# Desarrollo de una Metodología para la Propuesta de un Modelo Predictivo de Morbilidad Materna Extrema (MME) en Mujeres Embarazadas de los Departamentos de Bolívar, Córdoba, Magdalena, Sucre y Atlántico afiliadas a la EPS MUTUAL SER ESS, Mediante Técnicas de Machine Learning.

Elaborado por:
Oscar Ivan Echeverria Marrungo
Javier Callejas Cardozo
Jeshua David Junca Rojas

Universidad Ean
Escuela de Formación en Investigación
Seminario de Investigación de Pregrado
Elizabeth León Velásquez
Bogotá

10/06/2025

## Introducción

El presente proyecto desarrolla una propuesta metodológica para la construcción de un modelo predictivo de Morbilidad Materna Extrema (MME) en mujeres embarazadas afiliadas a la EPS MUTUAL SER ESS en los departamentos de Bolívar, Córdoba, Magdalena, Sucre y Atlántico, utilizando técnicas de Machine Learning. La investigación parte del análisis del contexto nacional y regional de la MME, identificando las principales problemáticas asociadas a la atención materna.

El capítulo 1 ofrece un resumen general del estudio. El capítulo 2 expone el problema de investigación, incluyendo los antecedentes globales, nacionales y regionales de la Morbilidad Materna Extrema, con énfasis en el contexto de la EPS MUTUAL SER ESS en los departamentos de Bolívar, Córdoba, Magdalena, Sucre y Atlántico. El capítulo 3 formula la pregunta de investigación. El capítulo 4 presenta el objetivo general, los objetivos específicos y la justificación del estudio. El capítulo 5 introduce los fundamentos conceptuales. El capítulo 6 desarrolla el marco teórico, abordando las variables clínicas, sociodemográficas y de atención médica, así como las técnicas y procesos de evaluación de modelos de Machine Learning. El capítulo 7 describe la metodología del estudio, detallando el enfoque, diseño, población, muestra, operacionalización de variables y técnicas de análisis de datos. El capítulo 8 presenta el análisis y discusión de los resultados, incluyendo la exploración de variables, correlaciones, tratamiento del desbalance, selección de características y revisión del estado del arte. Los capítulos 9 y 10 recogen, respectivamente, las recomendaciones y las lecciones aprendidas derivadas del proceso investigativo, seguidos por las conclusiones generales del estudio.

**Contenido**

## Figuras

**Tablas**

## 1. Resumen

Este estudio propone una metodología para el desarrollo de un modelo predictivo basado en machine learning con el fin de anticipar el riesgo de Morbilidad Materna Extrema (MME) en mujeres embarazadas afiliadas a la EPS MUTUAL SER ESS en cinco departamentos del Caribe colombiano. La MME representa un desafío crítico de salud pública asociado a desigualdades territoriales y deficiencias en la atención prenatal.

A partir de 88.810 registros clínicos y sociodemográficos, se aplicó una metodología integral que incluyó análisis de calidad de datos, selección de variables mediante técnicas estadísticas y algoritmos de aprendizaje automático (regresión logística y random forest), tratamiento del desbalance de clases (19:1) y validación geográfica. Se identificaron nueve variables predictoras clave, mientras que se excluyó la variable "documento" por su efecto espurio.

Los resultados muestran que es posible construir modelos predictivos robustos y contextualizados que apoyen la toma de decisiones clínicas y de salud pública. Esta metodología permite transitar desde un enfoque reactivo hacia uno preventivo, alineado con los Objetivos de Desarrollo Sostenible (ODS) y la reducción de la mortalidad materna evitable.

**Palabras clave:** Morbilidad Materna Extrema, Machine Learning, modelo predictivo, salud pública, ODS.

# 2. Problema de Investigación

Las complicaciones relacionadas con la gestación continúan siendo un problema de salud pública en Colombia. Entre ellas, la Morbilidad Materna Extrema (MME) se destaca como una condición en la que una mujer enfrenta complicaciones graves durante el embarazo, el parto o hasta 42 días después de la terminación del embarazo, poniendo en riesgo su vida y la del bebé, aunque logra sobrevivir a la situación crítica (Narváez Díaz & Caro Caro, 2024).

En 2021, Colombia registró 30.102 casos de MME, un aumento del 23,1 % en comparación con 2020, según el Sistema de Vigilancia en Salud Pública (SIVIGILA) (Instituto Nacional de Salud, 2021). En el departamento de Bolívar (departamento con mayor población representativa de la EPS analizada), la situación es preocupante debido a desigualdades en el acceso a la salud y deficiencias en la detección temprana de factores de riesgo. En 2020, se reportaron 585 casos de MME en esta región, con un saldo de 33 muertes perinatales (Instituto Nacional de Salud, 2020). Durante 2024, la EPS MUTUAL SER ESS reportó 827 casos, con Cartagena concentrando el 28 % de estos eventos (Instituto Nacional de Salud, 2024).

## 2.1. Antecedentes del Problema

La MME afecta a millones de mujeres en el mundo. Según la Organización Mundial de la Salud (OMS), cada año cerca de 295,000 mujeres mueren debido a complicaciones en el embarazo y el parto, y por cada muerte materna, aproximadamente 30 mujeres experimentan MME, lo que equivale a más de 8.8 millones de casos anuales, muchos de ellos prevenibles (OMS, 2020).

### 2.1.1. *Contexto Global y Nacional de la MME*

En América Latina, la MME es un indicador preocupante de la calidad de los servicios de salud materna. Según la Comisión Económica para América Latina y el Caribe (CEPAL), las tasas de MME y mortalidad materna están correlacionadas con inequidades sociales, acceso

desigual a servicios de salud y deficiencias en los sistemas de información (CEPAL, 2021). En Colombia, a pesar de los avances en cobertura de salud, persisten brechas en la detección temprana y el manejo adecuado de complicaciones obstétricas.

El Instituto Nacional de Salud (INS) reportó que en 2021 se registraron 30,102 casos de MME en el país. Las principales causas fueron la hemorragia obstétrica severa (37.2 %), los trastornos hipertensivos del embarazo (30.8 %), la sepsis materna (15.5 %) y otras complicaciones graves (16.5 %) (INS, 2022).

### 2.1.2. *Entorno Sectorial: La MME en los departamentos de Bolívar, Córdoba, Magdalena Sucre y Atlántico, y la EPS MUTUAL SER ESS*

La Región Caribe, donde se concentra el mayor número de afiliados a la EPS MUTUAL SER ESS, presenta altas tasas de Morbilidad Materna Extrema (MME), lo que refleja desigualdades en la atención materna y dificultades en la detección temprana de riesgos. En el año 2020 se reportaron 585 casos, con 33 muertes perinatales (INS, 2020). Al corte del 24 de mayo de 2025, la EPS MUTUAL SER ESS había identificado un total de 744 casos, de los cuales el 36 % correspondían al departamento de Bolívar, el 20 % a Córdoba, el 8,74 % a Sucre, el 8,20 % a Magdalena y el 20 % al Atlántico (INS, 2025).

Entre las principales causas de la alta incidencia de MME en los departamentos mencionados afiliados a MUTUAL SER ESS, se destacan:

#### 2.1.2.1. *Brechas en la calidad de la atención prenatal*

- Baja cobertura de controles prenatales efectivos: muchas mujeres no completan los ocho controles recomendados por la OMS (OMS, 2019).
- Falta de capacitación en emergencias obstétricas: la MME disminuye cuando el personal de salud recibe formación especializada (Bauserman et al., 2021).

2.1.2.2. *Deficiencias en registros clínicos y acceso a datos*

- Fragmentación de la información: la falta de interoperabilidad de los sistemas de salud en Colombia dificulta la toma de decisiones basada en datos (MinSalud, 2023).
- Subregistro de eventos críticos: muchos casos de MME no se documentan correctamente, limitando el análisis predictivo y la implementación de alertas tempranas (Muñoz et al., 2013).

2.1.2.3. *Factores sociodemográficos y desigualdades regionales*

- Barreras geográficas y económicas: el acceso limitado a hospitales de alta complejidad aumenta el riesgo de desenlaces adversos (INS, 2023).
- Enfermedades preexistentes no controladas: afecciones como diabetes gestacional, hipertensión y anemia severa incrementan el riesgo de complicaciones maternas y perinatales (Narváez Díaz & Caro Caro, 2024).

2.1.2.4. Impacto de la Tecnología en la Predicción de la MME

El uso de tecnologías avanzadas como la inteligencia artificial (IA) y el aprendizaje automático (ML) ha demostrado ser eficaz en la predicción de complicaciones obstétricas. Estudios recientes muestran que modelos de ML han alcanzado precisiones superiores al 85% en la predicción de preeclampsia y hemorragia posparto (Say et al., 2020; Souza et al., 2021).

