	0.985	0.662	0.167	0.441	0.231
rf	0.906	0.834	0.612	0.091	0.158
xgboost	0.929	0.835	0.545	0.108	0.180
knn	0.988	0.681	0.172	1.000	0.288
nb	0.267	0.800	**0.928**	0.018	0.035
SMOTE					
lr	0.840	0.835	0.713	0.062	0.114
svm	0.975	0.000	0.206	0.178	0.190
dt	0.986	0.759	0.138	0.556	0.216
rf	0.899	0.852	0.665	0.092	0.161
xgboost	0.925	0.844	0.617	0.115	0.194
knn	0.982	0.642	0.292	0.370	0.319
nb	0.658	0.808	**0.785**	0.034	0.065

Fuente: Elaboración propia

Bajo este enfoque, la optimización de los modelos permite aumentar la Sensibilidad de los modelos para identificar correctamente los casos de MI, minimizando la identificación de falsos negativos. Con respecto al Sobremuestreo, **NB** obtuvo la Sensibilidad más alta de la tabla (92,8%). Si bien NB demostró ser efectivo en la identificación de casos positivos, su baja Precisión (1,8%) resultó en un F1-score considerablemente más bajo

(3,5%). Este escenario refleja la dificultad de equilibrar Sensibilidad y Precisión en problemas de clasificación desequilibrada.

Tabla 23. Métricas promedio de modelos optimizados con F1-Score.

Método y Modelo	Promedio Exactitud	Promedio AUC	Promedio Sensibilidad	Promedio Precisión	Promedio F1
Sin balanceo					
lr	0,986	0,841	0,130	0,526	0,203
svm	0,975	0,000	0,201	0,177	0,187
dt	0,986	0,761	0,149	0,685	0,233
rf	0,959	0,862	**0,613**	**0,200**	**0,301**
xgboost	0,988	0,850	**0,360**	**0,650**	**0,456**
knn	0,989	0,680	0,259	1,000	0,405
nb	0,963	0,819	0,412	0,178	0,248
Submuestreo					
lr	0,986	0,842	0,129	0,524	0,201
svm	0,975	0,000	0,239	0,201	0,216
dt	0,985	0,705	0,167	0,420	0,232
rf	0,969	0,834	**0,507**	**0,240**	**0,325**
xgboost	0,987	0,807	**0,344**	**0,603**	**0,430**
knn	0,989	0,648	0,220	1,000	0,359
nb	0,961	0,830	0,450	0,172	0,248
Sobremuestreo					
lr	0,986	0,824	0,106	0,475	0,172
svm	0,974	0,000	0,182	0,144	0,159
dt	0,985	0,662	0,167	0,441	0,231
rf	0,959	0,814	0,459	0,166	0,243
xgboost	0,989	0,793	**0,315**	**0,746**	**0,435**
knn	0,988	0,681	0,172	1,000	0,288
nb	0,954	0,821	0,455	0,146	0,220
SMOTE					
lr	0,986	0,837	0,120	0,558	0,187
svm	0,975	0,000	0,206	0,178	0,190
dt	0,986	0,759	0,138	0,556	0,216
rf	0,975	0,823	0,421	0,273	0,328
xgboost	0,987	0,842	**0,326**	**0,577**	**0,413**
knn	0,989	0,652	0,249	0,980	0,391
nb	0,953	0,831	0,460	0,147	0,222

Fuente: Elaboración propia

Bajo la métrica F1-Score, la optimización indicó que los modelos con los mejores promedios resultaron ser **XGBoost**, con 43% para Submuestreo, 43,5% para Sobremuestreo y 41,3% para el SMOTE; esto quiere decir que mediante la optimización, este modelo logra un rendimiento robusto encontrando un balance entre precisión y sensibilidad, es decir minimizando tanto los falsos positivos como los falsos negativos. Cabe destacar que en estos modelos, la Sensibilidad más alta se alcanzó en el escenario sin balanceo de datos (36%), mientas que la Precisión estuvo por encima del Recall en todos los escenarios, contribuyendo a los valores elevados del F1-score mencionados previamente.

Después de la búsqueda aleatoria anterior, se aplicó la Optimización por Rejilla para realizar un afinamiento adicional alrededor de las áreas donde se obtuvieron los mejores resultados. Cabe anotar que para los escenarios anteriores, la métrica AUC para SVM (Support Vector Machine) fue nula y sus métricas estuvieron por debajo del promedio, lo que indica que el modelo tuvo dificultades persistentes para distinguir entre las clases; esto sugiere que SVM puede no ser la mejor opción para estos datos incluso con la búsqueda aleatoria de hiperparámetros que lo mejoren. Es por esto que, para la optimización por rejilla este modelo no se tuvo en cuenta.

5.4.1. Optimización por Rejilla

Para llevar a cabo el proceso de Optimización por Rejilla se utilizó la librería GridSearchCV. Esta herramienta es esencial en el aprendizaje automático; opera en un ámbito de exploración y optimización, donde primero define una grilla de hiperparámetros. Luego, entrena cada modelo posible, establecido mediante cada combinación de estos hiperparámetros, guiado por una métrica predeterminada. Con la validación cruzada, asegura una evaluación equilibrada y consistente, evitando la dependencia de un único conjunto de datos. Al final de esta búsqueda exhaustiva, emerge con el mejor modelo, aquel que se destaca en rendimiento según la métrica elegida, marcando el final de su misión con una decisión informada y precisa.

Para su uso se llevó a cabo la siguiente metodología:

i. Definición del Pipeline: Establecimos un pipeline en Scikit-Learn para cada modelo de clasificación. Este pipeline integró básicamente la aplicación de las instancias de modelado de cada uno de los 7 modelos a aplicar.
ii. Configuración de la Grilla de Hiperparámetros: Para cada modelo, definimos una grilla de hiperparámetros. Esta grilla incluyó diversas combinaciones de parámetros que consideramos relevantes para la optimización del modelo. Ver Anexo # 1.
iii. Aplicación de GridSearchCV: Implementamos GridSearchCV, una herramienta de Scikit-Learn que facilitó la búsqueda exhaustiva a través de la grilla de hiperparámetros. Configuramos GridSearchCV con validación cruzada, utilizando el número de pliegues definido para garantizar la robustez de los resultados:
    a. Número de corridas de validación cruzada igual a 10
    b. Métricas a optimizar F1-score y Recall.
iv. Selección del Mejor Modelo y Hiperparámetros: Una vez finalizada la búsqueda, GridSearchCV nos proporcionó el conjunto de hiperparámetros que logró el mejor rendimiento, basándonos en la métrica de puntuación que habíamos elegido.
v. Evaluación del Modelo Optimizado: Con el modelo ya optimizado, procedimos a evaluar su rendimiento en el conjunto de prueba. Calculamos métricas clave como

la Exactitud, Precisión, Sensibilidad (Recall) y el puntaje F1-Score para obtener una visión integral del rendimiento del modelo.

Las siguientes tablas revelan los resultados de la Optimización por Rejilla.

Tabla 24. Optimización por grilla de hiperparámetros - Sensibilidad.

Modelo	Subconjunto	Hiperparámetros	Exactitud	Sensibilidad	Precisión	F1
NB	Desbalanceado	_var_smoothing: 1e-09	0,295	**0,943**	0,019	0,037
NB	Sobremuestreo	_var_smoothing: 1e-09	0,186	**0,943**	0,017	0,032
NB	Smote	_var_smoothing: 1e-09	0,339	**0,800**	0,017	0,034
XGB	Submuestreo	_learning_rate: 0.1, _max_depth: 3, _n_estimators: 200, _subsample: 0.7	0,753	**0,757**	0,043	0,082
RF	Submuestreo	_max_depth: 20, _min_samples_leaf: 1, _min_samples_split: 5, _n_estimators: 100	0,823	**0,729**	0,058	0,107
LR	Submuestreo	_C: 10, _penalty: l2, _solver: liblinear	0,801	0,714	0,050	0,094
DT	Submuestreo	_max_depth: 20, _min_samples_leaf: 1, _min_samples_split: 2	0,692	0,714	0,033	0,063
KNN	Submuestreo	_metric: euclidean, _n_neighbors: 5, _weights: distance	0,689	0,686	0,031	0,060
LR	Sobremuestreo	_C: 10, _penalty: l2, _solver: newton-cg	0,846	0,657	0,060	0,110
NB	Submuestreo	_var_smoothing: 1e-07	0,929	0,586	0,115	0,192
XGB	Sobremuestreo	_learning_rate: 0.01, _max_depth: 10, _n_estimators: 100, _subsample: 0.7	0,952	0,486	0,147	0,226
DT	Smote	_max_depth: 20, _min_samples_leaf: 1, _min_samples_split: 2	0,973	0,443	0,256	0,325
DT	Sobremuestreo	_max_depth: None, _min_samples_leaf: 1, _min_samples_split: 2	0,982	0,386	0,391	0,388
LR	Smote	_C: 10, _penalty: l2, _solver: newton-cg	0,983	0,357	0,410	0,382
DT	Desbalanceado	_max_depth: None, _min_samples_leaf: 1, _min_samples_split: 2	0,979	0,357	0,305	0,329
KNN	Smote	_metric: manhattan, _n_neighbors: 3, _weights: uniform	0,958	0,329	0,130	0,186

Modelo	Subconjunto	Hiperparámetros	Exactitud	Sensibilidad	Precisión	F1
XGB	Desbalanceado	_learning_rate: 0.2, _max_depth: 6, _n_estimators: 200, _subsample: 1	0,989	0,314	0,786	0,449
RF	Desbalanceado	_max_depth: None, _min_samples_leaf: 1, _min_samples_split: 2, _n_estimators: 200	0,989	0,286	0,952	0,440
RF	Sobremuestreo	_max_depth: None, _min_samples_leaf: 1, _min_samples_split: 2, _n_estimators: 100	0,990	0,286	1,000	0,444
XGB	Smote	_learning_rate: 0.1, _max_depth: 6, _n_estimators: 100, _subsample: 0.9	0,987	0,286	0,588	0,385
KNN	Desbalanceado	_metric: euclidean, _ n_neighbors: 3, _weights: distance	0,987	0,286	0,645	0,396
KNN	Sobremuestreo	_metric: euclidean, _n_neighbors: 3, _weights: uniform	0,973	0,286	0,204	0,238
LR	Desbalanceado	_C: 0.1, _ penalty: l2, _solver: newton-cg	0,987	0,171	0,706	0,276
RF	Smote	_max_depth: 20, _min_samples_leaf: 4, _min_samples_split: 10, _ n_estimators: 200	0,986	0,143	0,625	0,233

Fuente: Elaboración propia

El modelo Naive Bayes en datos desbalanceados logra una alta Sensibilidad del 94.3%, lo que significa que es eficiente para detectar los casos críticos, pero su Precisión es muy baja (1,9%). Esto puede resultar en un mayor número de falsos positivos, lo que puede ser crítico en un contexto clínico. En otras palabras, este modelo tiende a identificar muchos casos positivos, pero también genera una cantidad significativa de falsos positivos, lo que debe considerarse en un entorno de despliegue, de acuerdo con lo que implique el hecho de clasificar dentro del riesgo alto de mortalidad a un NV con BPNT que no lo esté sufriendo realmente.

Por otro lado, los modelos XGBoost y Random Forest en datos de submuestreo mantienen un buen equilibrio entre Sensibilidad, Precisión y F1-score. Tienen una sensibilidad del 75.7% y 72.9% respectivamente, lo que significa que aún son capaces de identificar un porcentaje sustancial de casos críticos. Además, tienen una Precisión más aceptable que el modelo Naive Bayes en datos desbalanceados, lo que reduce la probabilidad de falsos positivos. Estos modelos podrían ser preferibles en un entorno de atención médica donde la reducción de falsos positivos es crítica.