A pesar de estos avances, en Colombia y particularmente en los departamentos de Bolívar, Córdoba, Magdalena Sucre y Atlántico, no existen modelos de IA diseñados específicamente para la detección temprana de MME. Su implementación en la EPS MUTUAL SER ESS podría optimizar la toma de decisiones clínicas y reducir la carga de MME en la región.

## 2.2. Descripción del Problema

La MME representa un problema de salud pública con un impacto significativo en la calidad y seguridad de la atención materna. Las complicaciones graves durante el embarazo, parto y puerperio siguen siendo frecuentes, especialmente en regiones con acceso limitado a servicios de salud adecuados.

Uno de los principales desafíos es la capacidad de predecir con precisión qué mujeres tienen mayor riesgo de desarrollar MME. Los métodos tradicionales de evaluación del riesgo suelen ser imprecisos, por lo que el uso de IA y ML representa una oportunidad para mejorar la detección temprana.

Actualmente, la implementación de estos modelos predictivos en Colombia es limitada debido a la disponibilidad y calidad de los datos y la falta de integración con los sistemas de salud. Esta investigación busca proponer una metodología para la implementación de un modelo predictivo de MME para optimizar la atención materna en los departamentos de Bolívar, Córdoba, Magdalena Sucre y Atlántico las afiliadas de la EPS MUTUAL SER ESS.

# 3. Pregunta de Investigación

Teniendo en cuenta lo expuesto anteriormente es primordial preguntar*: ¿Es posible predecir el riesgo de Morbilidad Materna Extrema en mujeres embarazadas afiliadas a la EPS MUTUAL SER ESS, ubicadas en los departamentos de Bolívar, Córdoba, Magdalena, Sucre y Atlántico (Colombia), mediante el uso de algoritmos de Machine Learning, a partir de datos clínicos y demográficos disponibles, y adaptando una metodología adecuada para el desarrollo de modelos predictivos?*

# 4. Objetivos

### 4.1. Objetivo general

Diseñar una metodología para la propuesta de un modelo predictivo basado en técnicas de Machine Learning para prevenir la Morbilidad Materna Extrema en mujeres afiliadas a la EPS MUTUAL SER ESS en los departamentos de Bolívar, Córdoba, Magdalena Sucre y Atlántico.

### 4.2. Objetivos específicos

1. Analizar los datos clínicos y sociodemográficos relevantes para el análisis de Morbilidad Materna Extrema de la población objeto de estudio, garantizando su calidad y consistencia.
2. Identificar patrones en los datos mediante análisis estadístico y técnicas de Machine Learning, seleccionando las variables más relevantes para la predicción del riesgo de Morbilidad Materna Extrema.
3. Proponer un enfoque metodológico para el desarrollo de un modelo predictivo utilizando técnicas de Machine Learning, para la predicción del riesgo de la Morbilidad Materna Extrema.

## 5. Justificación

Existe un creciente interés en el análisis de la Morbilidad Materna Extrema (MME) como un indicador clave para medir la calidad del cuidado materno. Se reconoce ampliamente que la vigilancia epidemiológica constante de la MME es una estrategia fundamental para reducir la tasa de mortalidad materna. Por ello, es esencial fortalecer los sistemas de vigilancia e información tanto en la región latinoamericana como en Colombia (Ortiz Lizcano, Quintero Jaramillo, Mejía López, Romero Vélez, & Ospino Rodríguez, 2010)

La reducción de la MME está directamente relacionada con el Objetivo de Desarrollo Sostenible (ODS) número 3 de las Naciones Unidas, que busca "Garantizar una vida sana y promover el bienestar para todas las edades". En particular, la meta 3.1 de este objetivo plantea

reducir la tasa mundial de mortalidad materna a menos de 70 por cada 100.000 nacidos vivos para el año 2030. (Organización Mundial de la Salud, 2023); sin embargo, a pesar de los avances en este campo, no se ha logrado disminuir de manera significativa la incidencia de casos recurrentes de MME y mortalidad materna. En este sentido, la creación o implementación de una herramienta para la identificación temprana de riesgos permitiría detectar oportunamente a las pacientes en situación de vulnerabilidad. Esto facilitaría el seguimiento y análisis de cada caso, evitando el aumento de complicaciones que deterioran los indicadores de salud materna, los cuales siguen siendo elevados a nivel regional, especialmente en la región Caribe y en el contexto nacional (Rodríguez, 2017).

El desarrollo de un modelo predictivo para la identificación temprana del riesgo de MME en mujeres embarazadas de los departamentos mencionados contribuirá a optimizar la atención médica y a reducir las complicaciones materno-perinatales. Este proyecto beneficiará a la EPS MUTUAL SER ESS al proporcionar una herramienta de apoyo basada en datos, que mejorará la detección temprana de riesgos y facilitará intervenciones oportunas, a través de una metodología adaptada con técnicas de Machine Learning. De esta manera, se favorecerá una mejor planificación y asignación de recursos en los servicios de salud.

La investigación enfocada en la línea de investigación en ingeniería en específico la ciencia de datos aplica técnicas avanzadas de análisis y modelado de datos para identificar patrones de riesgo con mayor precisión, garantizando eficiencia en su implementación. Su valor metodológico radica en la propuesta de un modelo interpretable y adaptable a entornos clínicos reales. En términos prácticos, su integración fortalecerá la gestión y optimización de los procesos en salud, promoviendo estrategias preventivas sostenibles y una mejor administración de los recursos disponibles.

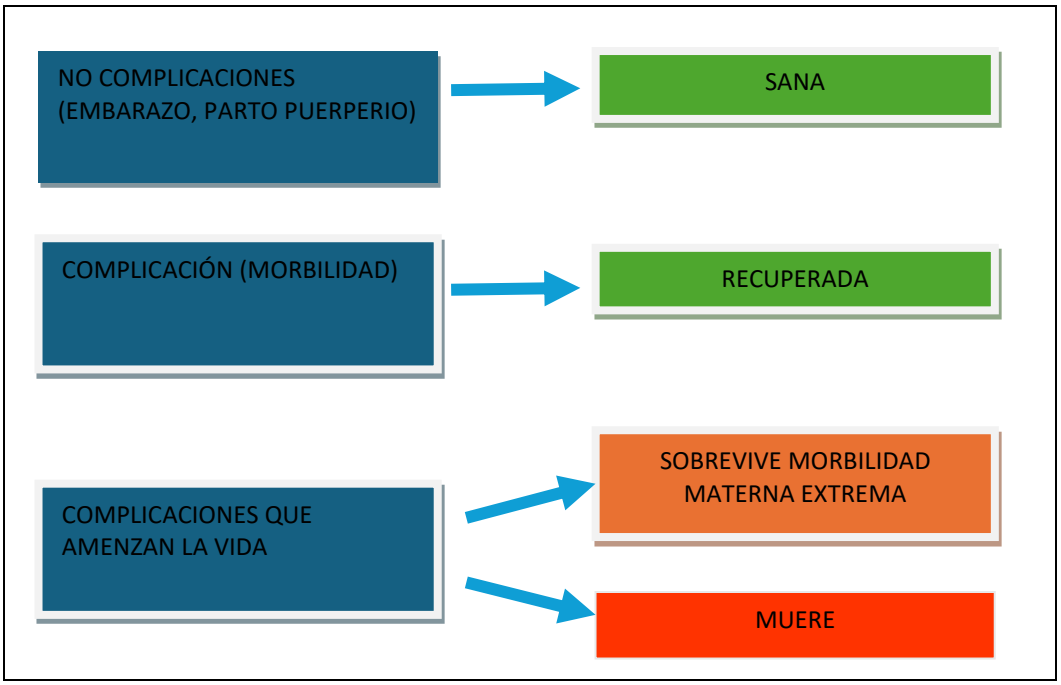
## 6. Marco Teórico

Para comprender la MME, es fundamental entender que, durante el embarazo, el proceso de salud y enfermedad se desarrolla a lo largo de una secuencia de eventos que van desde la salud

plena hasta la muerte. Dentro de esta secuencia, un embarazo puede clasificarse como no complicado, complicado (morbilidad), severamente complicado (morbilidad severa) o como una condición que amenaza la vida de la gestante y del bebé. En este último caso, las mujeres pueden recuperarse, desarrollar una incapacidad temporal o permanente, o, en el peor de los casos, fallecer. Aquellas gestantes que sobreviven a una complicación que ponía en riesgo su vida son consideradas casos de MME. (Ortiz Lizcano , Quintero Jaramillo, Mejía Loéz, Romero Vélez, & ospino Rodriguez, 2010) a continuación en el Figura 1, se muestra la secuencia entre los extremos de salud y muerte durante la gestación, permitiendo así identificar que espectro se encuentra la MME.

**Figura 1**

*Secuencia entre los extremos de salud y muerte durante el embarazo*




*Fuente:* Adaptado de Ortiz Lizcano , Quintero Jaramillo, Mejía Loéz, Romero Vélez, & ospino Rodriguez (2010)

*Nota*: Grafico que explica los extremos de condiciones para identificar que en MME.