Tabla 25. Optimización por grilla de hiperparámetros – F1-Score

Modelo	Subconjunto	Hiperparámetros	Exactitud	Sensibilidad	Precisión	F1
XBG	Smote	learning_rate: 0.2, max_depth: 10, n_estimators: 200, subsample: 0.9	0,989	0,371	0,812	**0,509**
XBG	Desbalanceado	learning_rate: 0.2, max_depth: 6, n_estimators: 200, subsample: 1	0,988	0,314	0,785	**0,449**
RF	Desbalanceado	max_depth: None, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 100	0,989	0,271	0,950	**0,422**
XBG	Sobremuestreo	learning_rate: 0.2, max_depth: 10, n_estimators: 200, subsample: 1	0,986	0,328	0,589	**0,422**
RF	Sobremuestreo	max_depth: None, min_samples_leaf: 1, min_samples_split: 5, n_estimators: 100	0,989	0,271	0,904	**0,418**
NB	Desbalanceado	metric: euclidean, n_neighbors: 5, weights: distance	0,989	0,271	0,864	0,413
KNN	Desbalanceado	metric: euclidean, neighbors: 5, weights: distance	0,989	0,271	0,864	0,413
RL	Smote	C: 10 penalty: l2 solver: newton-cg	0,983	0,357	0,409	0,381
DT	Sobremuestreo	max_depth: None, min_samples_leaf: 1, min_samples_split: 5	0,980	0,371	0,346	0,358
RL	Desbalanceado	C: 0.1 penalty: l2 solver: liblinear	0,987	0,171	0,750	0,279
DT	Desbalanceado	max_depth: 10, min_samples_leaf: 1, min_samples_split: 5	0,985	0,185	0,464	0,265
NB	Sobremuestreo	metric: manhattan, n_neighbors: 3, weights: distance	0,976	0,286	0,233	0,256
KNN	Sobremuestreo	metric: manhattan, neighbors: 3, weights: distance	0,976	0,286	0,233	0,256
RF	Smote	max_depth: None, min_samples_leaf: 2, min_samples_split: 2, n_estimators: 100	0,987	0,142	0,833	0,243
DT	Smote	max_depth: 20, min_samples_leaf: 2, min_samples_split: 10	0,981	0,200	0,291	0,237
NB	Smote	metric: manhattan, n_neighbors: 3, weights: distance	0,963	0,329	0,146	0,203
KNN	Smote	metric: manhattan, neighbors: 3, weights: distance	0,963	0,329	0,146	0,203

RL	Sobremuestreo	C: 1 penalty: l2 solver: newton-cg	0,846	0,657	0,060	0,110
RF	Submuestreo	max_depth: 20, min_samples_leaf: 1, min_samples_split: 5, n_estimators: 200	0,812	0,742	0,055	0,102
RL	Submuestreo	C: 1 penalty: l2 solver: lbfgs	0,802	0,685	0,049	0,091
XBG	Submuestreo	learning_rate: 0.1, max_depth: 3, n_estimators: 200, subsample: 0.7	0,753	0,757	0,043	0,081
NB	Submuestreo	metric: manhattan, n_neighbors: 11, weights: distance	0,781	0,643	0,042	0,079
KNN	Submuestreo	"metric: manhattan, neighbors: 11, weights: distance	0,781	0,643	0,042	0,079
DT	Submuestreo	max_depth: None, min_samples_leaf: 2, min_samples_split: 2	0,701	0,614	0,029	0,056

Fuente: Elaboración propia

Los modelos de XGBoost en diferentes escenarios, Desbalanceado, Sobremuestreo y SMOTE, lograron un F1-Score notable, superando el 42%. Esto indica que estos modelos pudieron equilibrar de manera efectiva tanto la capacidad para identificar casos positivos (Sensibilidad) como la Precisión en la clasificación. En particular, el modelo XGBoost con SMOTE alcanzó el F1-Score más alto, llegando a 50,9%. Esto sugiere que este modelo tiene un rendimiento sólido en la detección de casos de bajo peso al nacer que experimentan muertes en su primer año de vida, lo que es de gran relevancia en un entorno de atención médica.

Una vez entrenado nuestro modelo de XGBoost con los datos y los hiperparámetros especificados, procedemos a realizar un análisis de importancia de características. El modelo fue entrenado con los rendimientos obtenidos por la optimización por grilla de hiperparámetros de Sensibilidad.

Tabla 26. Modelo XGBoost y su rendimiento

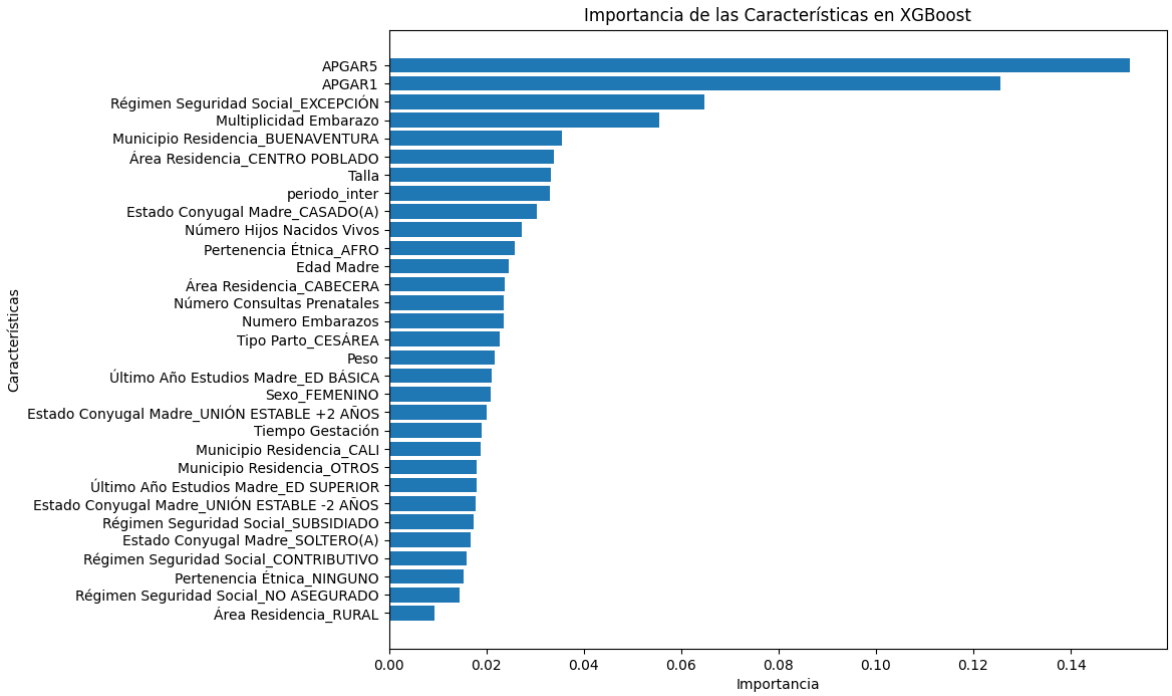
Modelo	Subconjunto	Hiperparámetros	Exactitud	Sensibilidad	Precisión	F1
XGB	Submuestreo	_learning_rate: 0.1, _max_depth: 3, _n_estimators: 200, _subsample: 0.7	0,753	0,757	0,043	0,082

Fuente: Elaboración propia

Este análisis nos permitió identificar qué características tuvieron un impacto más significativo en las predicciones del modelo; las características más importantes son

aquellas que contribuyen de manera crucial a la toma de decisiones del modelo. En el siguiente gráfico, se muestra la importancia de las características de manera ordenada, destacando las más influyentes en la capacidad predictiva del modelo. Además, se filtraron las características cuya importancia es igual a cero o nula, lo que permitió centrarnos en las más relevantes.