Teniendo en cuenta lo anterior, y reconociendo que esta condición ocurre con mayor frecuencia que la muerte materna, su estudio permite analizar un número más amplio de casos y

facilita una cuantificación más precisa de los factores de riesgo. (Ortiz Lizcano, Quintero Jaramillo, Mejía Loéz, Romero Vélez, & Ospino Rodríguez, 2010*)*. Estas complicaciones pueden estar relacionadas con:

- Hemorragias severas
- Trastornos hipertensivos del embarazo
- Infecciones sistemáticas o patológicas agravadas por la gestación. (Ministerio de Salud y Protección Social de Colombia, 2021).

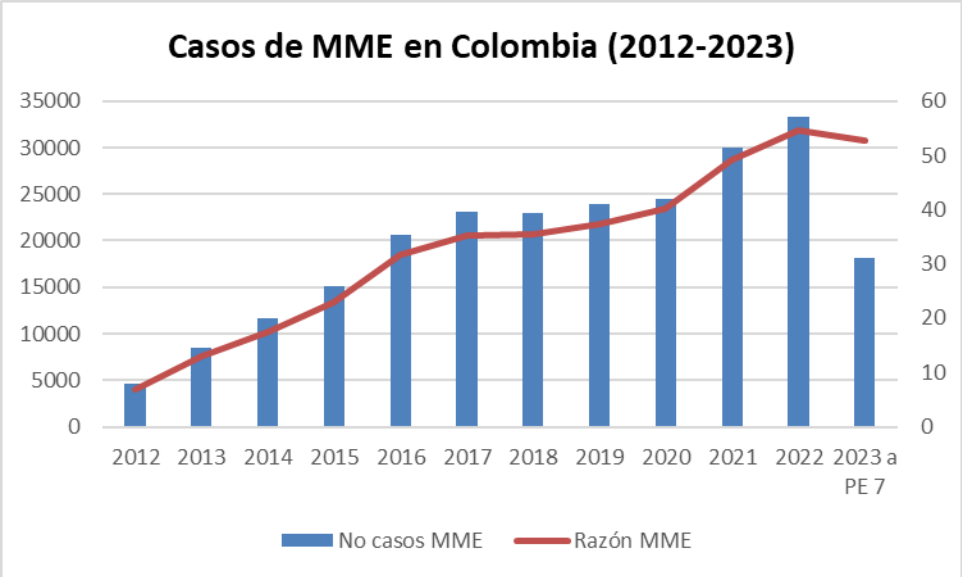
En el análisis de salud materna y sus posibles afectaciones requiere considerar diversos factores, incluyendo:

- Factores biológicos y clínicos: Hipertensión, diabetes gestacional, antecedentes obstétricos.
- Factores sociales y económicos: Nivel educativo, acceso a servicios de salud, apoyo familiar.
- Factores geográficos: Distancia a centros de salud y calidad de la atención disponible.

En el contexto en específico de Colombia, la MME ha sido identificada como un indicador de la calidad de atención materna y neonatal entre el 2012 al 2023 se evidenciaron 218,263 casos identificados a nivel nacional (Instituto Nacional de Salud, 2024)véase la siguiente Figura 2 relaciona la a razón de tendencia y la variación anual de los casos registrados por MME, la tendencia se mantiene en alza durante la década analizada, este incremento mantenido entre los años relacionados y subsiguientes sugiere la necesidad de mejorar los procesos de detención temprana y atención oportuna.

**Figura 2**

*Tendencia de la razón de MME, según notificación al Sivigila, Colombia, 2012 a periodo epidemiológico VII de 2023*[1].




*Fuente:* Figura adaptada del Instituto Nacional de Salud (2023).

Ahora bien, en lo que respecta a la presente investigación, es importante traer a colación las cifras de los departamentos de Bolívar, Córdoba, Magdalena, Atlántico y Sucre que presenta una de las tasas más altas de MME en Colombia, según el INS (2020). en el 2020 se reportaron entre estos 5 departamento un total de 2619 casos confirmados de MME, representando un 10% de los casos de tal año. Esta situación se ve agravada por factores como el acceso desigual a los servicios de salud, la ruralidad y las condiciones socioeconómicas de la población. Cartagena, como ciudad principal del departamento de Bolívar cuto más casos se encuentran, concentra el 28 % de los casos de MME registrados en la región.

[1] Tenga en cuenta que en Colombia la vigilancia de la MME inició en 2012. La razón se calcula por cada 1.000 nacidos vivos. Además, considere que la semana epidemiológica 7 corresponde al período del 16 al 22 de julio de 2023.

Dado que esta investigación se enfocará en la población afiliada a la EPS[2] MUTUAL SER ESS [3], específicamente en los departamentos de Bolívar, Córdoba, Magdalena Sucre y Atlántico, es relevante destacar que esta EPS es una de las principales aseguradoras del régimen subsidiado en la región Caribe.

Algunos de los principales retos en la atención materna en los departamentos de Bolívar, Córdoba, Magdalena Sucre y Atlántico y en MUTUAL SER ESS incluyen:

- Limitaciones en la infraestructura hospitalaria: En muchas zonas rurales, las gestantes deben trasladarse largas distancias para acceder a servicios especializados.
- Baja educación en salud materna: Muchas pacientes desconocen los síntomas de riesgo, lo que retrasa la búsqueda de atención médica. Esto está vinculado a la carencia de información sobre complicaciones del embarazo y a un débil acceso a la educación en salud sexual y reproductiva.
- Deficiencias en la calidad de la atención: Fallas en la infraestructura, falta de capacitación del personal y deficiencias en los protocolos de atención comprometen la detección y el manejo oportuno de la MME (Muñoz, Rojas & Torrez Villa, 2013).
- Deficiencias en los registros clínicos: La falta de datos estructurados dificulta el análisis epidemiológico y la toma de decisiones basada en evidencia (Martínez & Pérez, 2021).

## 6.1. Variables clave para identificar las condiciones de MME en específico en los departamentos de Bolívar, Córdoba, Magdalena Sucre y Atlántico y las afiliadas a la EPS MUTUAL SER ESS

Para entender realmente cómo se presenta la MME en la región, es necesario mirar más allá de los números y acercarnos a la realidad que viven las mujeres embarazadas. Identificar

[2] Empresa prestadora de salud
[3] Mutual Ser EPS es una entidad prestadora de salud colombiana en operación desde 1996. Actualmente, cuenta con más de 4 millones de afiliados a nivel nacional, con una mayor presencia en ciudades como Cartagena, Barranquilla, Montería y Sincelejo. Su mayor representación se encuentra en el departamento de Bolívar.

claramente quiénes están en mayor riesgo implica reconocer aspectos personales, clínicos y del entorno donde ellas viven y se atienden.

### 6.1.1. *Variables sociodemográficas*

- **Edad de la madre:** Es sabido que las mujeres muy jóvenes (menores de 18 años) y aquellas mayores de 35 enfrentan mayores riesgos durante el embarazo.
- **Nivel educativo:** La educación es clave porque influye directamente en qué tan rápido la madre reconoce síntomas alarmantes y busca ayuda.
- **Lugar donde viven (urbano o rural):** No es lo mismo estar en Cartagena, con hospitales cerca, que, en una vereda apartada de Bolívar, donde el acceso a un especialista puede ser complicado.
- **Situación familiar y estado civil:** El apoyo de la familia o pareja también es crucial para que busquen atención médica de forma rápida.
- **Nivel socioeconómico:** Desafortunadamente, lo económico todavía decide qué tan fácil es llegar a un médico o acceder a tratamientos preventivos oportunamente.

### 6.1.2. *Variables clínicas, que son la historia y el presente de salud de cada madre*

- **Antecedentes obstétricos:**
  - Cuántos embarazos previos ha tenido.
  - Si tuvo abortos o partos prematuros antes.
  - Si ha sufrido hemorragias severas o preeclampsia anteriormente.
- **Enfermedades que ya tenía antes del embarazo:**
  - Presión alta crónica.
  - Diabetes (ya sea previa o gestacional).
  - Anemia severa.
  - Obesidad o sobrepeso significativo.
- **Condiciones durante el embarazo actual:**
  - Preeclampsia o hipertensión gestacional.

- Infecciones urinarias o vaginales recurrentes sin tratamiento oportuno.
- Problemas detectados en exámenes de sangre, como niveles bajos de hemoglobina.
- Problemas en el crecimiento del bebé o detección temprana de alguna complicación fetal.

### 6.1.3. *Variables e la atención médica que reciben las gestantes*

- Controles prenatales: no es lo mismo asistir puntualmente a los 8 controles recomendados por la OMS, que hacerlo de forma incompleta o esporádica.
- Acceso rápido a especialistas y centros de alta complejidad: clave para que cualquier complicación pueda atenderse a tiempo.
- Tiempo entre los síntomas y la atención médica recibida: vital para prevenir complicaciones graves.
- Calidad de los registros médicos: historias clínicas organizadas y completas son fundamentales para tomar buenas decisiones.
- Preparación del personal de salud: el entrenamiento y la capacitación continua hacen la diferencia entre detectar rápido una emergencia o que se complique aún más.

Reconocer y analizar estas variables desde la realidad colombiana y específicamente desde lo que viven día a día las madres en los departamentos de Bolívar, Córdoba, Magdalena Sucre y Atlántico y afiliadas a la EPS MUTUAL SER ESS permitirá crear un modelo predictivo útil, cercano y realmente capaz de salvar vidas

A continuación, se presentan algunas definiciones y marcos de referencia para comprender cómo construir un modelo predictivo utilizando técnicas de ML. Se tendrá en cuenta lo mencionado anteriormente, así como las categorías de datos necesarias para desarrollar una metodología clave que desencadene en un sistema fiable y confiable, capaz de generar predicciones precisas a partir de los datos disponibles de la data histórica de MME para la población objetivo de la presente investigación.

### *6.2. Técnicas de ML para modelos predictivos*

Para armar un modelo predictivo que funcione en nuestro contexto colombiano y especialmente en los departamentos mencionados objeto de estudio, podemos considerar algunas técnicas de ML, cada una con sus ventajas particulares:

- **Árboles de Decisión:** Son intuitivos y visuales, parecidos a esas preguntas que nos hace un médico experimentado para tomar decisiones rápidas y claras.
- **Random Forest (Bosques Aleatorios):** Como cuando consultamos varias opiniones antes de tomar una decisión importante; aquí, varios árboles trabajan juntos para tomar decisiones más robustas y precisas.
- **Regresión Logística:** Es sencilla pero poderosa, ideal para identificar claramente quién tiene más riesgo de desarrollar MME con base en sus características personales y clínicas.
- **Support Vector Machines (Máquinas de vectores de soporte - SVM):** Perfectas para datos complejos; algo así como separar claramente quién tiene alto riesgo y quién no, aunque la diferencia sea pequeña y sutil.
- **Gradient Boosting (XGBoost o LightGBM):** Técnicas más modernas y sofisticadas, son como esos jugadores que marcan la diferencia en un partido difícil, permitiendo predicciones rápidas y muy precisas.
- **Una red neuronal multinivel:** Es un modelo de inteligencia artificial con varias capas de neuronas que permite aprender patrones complejos en datos, útil para tareas como clasificación y predicción.
- **Un modelo aditivo generalizado (GAM):** es una extensión flexible de los modelos lineales que permite capturar relaciones no lineales entre las variables independientes y la variable dependiente, usando funciones suaves (como splines), pero manteniendo la interpretabilidad del modelo.

### 6.3. Optimización y evaluación de modelos ML predictivos

### 6.3.1. *Optimización*

La optimización de datos mediante aprendizaje automático (ML) es un proceso clave para garantizar la calidad y utilidad de la información, aspectos fundamentales en la toma de decisiones informadas. Este enfoque mejora la eficiencia y precisión de los datos en contextos analíticos y predictivos.

La optimización de datos mediante aprendizaje automático (ML)4 es un proceso clave para garantizar la calidad y utilidad de la información, aspectos fundamentales en la toma de decisiones informadas. Este enfoque mejora la eficiencia y precisión de los datos en contextos analíticos y predictivos.

La optimización de datos con ML se basa en metodologías computacionales que procesan, transforman y analizan la información para reducir su complejidad. Según Provost y Fawcett (2013), sus principales objetivos son:

- Mejorar la calidad de los datos.
- Reducir la dimensionalidad.
- Optimizar el rendimiento de los modelos predictivos.
- Automatizar tareas repetitivas.

Uno de los aspectos más relevantes en este proceso es el ajuste de hiperparámetros5 y la arquitectura del modelo para mejorar su desempeño. Entre las técnicas más utilizadas destacan:

[4] El aprendizaje automático busca desarrollar técnicas que permitan a las computadoras aprender y generalizar comportamientos a partir de la información suministrada. Tiene diversas aplicaciones, como motores de búsqueda, diagnósticos médicos y detección de fraude. Se relaciona estrechamente con la estadística computacional y se clasifica en dos tipos: supervisado (deduce una función a partir de datos de entrenamiento, los cuales incluyen un conjunto de entrada y sus respectivos resultados esperados.) y no supervisado (No cuenta con categorías de respuesta predefinidas y emplea técnicas de agrupamiento para identificar patrones y clases en los datos). (Rodríguez, 2017)

[5] Los hiperparámetros son variables establecidas antes del entrenamiento de un modelo de aprendizaje automático (ML) y afectan directamente su rendimiento. A diferencia de los parámetros, que se ajustan automáticamente durante el proceso de aprendizaje, los hiperparámetros se configuran manualmente antes de entrenar el modelo. Algunos ejemplos incluyen el número de nodos y capas en una red neuronal o la cantidad de ramificaciones en un árbol de decisión. Estos valores determinan aspectos clave como la arquitectura del modelo, la tasa de aprendizaje y su complejidad. (AWS, 2024)

6.3.1.1. *Búsqueda de hiperparámetros*

- **Búsqueda en cuadrícula (grid search):** A través de este método, se define una lista de hiperparámetros y una métrica de rendimiento. Luego, el algoritmo evalúa todas las combinaciones posibles para determinar la opción más adecuada. Es importante considerar que este proceso puede ser arduo y repetitivo, especialmente en sistemas complejos con un número elevado de hiperparámetros. (AWS, 2024)
- **Búsqueda aleatoria (random search):** Explora diferentes combinaciones de hiperparámetros dentro de un espacio definido, permitiendo encontrar configuraciones óptimas de manera más eficiente, funciona bien cuando un número relativamente pequeño de hiperparámetros. (Bergstra & Bengio, 2012).

6.3.1.2. Optimización bayesiana

En los últimos años, esta técnica ha ganado popularidad por su eficiencia en la exploración del espacio de hiperparámetros. Se basa en modelos probabilísticos fundamentados en el teorema de Bayes, el cual describe la probabilidad de que ocurra un evento según el conocimiento disponible. En el contexto de la optimización de hiperparámetros, el algoritmo construye un modelo probabilístico a partir de un conjunto de hiperparámetros y una métrica específica. Además, emplea análisis de regresión para seleccionar, de manera iterativa, el conjunto óptimo (AWS, 2024).

6.3.1.3. Regularización:

La regularización es un conjunto de métodos diseñados para reducir el sobreajuste en modelos de aprendizaje automático de ML. En la mayoría de los casos, compensa una ligera disminución en la precisión del entrenamiento con una mejor capacidad de generalización, es decir, la habilidad del modelo para hacer predicciones precisas en nuevos conjuntos de datos. Sin embargo, este enfoque implica un mayor error durante el entrenamiento, lo que significa que las

predicciones en dicho proceso pueden ser menos precisas, pero más confiables en los datos de prueba (IBM, 2023).

6.3.1.4. Análisis Exploratorio de Datos (EDA):

El Análisis Exploratorio de Datos (EDA), por sus siglas en inglés: Exploratory Data Analysis) Su objetivo principal es entender los datos antes de construir un modelo predictivo.

Es el proceso de:

- Examinar, resumir y visualizar los datos.
- Detectar patrones, relaciones, anomalías y errores.
- Formular hipótesis y obtener una comprensión más profunda del conjunto de datos.

6.3.1.5. Análisis de Correlación:

El análisis de correlación es una técnica estadística que se utiliza para medir y entender la relación entre dos o más variables numéricas.

Mide:

- La fuerza de la relación entre dos variables.
- La dirección de esa relación (positiva o negativa).

6.3.1.6. Selección de características (Feature Selection):

La selección de características es un proceso clave en el aprendizaje automático que permite reducir el número de variables de entrada, seleccionando aquellas más relevantes para mejorar la eficiencia y precisión del modelo. Este proceso se realiza antes del entrenamiento y puede incluir ingeniería de características para modificar los datos. Algunos algoritmos automatizan la

selección de características, mientras que en otros se pueden ajustar parámetros manualmente. Se asigna una puntuación a cada atributo, y solo aquellos con las mejores puntuaciones son utilizados en el modelo. Si una característica no cumple con el umbral de selección, puede emplearse en la predicción, pero su influencia se basará únicamente en estadísticas globales.

Entre los métodos dentro de esta métrica de selección de variables que se tomarán en cuenta en esta investigación serán:

- Análisis Chi Cuadrado: Prueba estadística que permite evaluar si existe una asociación significativa entre dos o más variables categóricas. Su uso principal se da en tablas de contingencia, donde se comparan las frecuencias observadas con las esperadas bajo la hipótesis nula de que no existe asociación entre las variables.