Gráfica 3. Características del modelo entrenado XGBoost



Fuente: Elaboración propia

Como se observa, las variables "APGAR5" y "APGAR1" fueron las más importantes pues son indicadores de evaluaciones que se realizan minutos después del nacimiento y son vitales para determinar la viabilidad y la salud inmediata del neonato; esto resulta clave entonces en el pronóstico de supervivencia en individuos con bajo peso al nacer. La "Multiplicidad Embarazo" y "periodo_inter" también son importantes, posiblemente reflejando cómo las circunstancias del embarazo, como embarazos múltiples o el intervalo entre nacimientos, pueden influir en el riesgo de complicaciones que podrían llevar a defunciones en recién nacidos con bajo peso. Los aspectos relacionados con el "Régimen de Seguridad Social" sugieren que el acceso a los servicios de salud y la cobertura de seguro son factores significativos que afectan los resultados de salud en los recién nacidos con bajo peso.

6. CONCLUSIONES Y TRABAJOS FUTUROS

6.1. CONCLUSIONES

Al analizar el desempeño de diversos modelos de aprendizaje automático en la tarea de predecir si un nacido vivo con BPNT experimentará una muerte en su primer año de vida, podemos destacar varios hallazgos importantes:

- Desafío del desbalanceo de datos: La naturaleza desbalanceada de los datos, con una escasez de casos positivos en comparación con los negativos, planteó un desafío importante. En particular, el objetivo era maximizar la capacidad de detectar casos positivos (Sensibilidad). Sorprendentemente, en tres de los seis mejores modelos optimizados mediante GridSearchCV, tanto para F1-Score como para Sensibilidad, los datos desbalanceados obtuvieron los mejores resultados. Esto sugiere que es válido considerar los datos tal como se presentan, centrándose más en las técnicas de modelado, que en el balanceo de datos.

- Técnica de Submuestreo: El uso de Submuestreo, que implica reducir el tamaño de la clase mayoritaria, resultó en un aumento significativo en la Sensibilidad para la mayoría de los modelos. Esto indica que al equilibrar las clases, los modelos pudieron identificar mejor la clase minoritaria, lo cual es fundamental para la detección de casos críticos. Sin embargo, es importante destacar que este enfoque puede llevar a una disminución en la Precisión, lo que a su vez afecta el valor del F1-Score.

- Selección de modelos: Dentro de los modelos evaluados, Naive Bayes (NB), XGBoost (XGB) y Random Forest (RF) destacaron en términos de Sensibilidad. Esto sugiere que estos modelos podrían ser adecuados para identificar casos críticos en un entorno clínico.

- Ajuste de hiperparámetros: El proceso de ajuste de hiperparámetros desempeñó un papel esencial en el rendimiento de los modelos. Parámetros específicos, como "var_smoothing" en NB, "_learning_rate" y "_max_depth" en XGB, y "_max_depth" y

"_min_samples_split" en RF, resultaron fundamentales para equilibrar la sensibilidad y la precisión de los modelos.

- Tiempo de ejecución: El tiempo de entrenamiento y evaluación de Random Forest (RF) en los cuatro datasets, que totalizó 59.5 minutos, es un factor importante a considerar en el proceso de desarrollo de modelos. Si bien RF demostró un buen rendimiento en términos de métricas como Sensibilidad y Precisión, es esencial tener en cuenta que su tiempo de ejecución puede ser significativo. La elección de este modelo en un contexto de despliegue debe considerarse teniendo en cuenta que en situaciones de configuración del modelo donde los datos de entrada cambian con el tiempo, o se requieren mantenimiento y verificación del mismo, que conllevan la necesidad de reentrenar el modelo para adaptarse a los nuevos patrones.

- Consideraciones clínicas: Dado nuestro perfil como Científicos de Datos con conocimiento clínico medio pero no expertos en salud pública, es crucial destacar que, en un entorno de atención médica, la elección del modelo y la métrica de optimización deben evaluarse cuidadosamente en función de las implicaciones clínicas de los falsos negativos y positivos.

6.2. TRABAJO FUTUROS

La investigación realizada sobre la predicción de la mortalidad infantil asociada al bajo peso al nacer a término (BPNT-MI) destaca la importancia de anticiparse a este tipo de fenómenos mediante el desarrollo continuo de modelos de aprendizaje automático. En este marco, es indispensable profundizar en la utilización de datos provenientes del Portal Alto Riesgo Obstétrico (ARO) de la Secretaría de Salud Pública Distrital de Cali, el cual concentra información detallada sobre gestantes de alto riesgo, incluyendo aspectos sociodemográficos, complicaciones durante el embarazo y el parto, y datos actualizados sobre el proceso de atención durante la etapa gestacional.

La integración de datos provenientes del Registro Único de Afiliados al Sistema de Salud Nacimientos y Defunciones RUAF- ND y el Portal ARO se presenta como una herramienta de gran relevancia para el sistema de salud. Esta integración no solo optimiza el flujo de información, sino que también contribuye a una toma de decisiones más eficiente y precisa tanto a nivel asistencial como en el diseño de programas institucionales de control. Desde el punto de vista asistencial, la integración de datos permite a los profesionales de la salud realizar estimaciones robustas y tempranas sobre el estado de salud de las gestantes y los recién nacidos, posibilitando la detección

anticipada de posibles complicaciones, la implementación de intervenciones preventivas y la planificación de la atención médica de manera personalizada.

En el ámbito de la investigación, la inclusión de datos del RUAF-ND y del Portal ARO para evaluar la generalización de los modelos de predicción del bajo peso al nacer podría mejorar la confiabilidad y solidez de dichos modelos. Además, es evidente la necesidad de continuar colaborando con las secretarías de salud departamental y distrital, ya que son las proveedoras de los conjuntos de datos para implementar mejores controles en el levantamiento de información, abordando limitaciones como la idoneidad de la atención prenatal y mitigando la falta de registros.