  Formula estadística:

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

  Donde:

  $x^2 = Valor\ del\ estadistico\ Chi - cuadrado$

  $\sum$= la suma sobre todas las categorías o celdas de la tabla

  $O_i = frecuencia\ observada\ en\ la\ categoría\ o\ celda\ I$

  $E_i =$ Frecuencia esperada en la categoría o celda i bajo la hipótesis nula

  Por otro lado, el cálculo de frecuencias esperadas se opera de la siguiente manera:

$$E_{ij} = \frac{(Total\ de\ la\ fila\ i) x\ (Total\ de\ la\ columna\ j)}{Total\ general}$$

- Eliminación Recursiva de Características (RFE): por sus siglas en inglés Recursive Feature Elimination, es un algoritmo de selección de características wrapper. Esto significa que se basa en el rendimiento de un modelo predictivo específico para evaluar y seleccionar las características más relevantes de un conjunto de datos (Guyon et al., 2002).

- El principio fundamental de RFE es entrenar repetidamente un modelo, eliminar las características menos importantes en cada iteración y luego volver a entrenar el modelo con el subconjunto restante de características. Este proceso recursivo permite identificar el subconjunto de características que ofrece el mejor rendimiento predictivo para el modelo elegido.

### 6.3.2. *Evaluación de Modelos de Machine Learning*

Para garantizar la efectividad de un modelo predictivo, es fundamental aplicar métricas de evaluación adecuadas. En el caso de la predicción de MME, donde se busca identificar gestantes con alto riesgo, las métricas más utilizadas son:

#### 6.3.2.1. Validación cruzada

Esta técnica permite evaluar el rendimiento de un modelo predictivo de aprendizaje automático (ML). A través del proceso de validación, se determina si los resultados obtenidos describen adecuadamente las relaciones hipotéticas entre las variables. Para llevar a cabo esta evaluación, es necesario reservar de antemano una parte de los datos del conjunto de entrenamiento, que se utilizarán posteriormente para probar el modelo.

Entre las técnicas más utilizadas en este método se encuentra **Train-Test Split**, que consiste en dividir aleatoriamente el conjunto de datos, reservando entre un 70 % y 80 % para el entrenamiento y el resto para la validación. Sin embargo, cuando los datos son limitados, este enfoque puede no ser el más eficiente.

En estos casos, el método **K-Folds Cross Validation** resulta más adecuado, ya que garantiza que todas las observaciones tengan la oportunidad de aparecer tanto en el conjunto de entrenamiento como en el de prueba. Este método divide el conjunto de datos en *K* grupos o *folds*. Un valor típico de *K* oscila entre 5 y 10, dependiendo del tamaño del conjunto de datos. El procedimiento consiste en entrenar el modelo utilizando *K-1* folds y validarlo con el fold restante. Se registran las puntuaciones y errores en cada iteración, repitiendo el proceso hasta que cada fold haya sido utilizado para la validación. Finalmente, el promedio de las puntuaciones obtenidas se emplea como métrica de rendimiento del modelo (DataScientest, 2024)

#### 6.3.2.2. *Precisión (Accuracy)*

La precisión mide la proporción de predicciones correctas realizadas por el modelo sobre un conjunto de datos. Puede considerarse como una medida de exactitud y, generalmente, se calcula utilizando un conjunto de pruebas independientes que no fueron empleados durante el proceso de aprendizaje. Su formulación está regida por la siguiente ecuación (Diaz, 2025):

$$Presición = \frac{VP}{(VP + FP)}$$

Donde:

- **VP:** Verdaderos positivos
- **FN:** Falsos negativos
- **FP:** Falsos positivos

#### 6.3.2.3. Sensibilidad (Recall):

También llamada razón de verdaderos positivos (VPR), mide la capacidad del modelo para identificar correctamente los casos positivos (gestantes con patología MME). Se define mediante la ecuación:

$$VPR = \frac{VP}{(VP + FN)}$$

Donde:

- **VP:** Verdaderos positivos
- **FN:** Falsos negativos

6.3.2.4. F1 Score

Representa el equilibrio entre precisión y sensibilidad, proporcionando una evaluación más justa del rendimiento del modelo. Se calcula como un promedio ponderado de ambas medidas, según la siguiente ecuación. El valor más favorable es aquel que tiende a 1. Cabe destacar que este método se utiliza cuando el conjunto de datos está desbalanceado, ya que combina precisión y recall en una sola métrica (Diaz, 2025), véase la siguiente ecuación:

$$F1 = 2\,x\,\frac{Presición\ \ x\ Recall}{Presión + Recall}$$

6.3.2.5. Área bajo la curva ROC (AUC-ROC):

La curva ROC (Receiver Operating Characteristic) representa gráficamente la relación entre la sensibilidad y la especificidad en un sistema de clasificación binario a medida que varía el umbral de discriminación. En otras palabras, muestra la proporción de verdaderos positivos frente a falsos positivos, lo que facilita la evaluación del desempeño del modelo (Rodriguez, 2017).

Para comprender la curva ROC, es necesario conocer algunos conceptos clave:

- **Atributo:** También llamado campo, variable o característica, es una cantidad que describe un ejemplo y tiene un dominio definido. Los dominios más comunes son categórico, nominal, ordinal y continuo.
- **Matriz de confusión:** Es una tabla que muestra las clasificaciones previstas y reales de un modelo. Su tamaño es L×LL \times LL×L, donde LLL representa el número de posibles valores de la etiqueta. La interpretación de esta matriz se facilita mediante su representación gráfica, lo que da origen a la curva ROC.

# 7. Metodología

## *7.1. Enfoque, Alcance y Diseño de la Investigación*

Esta investigación se desarrollará en dos fases metodológicas principales. ***La primera fase***, con un enfoque cuantitativo, se centrará en el análisis de las variables clínicas y sociodemográficas asociadas a la MME, abarcando los objetivos 1 y 2. Se aplicarán técnicas estadísticas para asegurar la calidad y consistencia de los datos. ***En la segunda fase*** se orientará a proponer una metodología detallada para el desarrollo de un modelo predictivo de riesgo de MME, en concordancia con el objetivo 3, empleando técnicas ML.

***En la primera fase***, el análisis se enfocará en variables tanto numéricas como categóricas relevantes para la salud materna. Se garantizará la consistencia y adecuación de las fuentes de datos, incluyendo RIPS[6], SIVIGILA y bases internas de EPS MUTUAL SER, para cumplir con el primer objetivo de analizar exhaustivamente los datos clínicos y sociodemográficos.

***La segunda fase*** consistirá en la formulación de una metodología robusta para la construcción de un modelo predictivo de riesgo de MME. Se explorarán diversos algoritmos de *ML* y se establecerán criterios para la evaluación de su desempeño, abordando así el tercer objetivo de la investigación.

[6] Registros individuales de Prestación de Servicios de la salud, conjunto de datos que documentan cada servicio de salud prestado en Colombia del Ministerio de salud y Protección Social de Colombia.

El diseño de investigación para la primera fase será no experimental, descriptivo y de corte transversal, ya que no se manipularán variables y la recolección de datos se realizará en un único momento temporal, respondiendo al primer objetivo de caracterización.

## 7.2. Definición y Operacionalización de Variables

### 7.2.1. *Identificación inicial de Variables*

Para caracterizar inicialmente el perfil clínico y sociodemográfico de las gestantes, se definirán variables que cumplan con dos criterios fundamentales: (i) su asociación con el riesgo de MME, respaldada por la literatura científica, las directrices de la OMS y el Ministerio de Salud de Colombia, y (ii) su disponibilidad y consistencia en las fuentes de datos empleadas. La selección de estas variables permitirá un análisis detallado de los datos clínicos y sociodemográficos, asegurando la calidad y consistencia necesarias para alcanzar el primer objetivo de la investigación.

### 7.2.2. *Variables Clínicas*

Las variables clínicas se centrarán en condiciones patológicas reconocidas como factores de riesgo para la MME. Estas se identificarán mediante los códigos de la CIE-10[7], asegurando una clasificación diagnóstica estandarizada, válida y comparable, véase en la Tabla 1.

**Tabla 1**

*Variables clínicas preliminares asociadas a MME*

[7] Clasificación Internacional de Enfermedades décima edición. Son un sistema organizado desarrollado por la OMS que pretende clasificar y codificar enfermedades, trastornos, síntomas, hallazgos clínicos causas externas de enfermedad y factores sociales que influyen en la salud. Se componen por una letra seguida de un número.

| Variable clínica | Descripción general | Códigos CIE-10 asociados |
|---|---|---|
| **Preeclampsia y eclampsia** | Trastornos hipertensivos específicos del embarazo | O14, O15 |
| **Hipertensión gestacional** | Hipertensión sin proteinuria durante el embarazo | O13 |
| **Diabetes gestacional** | Intolerancia a la glucosa detectada durante la gestación | O24.4, O24.9 |
| **Hemorragias obstétricas** | Hemorragias antes, durante o después del parto | O44, O45, O46, O67, O72 |
| **Infecciones graves** | Infecciones sistémicas durante el embarazo, parto o puerperio | O23, O75.3, A41, N39.0 |
| **Trastornos del parto** | Complicaciones durante el trabajo de parto | O62, O63, O64, O66 |

Fuente: Elaboración Propia

7.2.3. *Variables Sociodemográficas*

Estas variables proporcionarán el contexto social de las gestantes, facilitando el análisis de los determinantes sociales asociados al riesgo materno. Se clasificarán como cuantitativas o categóricas[8] según su naturaleza véase la Tabla 2, donde se agrupan este tipo de variables que se tomaran como eje inicial.

**Tabla 2**

*Variables sociodemográficas preliminares*

| Variable | Descripción | Tipo de variable |
|---|---|---|
| **Edad** | Edad en años cumplidos al momento del parto o evento | Cuantitativa |
| **Nivel educativo** | Nivel máximo alcanzado (sin educación, básica, etc.) | Categórica |
| **Área de residencia** | Zona geográfica: rural o urbana | Categórica |
| **Régimen de afiliación** | Contributivo, subsidiado, otro | Categórica |
| **Grupo étnico** | Autorreconocimiento (indígena, afrodescendiente, etc.) | Categórica |
| **Número de embarazos previos** | Total, de gestaciones anteriores | Cuantitativa |

Fuente: Elaboración Propia

[8] Representa grupo o categorías y no tiene un orden numérico, expresan cualidades o atributos.

Estas variables constituirán los atributos fundamentales para identificar patrones de riesgo en la población de estudio, lo que permitirá cumplir con el segundo objetivo: Una adecuada definición, clasificación y codificación de estas variables es esencial, ya que facilitará no solo el procesamiento de los datos, sino también la correcta aplicación de los métodos analíticos y predictivos.

### 7.2.4. *Diccionario de Variables*

A continuación, en Tabla 3, se presenta el diccionario de datos que conforma la base de datos o dataset construida, con el fin de facilitar el análisis.

Cada una de las variables de la base de datos inicial (correspondiente al muestreo completo) fue utilizada para el análisis exploratorio y el desarrollo del modelo de clasificación de riesgo de morbilidad materna extrema. Las variables totales consideradas en el estudio.

**Tabla 3**
*Diccionario de variables*

| VARIABLE | DESCRIPCIÓN |
|---|---|
| **FUM** | Fecha de la Última Menstruación, punto de partida para estimar la edad gestacional. |
| **FPP** | Fecha Probable de Parto, calculada a partir de la FUM. |
| **SEMANA_GESTACIONAL** | Número de semanas de gestación al momento del registro. |
| **EDAD** | Edad de la paciente en años. |
| **MAYOR_35** | Indicador (por ejemplo, 1/0 o Sí/No) que señala si la paciente tiene 35 años o más. |
| **NUMEROS_PARTOS_CESARIAS** | Total, de partos y cesáreas realizados previamente. |
| **ABORTO** | Número de abortos previos. |
| **VIVOS** | Número de partos que resultaron en nacidos vivos. |
| **MUERTOS** | Número de partos que resultaron en muertes (intrauterinas o neonatales). |
| **RIESGO_PREECLAMPSIA** | Indicador de riesgo para desarrollar preeclampsia. |
| **NUMEROS_CONTROLES_PRENATALES** | Cantidad de controles o visitas prenatales realizados. |
| **IMC** | Índice de Masa Corporal de la paciente. |

| VARIABLE | DESCRIPCIÓN |
|---|---|
| **RIESGO** | Valor o clasificación de riesgo general derivado de la combinación de indicadores clínicos. |
| **CONSULTA_URGENCIA_ULTIMOS_30_DIAS** | Número de consultas de urgencia realizadas en los últimos 30 días. |
| **NACIONALIDAD_PROCEDENCIA** | Nacionalidad o procedencia de la paciente. |
| **CODIGO_OCUPACION** | Código que clasifica la ocupación de la paciente. |
| **NIVEL_EDUCATIVO** | Nivel educativo alcanzado. |
| **AFIC_GRUPO_ETNICO** | Afiliación o pertenencia a un grupo étnico. |
| **AFIN_NIVEL_SISBEN** | Nivel del SISBEN asignado, referente a la clasificación socioeconómica. |
| **AFIN_GRUPO_POBLACIONAL** | Clasificación según el grupo poblacional al que pertenece la paciente. |
| **AFIC_ZONA** | Zona de afiliación o residencia (por ejemplo, urbana o rural). |
| **TIPO_DE_CASO** | Clasificación del caso según criterios predefinidos. |
| **COD_MUNICIPIO** | Código del municipio de residencia o atención. |
| **DIFERENCIA_FIP_FUM** | Diferencia en días entre la Fecha de Inicio del Parto (FIP) y la FUM, que puede ayudar a validar la cronología de eventos. |
| **HIPERTENSION** | Indicador de presencia de hipertensión en la paciente. |
| **VIH_MATERNO_CONFIRMADO** | Indicador que confirma la infección por VIH en la paciente. |
| **TAMIZAJE_SIFILIS** | Resultado o indicación de la realización del tamizaje para sífilis. |
| **TAMIZAJE_VIH** | Resultado o indicación de la realización del tamizaje para VIH. |
| **TAMIZAJE_HEPATITIS** | Resultado o indicación de la realización del tamizaje para hepatitis. |
| **DIAGNOSTICOS** | Diagnósticos clínicos asociados al caso, que pueden incluir complicaciones o condiciones preexistentes. |
| **HEMOGLOBINA** | Nivel de hemoglobina medido, utilizado como indicador de anemia. |
| **FECHA_HB** | Fecha en la que se realizó la medición de hemoglobina. |
| **GLUCOSA_PRE** | Nivel de glucosa en ayunas o prepandial. |
| **GLUCOSA_1_HORA** | Nivel de glucosa medido a 1 hora tras la administración de glucosa. |
| **GLUCOSA_2_HORA** | Nivel de glucosa medido a 2 horas tras la administración de glucosa. |
| **FECHA_GLUCOSA** | Fecha en la que se realizó la prueba de glucosa. |

| VARIABLE | DESCRIPCIÓN |
|---|---|
| **HEMORRAGIA** | Indicador de la ocurrencia de hemorragia durante el embarazo o el parto. |
| **POLIHIDRAMNIOS** | Indicador de la presencia de polihidramnios (exceso de líquido amniótico). |
| **MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA** | Registro de antecedentes de mortalidad perinatal o neonatal tardía. |
| **TRIMESTRE** | Trimestre del embarazo en el que se encuentra la paciente. |
| **ETIQUETA_MORBILIDAD** | Etiqueta objetivo que indica la presencia o riesgo de morbilidad materna extrema, basada en criterios predefinidos. |

Fuente: Elaboración Propia

### 7.3. Población y Muestra

La población de estudio estará conformada por mujeres afiliadas a la EPS MUTUAL SER ESS, residentes en los departamentos de Bolívar, Córdoba, Magdalena, Sucre y Atlántico, que hayan estado embarazadas durante el período comprendido entre noviembre de 2019 y febrero de 2025. La base de datos incluye un total de 88.810 registros.

Se realizará un muestreo censal, incluyendo todos los registros que cumplan con los criterios de inclusión y exclusión previamente definidos. La obtención de los datos se llevó a cabo a partir de la información suministrada por la EPS y mediante consultas directas al sistema SIVIGILA.

### 7.4. Técnicas de Análisis de Datos

La identificación de patrones y la selección de variables relevantes para la predicción de MME se llevarán a cabo mediante diversas técnicas estadísticas, abordando el objetivo 2. Las siguientes etapas se proponen para alcanzar estos objetivos:

Se aplicarán las siguientes técnicas:

- **Análisis Exploratorio de Datos (EDA):** Inicialmente, se realizará un análisis exploratorio de datos para describir las características de la muestra y detectar patrones

preliminares, utilizando herramientas visuales como histogramas, diagramas de caja, gráficos de dispersión y mapas de calor.

- **Análisis de Correlación:** Para evaluar la relación entre las variables explicativas y la condición de MME, se calcularán los coeficientes de correlación de Pearson (para variables numéricas) y Spearman (para variables ordinales o no lineales). Adicionalmente, se examinará la colinealidad entre variables para mitigar redundancias en el modelo.
- **Selección de Características (*Feature Selection*):** Se emplearán métodos de selección de características para identificar las variables con mayor poder predictivo:

    - **Análisis Chi-cuadrado:** Para evaluar la asociación entre variables categóricas y la variable objetivo (MME).
    - **Importancia de variables en Random Forest:** Para determinar la relevancia de cada variable en la predicción mediante este algoritmo.

Estas técnicas permitirán cumplir con el segundo objetivo, asegurando la identificación de las variables más relevantes para la predicción del riesgo de MME, y a su vez, mejorarán la interpretabilidad del modelo y prevendrán el sobreajuste.

### 7.5. Propuesta Metodológica para el Futuro Desarrollo del Modelo Predictivo (segunda etapa)

Para el desarrollo del objetivo 3 se propone seguir la siguiente metodología:

- **Etapa 1. Revisión Bibliográfica Especializada:** Búsqueda sistemática en bases de datos científicas (PubMed, Scopus, IEEE, Google Scholar; etc.) de estudios que apliquen modelos de ML en predicción de riesgo en salud. Se documentarán tipos de modelos, métricas de desempeño, limitaciones y buenas prácticas.