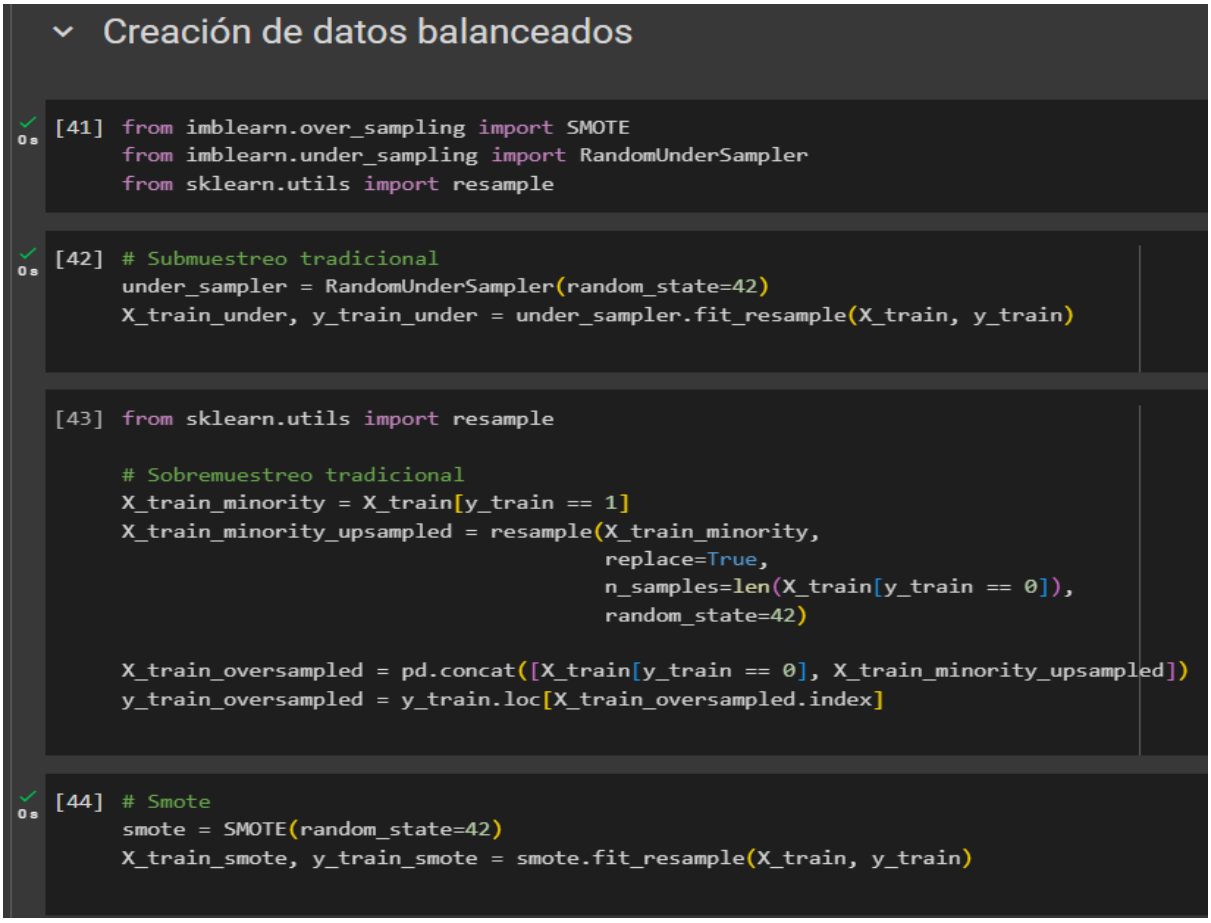
Como perspectiva futura, se sugiere explorar también la predicción del bajo peso al nacer y la mortalidad infantil como una alternativa adicional de investigación, ampliando así el enfoque y la aplicación práctica de los modelos desarrollados. Este nuevo horizonte de trabajo podría contribuir significativamente a la prevención y atención temprana de complicaciones relacionadas con el peso al nacer y la mortalidad infantil, fortaleciendo la eficacia de las intervenciones médicas y mejorando los resultados de salud infantil.

En trabajos futuros se podrían agregar otras variables de tipo médico en una articulación de científicos de datos, médicos y otros profesionales expertos en el comportamiento de la MI y el BPN, así mismo profundizar en el análisis el comportamiento de estos dos fenómenos con la selección de variables mediante métodos estadísticos.

7. ANEXOS

Anexo # 1. Técnicas y funciones para el Balanceo de Datos en Py.Caret

1. Técnicas de balanceo de los datos:



Creación de datos balanceados

```
[41] from imblearn.over_sampling import SMOTE
     from imblearn.under_sampling import RandomUnderSampler
     from sklearn.utils import resample
```

```
[42] # Submuestreo tradicional
     under_sampler = RandomUnderSampler(random_state=42)
     X_train_under, y_train_under = under_sampler.fit_resample(X_train, y_train)
```

```
[43] from sklearn.utils import resample

     # Sobremuestreo tradicional
     X_train_minority = X_train[y_train == 1]
     X_train_minority_upsampled = resample(X_train_minority,
                                           replace=True,
                                           n_samples=len(X_train[y_train == 0]),
                                           random_state=42)

     X_train_oversampled = pd.concat([X_train[y_train == 0], X_train_minority_upsampled])
     y_train_oversampled = y_train.loc[X_train_oversampled.index]
```

```
[44] # Smote
     smote = SMOTE(random_state=42)
     X_train_smote, y_train_smote = smote.fit_resample(X_train, y_train)
```

2. Función para la optimización de parámetros por búsqueda aleatoria con Py.Caret

```
metricas_unb = pd.DataFrame()
model_params_unb = {}

# Lista de modelos para afinar
models = ['lr', 'svm', 'dt', 'rf', 'xgboost', 'knn', 'nb']

for m in models:
    model = create_model(m)
    tuned = tune_model(model, n_iter=20, optimize='Recall', search_algorithm='random', early_stopping=True, choose_better=True)

    # Obtener y almacenar las métricas medias
    mean_metrics = pull().loc['Mean']
    metricas_unb = metricas_unb.append(mean_metrics, ignore_index=True)

    # Obtener y almacenar los hiperparámetros del modelo afinado
    model_params_unb[m] = tuned.get_params()
```