- **Etapa 2. Selección y Justificación de Algoritmos Potenciales:**

Basándose en la revisión y criterios técnicos, se propone considerar los siguientes algoritmos:

- Random Forest
- Regresión Logística

- **Etapa 3. Definición de Criterios de Evaluación:** Se propondrán métricas de evaluación clave (precisión, sensibilidad y robustez).

### 7.6. Entregables de la Metodología Propuesta

Los productos derivados del proceso metodológico propuesto se exponen en la siguiente la Tabla 4.

**Tabla 4**

*Entregables por objetivo*

| Objetivo Específico | Fase Metodológica | Producto / Entregable | Formato |
|---|---|---|---|
| **Objetivo 1** | Análisis exploratorio y caracterización de datos | N/A | |
| **Objetivo 2** | Identificación de patrones y selección de variables | Reporte de correlaciones y redundancias | N/A |
| | | Informe de selección de características (feature selection) | Anexo 1 Análisis exploratorio estadístico/ Documento .PDF |
| | | Informe de análisis exploratorio (EDA) | |
| | | Matriz de correlación (variables y MME) | |

| Objetivo Específico | Fase Metodológica | Producto / Entregable | Formato |
|---|---|---|---|
| | Revisión bibliográfica | Estado del arte sobre modelos predictivos aplicados a problemas clínicos similares | NA |
| **Objetivo 3** | | | |
| | Metodología para el desarrollo del modelo | Documento metodológico con fases, técnicas y criterios para el desarrollo del modelo MME | /Anexo 2 Metodología propuesta Documento PDF |

Fuente: Elaboración Propia

## 8. Análisis y discusión de los resultados

Para el desarrollo de esta investigación se utilizaron datos suministrados por la EPS objeto de estudio correspondientes al período de noviembre de 2019 a febrero de 2025 y registros complementarios extraídos del SIVIGILA. La población objetivo, como se describió en el apartado 7.3, está conformada por mujeres afiliadas a la EPS MUTUAL SER ESS ubicadas en los departamentos de Bolívar, Córdoba, Magdalena, Sucre y Atlántico.

El conjunto de datos final integrado para el análisis contenía **88.810 registros**. Esta base fue consolidada mediante un proceso riguroso de recopilación, validación y depuración de información proveniente de múltiples fuentes: diagnósticos clínicos, historias médicas, resultados de laboratorio, eventos obstétricos reportados, y bases estructuradas internas de la EPS. La integración de estos datos permitió construir una base robusta para el modelado predictivo, garantizando representatividad regional y temporal.

Durante el preprocesamiento inicial, se evaluaron aspectos fundamentales como:

- **Estructura de los datos:** El DataFrame[9] original contenía más de 50 variables entre clínicas, sociodemográficas y de atención médica. Estas incluían tanto variables continuas (como edad, índice de masa corporal, niveles de hemoglobina y glucosa) como categóricas (nivel educativo, zona de residencia, grupo étnico, etc.).
- **Calidad de los datos:** Se identificaron celdas vacías, valores nulos, y casos con codificación atípica (por ejemplo, valores biológicamente imposibles en glucosa o semanas de gestación). Estos hallazgos motivaron una limpieza de datos que incluyó imputación selectiva, winsorización[10] y transformación de variables.
- **Balance de clases:** Un aspecto crítico detectado fue el desbalance significativo entre los casos positivos de Morbilidad Materna Extrema (MME) y los negativos. Aproximadamente el **5% del total de registros correspondían a casos positivos de MME**, mientras que el 95% restante no presentaban dicha condición. Este desbalance obliga a aplicar técnicas de muestreo (como SMOTE y undersampling) y métricas específicas durante el entrenamiento del modelo.
- **Visualización exploratoria:** A través de histogramas, boxplots y mapas de calor, se evidenciaron patrones clínicamente relevantes, como la mayor proporción de MME en mujeres mayores de 40 años y la alta correlación entre el número de partos/cesáreas y número de hijos vivos. Se identificaron outliers en variables como IMC y niveles de glucosa, los cuales fueron tratados apropiadamente para evitar distorsiones en el entrenamiento del modelo.

En conjunto, estos análisis permitieron validar la consistencia interna del conjunto de datos y establecer una base confiable para los procedimientos posteriores de ingeniería de características y selección de variables predictivas, cuya discusión detallada se expone en los apartados siguientes.

[9] Un DataFrame es una estructura de datos tabular en Python (especialmente en la librería pandas) que organiza la información en filas y columnas, similar a una hoja de cálculo o una tabla de base de datos, facilitando el análisis, manipulación y visualización de datos.

[10] winsorización es una técnica de tratamiento de outliers que consiste en limitar los valores extremos de una variable numérica a un cierto percentil.

La información recopilada se analizó mediante la categorización de los datos en cuatro grupos principales:

1. Embarazos sin complicaciones
2. Embarazos con eventualidades asociadas a diagnósticos específicos
3. Población con diagnóstico de MME
4. Embarazos que culminaron en aborto, parto o cesaría en diferentes etapas de la gestación

Esta clasificación permitió establecer un análisis comparativo y determinar patrones de comportamiento en los diferentes escenarios obstétricos estudiados.

### 8.1. Macrocategorización del estudio:

La clasificación anterior se organizó dentro de una macrocategorización principal basada en la presencia o ausencia de MME:

- **Casos positivos de MME**: Pacientes con diagnóstico confirmado de morbilidad materna extrema
- **Casos negativos de MME**: Pacientes sin evidencia de morbilidad materna extrema

Esta dicotomización constituyó la variable objetivo de control del estudio, permitiendo la comparación entre ambos grupos y estableciendo como unidad de análisis la distribución de casos según la presencia de MME.

### 8.2. Distribución de la muestra:

- Casos sin MME: 84.415
- Casos confirmados de MME: 4.395
- **Total, de registros analizados: 88.810**

A continuación, se presentan los casos positivos de MME identificados en los datos analizados por departamento (Tabla 5):

**Tabla 5**
*Cantidad de casos identificados de MME*

| NOMBRE DEPARTAMENTO | CANTIDAD | CANTIDAD_MME | CANTIDAD DE CASOS NO MME |
|---|---|---|---|
| BOLÍVAR | 30.266 | 1.489 | 28777 |
| CÓRDOBA | 21.112 | 955 | 20157 |
| ATLÁNTICO | 16.561 | 1.159 | 15402 |
| SUCRE | 11.219 | 345 | 10874 |
| MAGDALENA | 9.652 | 447 | 9205 |
| **TOTAL** | **88.810** | **4.395** | **84.415** |

*Fuente:* Elaboración propia

El procedimiento de total del tratamiento de la base de datos mediante el uso de las herramientas del lenguaje de programación en la herramienta Python 3.0[11] se encuentra en el Anexo 1 Análisis exploratorio estadístico

## 8.3. Análisis Exploratorio de Variables numéricas y categóricas

A continuación, se presentan los resultados del análisis exploratorio mediante histogramas de las 17 variables numéricas identificadas en la base de datos de 88.810 registros.