3. Función para la optimización de parámetros por grillas

```python
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

def apply_grid_search(X_train, y_train, X_test, y_test, pipeline, param_grid, cv=10, scoring_metric='f1'):

    # Aplicación del modelado

    grid_search = GridSearchCV(pipeline, param_grid, cv=cv, scoring=scoring_metric, verbose=1)
    grid_search.fit(X_train, y_train)

    best_params = grid_search.best_params_
    best_model = grid_search.best_estimator_

    # Predicciones
    y_pred = best_model.predict(X_test)

    # Calcular métricas de evaluación
    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred)
    recall = recall_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred)

    return {
        'best_params': best_params,
        'best_score': grid_search.best_score_,
        'accuracy': accuracy,
        'precision': precision,
        'recall': recall,
        'f1_score': f1
    }
```

4. Rejilla de hiperparámetros utilizados por modelo

Modelo	Parámetros de la Rejilla
Regresión Logística	'clf__C': [0.1, 1, 10], 'clf__penalty': ['l2'], 'clf__solver': ['newton-cg', 'lbfgs', 'liblinear']
Árbol de Decisión	'clf__max_depth': [None, 10, 20], 'clf__min_samples_split': [2, 5, 10], 'clf__min_samples_leaf': [1, 2, 4]
Bosque Aleatorio	'clf__n_estimators': [100, 200], 'clf__max_depth': [None, 10, 20], 'clf__min_samples_split': [2, 5, 10], 'clf__min_samples_leaf': [1, 2, 4]
XGBoost	'clf__n_estimators': [100, 200], 'clf__max_depth': [3, 6, 10], 'clf__learning_rate': [0.01, 0.1, 0.2], 'clf__subsample': [0.7, 0.9, 1]
k-Vecinos más Cercanos	'clf__n_neighbors': [3, 5, 11], 'clf__weights': ['uniform', 'distance'], 'clf__metric': ['euclidean', 'manhattan']
Naive Bayes	'clf__var_smoothing': [1e-9, 1e-8, 1e-7]

BIBLIOGRAFÍA

[1] M. n. d. n. 2025, «Organización Mundial de la Salud,» 30 Diciembre 2014. [En línea]. Available: https://www.who.int/es/publications/i/item/WHO-NMH-NHD-14.5. [Último acceso: 10 Noviembre 2023].

[2] Grupo Interinstitucional de las Naciones Unidas para la Estimación de la Mortalidad Infantil, Enero 2023. [En línea]. Available: https://data.unicef.org/resources/levels-and-trends-in-child-mortality/. [Último acceso: 10 Noviembre 2023].

[3] D. A. N. d. Estadistica, «Estadísticas Vitales (EEVV) Nacimientos en Colombia,» 21 Septiembre 2023. [En línea]. Available: https://www.dane.gov.co/files/operaciones/EEVV/bol-EEVV-Nacimientos-IItrim2023.pdf. [Último acceso: 10 Noviembre 2023].

[4] DANE, 21 Septiembre 2023. [En línea]. Available: https://www.dane.gov.co/files/operaciones/EEVV/bol-EEVV-Defunciones-IItrim2023.pdf. [Último acceso: 12 Noviembre 2023].

[5] «Determinantes y factores asociados con la tasa de Mortanidad Infantil: una comparación departamental y municipal,» DANE, 2021. [En línea]. Available: https://www.dane.gov.co/files/investigaciones/poblacion/informes-estadisticas-sociodemograficas/2021-09-23-Determinantes-factores-asociados-tasa-mortalidad-infantil-dptl-mpal.pdf. [Último acceso: 11 Noviembre 2022].

[6] M. S. Kramer, F. C. Barros, J. Kiley, S. Liu y K. S. Josephf, ¿La reduccion de la mortalidad infantil depende a la prevencion del bajo peso al nacer? Analisis de tendencias actuales en el conteniente americano, vol. 25, Revista del Hospital Materno Infantil Ramón Sardá, 2006, pp. 98-104.

[7] UNICEF, «Estado Mundial de la Infancia 2021. En mi mente: promover, proteger y cuidar la salud mental de la infancia".,» Octubre 2021. [En línea]. Available: https://www.unicef.org/es/informes/estado-mundial-de-la-infancia-2021. [Último acceso: 10 Noviembre 2023].

[8] OMS, «Mejorar la supervivencia y el bienestar de los recien nacidos. Nota Descriptiva.,» Septiembre 2020. [En línea]. Available: http://www.who.int/mediacentre/factsheets/fs333/es/ . [Último acceso: 28 Noviembre 2022].

[9] «Indicadores Básicos de salud 2023,» Ministerio de Salud y Protección Social, 2023. [En línea]. Available: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/ED/GCFI/indicadores-basicos-salud-2023.pdf. [Último acceso: 17 Noviembre 2023].

[10] «Asi vamos en Salud. Indicadores de Salud. Seguridad Alimentaria y Nutricional,» 10 Agoato 2023. [En línea]. Available: https://www.asivamosensalud.org/indicadores/seguridad-alimentaria-y-nutricional/prevalencia-de-bajo-peso-al-nacer. [Último acceso: 9 Noviembre 2023].

[11] I. N. d. Salud, «Boletín Epidemiológico Semanal BES,» 1-7 Agosto 2021. [En línea]. Available: https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2021_Boletin_epidemiologico_semana_31.pdf. [Último acceso: 12 Noviembre 2023].

[12] I. N. d. Salud, «Bajo peso al Nacer a Término. Periodo Epidemilógico IX, Colombia 2022.,» 26 Octubre 2022. [En línea]. Available: https://www.ins.gov.co/buscador-eventos/Informesdeevento/BAJO%20PESO%20AL%20NACER%20PE%20IX%202022.pdf#search=bajo%20peso%20al%20nacer. [Último acceso: 10 Noviembre 2023].

[13] A. S. d. Cali, «Bajo peso al nacer a término. Boletín Epidemiológico.,» 25 Febrero 2023. [En línea]. Available: https://www.cali.gov.co/salud/loader.php?lServicio=Tools2&lTipo=descargas&lFuncion=descargar&idFile=79061. [Último acceso: 17 Noviembre 2023].

[14] Y. M. Bonilla, «Un modelo logístico para la evolución de neonatos prematuros con bajo peso al nacer, atendidos en el hospital universitario del Valle, durante el periodo 2002 a 2010,» Universidad del Valle, Santiago de Cali, 2019.

[15] C. Castaño, L. S. Alvarez, B. Caicedo, I. C. Ruiz y S. Valencia-Aguirre, «Tendencia del bajo peso al nacer en recién nacidos a término y su relación con la pobreza y el desarrollo municipal en Colombia. 2000-2014,» *Revista Chilena de Nutrición,* vol. 47, nº 1, pp. 22-30, 2020.

[16] S. d. S. d. Cali, «Análisis de la situación integral de salud (ASIS),» 31 Diciembre 2021. [En línea]. Available: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/ED/PSP/asis-cali-distrito-2021.pd. [Último acceso: 1 Diciembre 2022].

[17] J. F. Mosnreal, M. R. Tun, J. R. Hernandez y L. E. Serralta, «Risk factors for low birth weight according to the multiple logistic regression model. A retrospective cohort study in José María Morelos municipality, Quintana Roo, Mexico,» 17 Enero 2018. [En línea]. Available: https://www.medwave.cl/medios/medwave/Enero-febrero2018/PDF/medwave-2018-01-7143.pdf. [Último acceso: 1 Diciembre 2022].

[18] L. S. Alvarez, «Los determinantes sociales de la salud: más allá de los factores de riesgo,» *Revista de Gerencia Politica de Salud,* vol. 8, nº 17, pp. 66-79, 2009.

[19] E. F. Quiroga, «Protocolo de vigilancia en salud pública: Bajo peso al nacer a término - evento 110 Version 3,» Instituto Nacional de Salud, 2020. [En línea]. Available: https://www.ins.gov.co/buscador-eventos/Lineamientos/Pro_Bajo%20peso%20al%20nacer.pdf. [Último acceso: 12 Noviembre 2022].

[20] FAO, FIDA, OPS, PMA y UNICEF, «America Latina y el Caribe. Panorama de la seguridad alimentaria y la nutrición. Estadisticas y tendencias,» 2023. [En línea]. Available: https://doi.org/10.4060/cc8514es. [Último acceso: 10 Octubre 2023].

[21] «Objetivos de desarrollo sostenible, Salud y Bienestar,» ONU, [En línea]. Available: https://www.un.org/sustainabledevelopment/es/health/. [Último acceso: 12 Noviembre 2023].

[22] C. A. Torres Ricaurte, «Bajo de peso al nacer y mortalidad infantil en Santiago de Cali, 2011-2014: Un análisis de factores sociodemográficos como herramienta para planificación del desarrollo,» 29 Noviembre 2017. [En línea]. Available: https://doi.org/10.57998/bdigital.handle.001.941. [Último acceso: 30 Junio 2023].

[23] «Encuesta Nacional de Demografía y Salud. Componente Demográfico. Tomo 1,» MINSALUD, Profamilia, 2015. [En línea]. Available: https://profamilia.org.co/docs/ENDS%20%20TOMO%20I.pdf. [Último acceso: 29 Junio 2023].

[24] A. Estrada, S. L. Restrepo, N. C. Ceballo y F. M. Santander, «Factores maternos relacionados con el peso al nacer de recién nacidos a término, Colombia, 2002-2011,» Noviembre 2016. [En línea]. Available: https://www.scielo.br/j/csp/a/FdHmLY3wjDzMZJhcTRQ5Rzc/?lang=es. [Último acceso: 29 Junio 2023].

[25] M. D. Giseselmann, «Educación, mortalidad infantil y bajo peso al nacer en Suecia 1973-1990: surgimiento de la paradoja del bajo peso al nacer. Vol.33, Num. 1,» Scandinavian Journal of Public Health, Enero 2005. [En línea]. Available: https://journals.sagepub.com/doi/10.1080/14034940410028352. [Último acceso: 15 Julio 2023].

[26] Centro de los Objetivos de Desarrollo Sostenible para América Latina y el Caribe (CODS), «Índice ODS 2021 para América Latina y el Caribe,» Agosto 2022. [En línea]. Available: https://cods.uniandes.edu.co/wp-content/uploads/2022/08/I%CC%81ndice-ODS-2021-para-Ame%CC%81rica-Latina-y-el-Caribe.pdf. [Último acceso: 17 Mayo 2023].

[27] OMS y OPS, «Política sobre la aplicación de la ciencia de datos en la salud pública mediante la inteligencia artificial y otras tecnologías emergentes,» 23 Septiembre 2021. [En línea]. Available: https://www.paho.org/es/file/93410/download?token=CYX06dv0. [Último acceso: 12 Agosto 2023].

[28] A. Gerón, Aprende Machine Learning con Scikit-Learn, Keras y TensorFlow. 2ª Edición, Madrid. España: Ediciones Anaya Multimedia, 2020, pp. 39-45.

[29] B. Mahesh, «Machine Learning Algorithms - A Review,» 2019. [En línea]. Available: https://www.ijsr.net/archive/v9i1/ART20203995.pdf. [Último acceso: 2 Noviembre 2023].

[30] J. A. Rodrigo, «Máquinas de Vector Soporte (Support Vector Machines, SVMs),» Abril 2017. [En línea]. Available: https://cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines. [Último acceso: 9 Noviembre 2023].

[31] F. Y. Osisanwo , J. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi y J. Akinjobi, «Supervised Machine Learning Algorithms: Classification and Comparison,» Seventh Sense Research Group, Junio 2017. [En línea]. Available: 10.14445/22312803/IJCTT-V48P126. [Último acceso: 9 Noviembre 2023].

[32] A. F. Alaminos-Fernandez, Árboles de decisión en R con Random Forest, Universidad de Alicante, España: OBETS – Ciencia Abierta Instituto de Desarrollo Social y Paz, 2023, p. 17.

[33] J. A. Rodrigo, «Árboles de decisión con Python: regresión y clasificación,» Octubre 2020. [En línea]. Available: https://cienciadedatos.net/documentos/py07_arboles_decision_python. [Último acceso: 17 Noviembre 2023].

[34] T. Chen y C. Guestrin, «XGBoost: A Scalable Tree Boosting System,» KDD '16: Actas de la 22ª Conferencia Internacional ACM SIGKDD sobre Descubrimiento de Conocimiento y Minería de Datos, 13 Agosto 2016. [En línea]. Available: https://dl.acm.org/doi/pdf/10.1145/2939672.2939785. [Último acceso: 16 Noviembre 2023].

[35] J. J. Espinoza-Zuñiga, «Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito,» Revista Ingenieria, Investigación y Tecnología, 28 Abril 2020. [En línea]. Available: https://doi.org/10.22201/fi.25940732e.2020.21.3.022. [Último acceso: 17 Noviembre 2023].

[36] G. Rebala, A. Ravi y S. Churiwala, An introduction to Machine Learnig, Switzerland: Sprinder Nature Switzerland, 2019.

[37] J. Torres, Python Deep Learning Introducción práctica con Keras y TensorFlow 2, Alemania: Marcombo, 2020.

[38] W. D. T. Y. L.-D. A. G. S. Ren Y, «Issue of Data Imbalance on Low Birthweight Baby Outcomes Prediction and Associated Risk Factors Identification: Establishment of Benchmarking Key Machine Learning Models With Data Rebalancing Strategies,» 31 Mayo 2023. [En línea]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10267797/. [Último acceso: 15 Septiembre 2023].

[39] Y. L. S. J. M. G. Y. H. G. B. Haixiang G, «Learning from class-imbalanced data: Review of methods and applications,» 1 Mayo 2017. [En línea]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0957417416307175. [Último acceso: 22 Septiembre 2023].

[40] M. C. Arrieta J, «Estudio Comparativo de Técnicas de Balanceo de Datos en el Aprendizaje de Múltiples Instancias,» 4 Septiembre 2015. [En línea]. Available: https://www.researchgate.net/publication/283642919_Estudio_Comparativo_de_Tecnicas_de_Balanceo_de_Datos_en_el_Aprendizaje_de_Multiples_Instancias. [Último acceso: 22 Octubre 2023].

[41] B. K. L. H. K. W. Chawla N, « SMOTE: Synthetic Minority Over-sampling Technique,» 2 Junio 2002. [En línea]. Available: https://www.researchgate.net/publication/220543125_SMOTE_Synthetic_Minority_Over-sampling_Technique. [Último acceso: 26 Octubre 2023].

[42] Z. J. W. W. Y. J. Pan T, «Learning imbalanced datasets based on SMOTE and Gaussian distribution,,» 21 Febrero 2020. [En línea]. Available: https://doi.org/10.1016/j.ins.2019.10.048.. [Último acceso: 23 Octubre 2023].

[43] I. Tomek, «Two Modifications of CNN,,» 1976. [En línea]. Available: http://dx.doi.org/10.1109/TSMC.1976.4309452. [Último acceso: 4 Noviembre 2023].

[44] A. Faruk, E. S. Cahyono, N. Eliyati y I. Arifieni, «Prediction and Classification of Low Birth Weight Data Using Machine Learning Techniques,» *Indonesian Journal of Science & Technology,* vol. 3, nº 1, pp. 18-22, 2018.

[45] N. Eliyati, A. Faruk, E. S. Kresnawati y I. Arifieni, «Support vector machines for classification of low birth weight in Indonesia,» de *Sriwijaya International Conference on Basic and Applied Science*, Palembang, Indonesia, 2018.

[46] H. Jeong, K. Min Moon y H. S. Jin, «Machine Learning Models for Predicting Mortality in 7472 Very Low Birth Weight Infants Using Data from a Nationwide Neonatal Network,» *Diagnostics,* vol. 12, nº 3, p. 625, 2022.

[47] W. Khan, N. Zaki, M. M. Masud y A. Ahmad, «Infant birth weight estimation and low birth weight classification in United Arab Emirates using machine learning algorithms,» 15 Julio 2022. [En línea]. Available: https://doi.org/10.1038/s41598-022-14393-6. [Último acceso: 8 Agosto 2023].

[48] M. Hajipour, N. Taherpour, H. Fateh, E. Yousefi, K. Etemad y et al, «Predictive Factors of Infant Mortality Using Data Mining in Iran,» J Comp Ped., 28 Febrero 2021. [En línea]. Available: https://doi.org/10.5812/compreped.108575.. [Último acceso: 30 Marzo 2023].

[49] D. Senthilkumar y S. Paulraj, «Prediction of Low Birth Weight Infants and Its Risk Factors Using Data Mining Techniques Senthilkumar,» 5 Marzo 2015. [En línea]. Available: https://ieomsociety.org/ieom_2015/papers/134.pdf. [Último acceso: 1 Noviembre 2023].

[50] E. Mfateneza, P. C. Rutayisiri, E. Biracyaza, S. Musafiri y W. G. Mpabuka, «Application of machine learning methods for predicting infant mortality in Rwanda: analysis of Rwanda demographic health survey 2014–15 dataset,» 4 Mayo 2022. [En línea]. Available: https://doi.org/10.1186/s12884-022-04699-8. [Último acceso: 27 Mayo 2023].

[51] S. J. Sawe, Machine Learning Prediction of Low Birth Weight in Kenya using Maternal Risk Factors", Rwanda, Africa: University of Rwanda, College of Business and Economics, 2022.

[52] Ministerio de Salud y Protección Social, «Manual del Usuario Modilo de Nacimientos y Defunciones Aplicativo Web Nacimientos y defunciones RUAF-ND,» Julio 2022. [En línea]. Available: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/ED/GCFI/manual-usuario-ruafnd-v2-jul2022.pdf. [Último acceso: 23 Noviembre 2023].

[53] A. Zabala-Garcia, H. Ortiz-Reyes, J. Salomon-Kuri, C. Padilla-Amigo y R. Preciado-Ruiz, «Periodo intergenésico: Revisión de la literatura,» Revista chilena de obstetricia y ginecología, Febrero 2018. [En línea]. Available: https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0717-75262018000100052. [Último acceso: 30 Noviembre 2023].

[54] E. R. Lara, «Modelos de Clasificación con datos no balanceados,» Junio 2018. [En línea]. Available: https://idus.us.es/bitstream/handle/11441/77518/Espinar%20Lara%20Roc%C3%ADo%20TFG.pdf?sequence=1&isAllowed=y. [Último acceso: 6 Diciembre 2023].

[55] A. P. Becerra, «Desbalanceo de datos en redes de clasificación binaria,» Septiembre 2021. [En línea]. Available: https://rua.ua.es/dspace/bitstream/10045/118112/1/Desbalanceo_de_datos_en_redes_de_clasificacion_binaria_Amoros_Becerra_Pablo.pdf. [Último acceso: 6 Diciembre 2023].

[56] S. E. Jimenez, Y. Hernandez y H. J. Ortiz, «Técnicas de Optimización de Hiperparámetros en Modelos de Aprendizaje Automático para Predicción de Enfermedades Cardiovasculares,» Noviembre 2022. [En línea]. Available: https://www.researchgate.net/publication/365360062_Tecnicas_de_Optimizacion_de_Hiperparametros_en_Modelos_de_Aprendizaje_Automatico_para_Prediccion_de_Enfermedades_Cardiovasculares. [Último acceso: 6 Diciembre 2023].