### 8.3.1. *Análisis de Distribución de Variables Numéricas y categóricas mediante Histogramas*

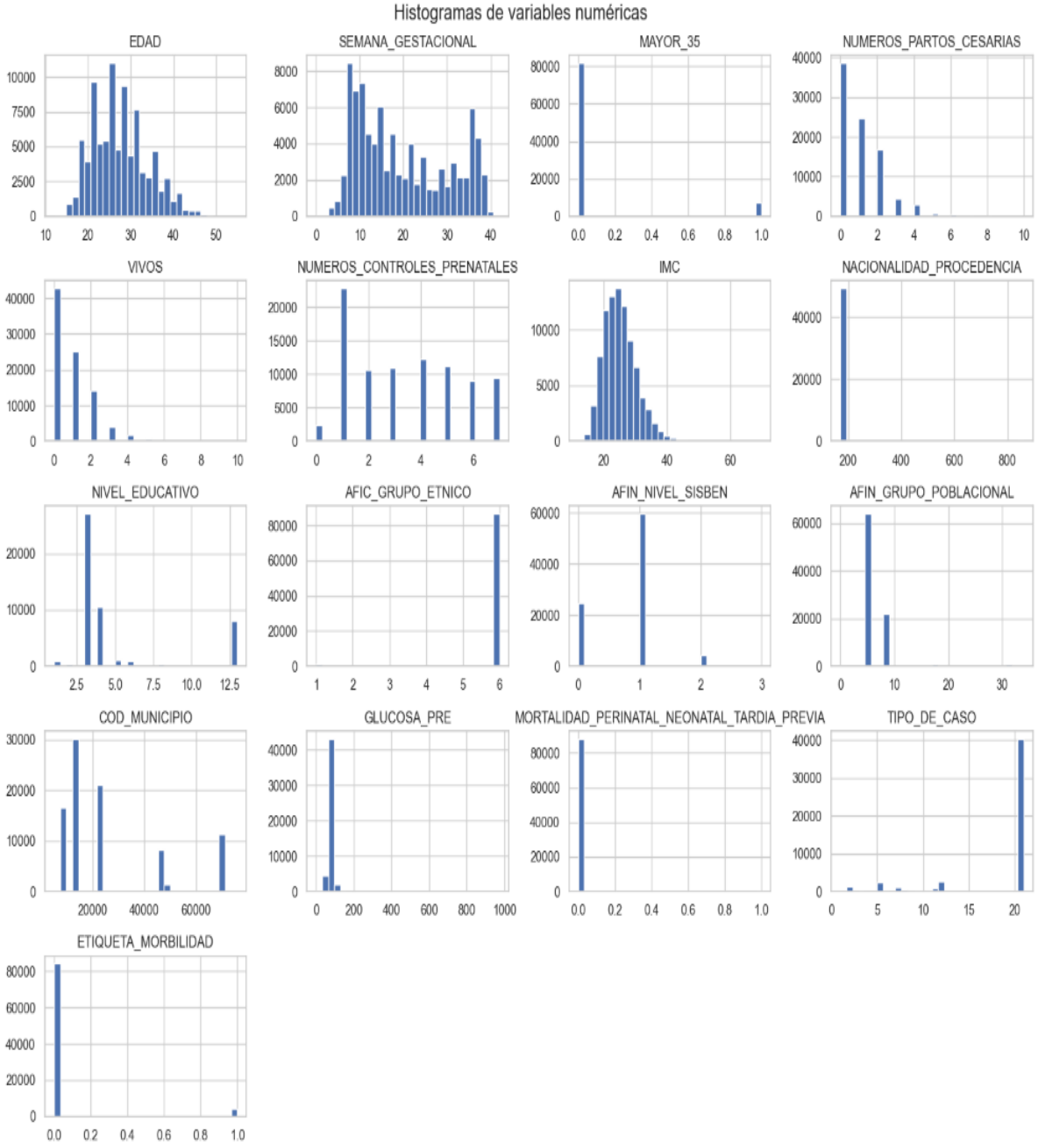
Con el objetivo de comprender la distribución y comportamiento de las variables numéricas en el conjunto de datos, se elaboraron histogramas que permitieron identificar patrones clave,

[11] Python 3.0 se usa en análisis de datos para limpiar, explorar y visualizar información, además de ayudar en la selección de características clave mediante librerías como pandas, NumPy y scikit-learn, facilitando así la preparación efectiva de datos para modelos de machine learning.

valores atípicos y posibles necesidades de transformación para el posterior modelado, que se presentan a continuación en la Figura 3:

**Figura 3**

*Histogramas de variables numéricas*




*Fuente:* Elaboración Propia

Identificados los resultados de los histogramas se realizó el siguiente análisis de comportamiento de las variables

8.3.1.1. Variables Demográficas (análisis con histogramas)

- **EDAD**: La distribución etaria de las gestantes se concentra principalmente entre los 20 y 35 años, patrón epidemiológicamente esperado en población gestante. Se observa asimetría negativa leve, con frecuencia considerablemente menor en mujeres mayores de 40 años, representando el comportamiento reproductivo típico de la población colombiana.
- **MAYOR_35**: Como variable categórica binaria, confirma que la mayoría de las gestantes tienen menos de 35 años. Este desbalance requiere consideración especial durante el análisis para evitar sesgos.

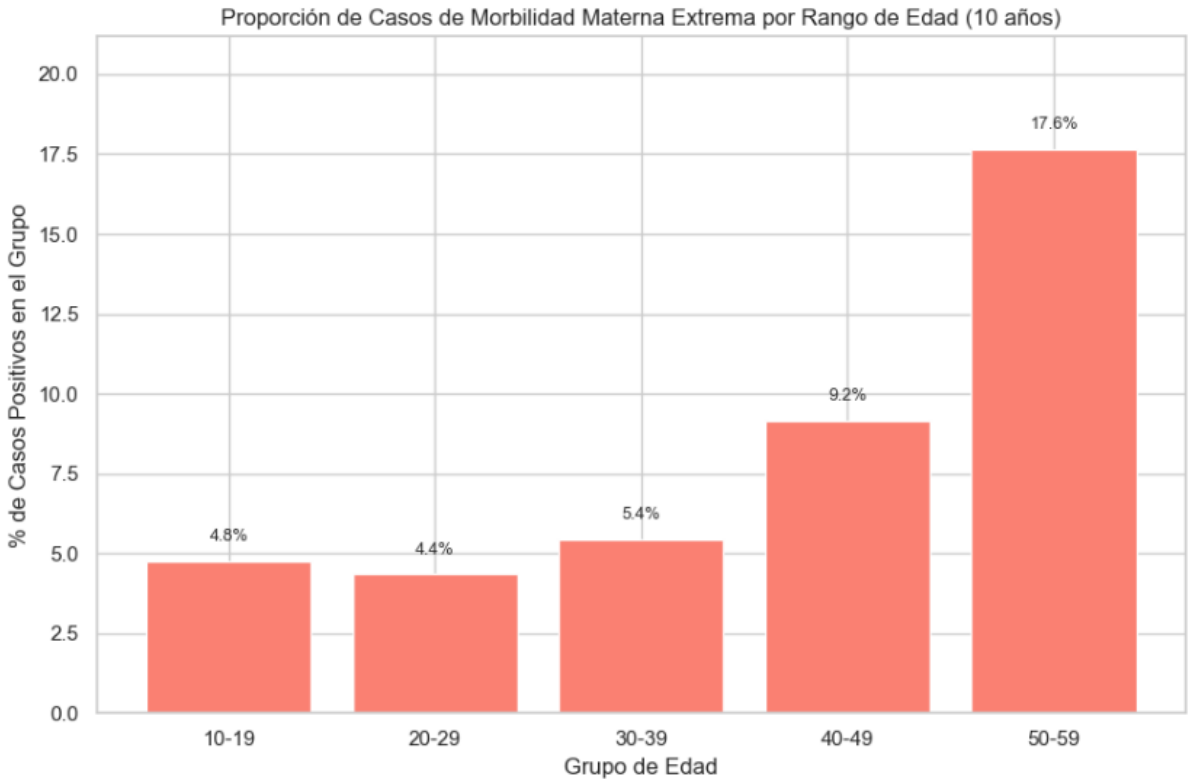
Dicho esto, se identificó por edades un análisis etario:

8.3.1.1.1. Análisis MME por Grupos Etarios

Para profundizar en el análisis de la variable edad, se realizó una estratificación de casos positivos de MME por intervalos decenales (10-19, 20-29, 30-39, 40-49 y 50-59 años), cuyos resultados se presentan en la Figura 4.

**Figura 4**

*Proporción de Casos de Morbilidad Materna Extrema por Rango de Edad*




*Fuente:* Elaboración propia

La Figura 4 evidencia un patrón crítico en la distribución de la morbilidad materna extrema según la edad, con implicaciones directas para la gestión del riesgo obstétrico y la formulación de políticas públicas en salud materna.

8.3.1.1.1.1. Hallazgos principales

El análisis revela un incremento exponencial en la proporción de casos MME conforme avanza la edad materna:

- **Grupos 10-39 años**: Mantienen tasas relativamente estables por debajo del 6%
- **Grupo 40-49 años**: Experimenta duplicación del riesgo (9.2%)

- **Grupo 50-59 años**: Presenta la tasa más crítica con 17.6% de casos MME

8.3.1.1.1.2. Interpretación Clínica:

Este patrón epidemiológico se explica por múltiples factores fisiopatológicos asociados al envejecimiento reproductivo:

1. **Complicaciones específicas del embarazo tardío**: Mayor incidencia de hipertensión gestacional, diabetes gestacional, preeclampsia y necesidad de cesáreas de emergencia.
2. **Comorbilidades preexistentes**: Acumulación de condiciones crónicas (hipertensión, diabetes tipo 2, enfermedades cardiovasculares) que incrementan la vulnerabilidad materna.
3. **Cambios fisiológicos relacionados con la edad**: Disminución de la reserva funcional de órganos vitales que compromete la adaptación a los cambios hemodinámicos del embarazo.

8.3.1.1.1.3. Implicaciones para la Atención Obstétrica:

Aunque el grupo de 50-59 años representa una frecuencia absoluta baja de embarazos, la proporción crítica de complicaciones graves (17.6%) demanda:

- Protocolos de atención diferenciada para embarazos en edad materna avanzada
- Seguimiento especializado multidisciplinario
- Estrategias de intervención preventiva priorizadas

En contraste, los grupos etarios de 20-39 años, considerados como rango reproductivo óptimo, confirman menor carga de riesgo, respaldando la evidencia clínica sobre el momento ideal para la gestación desde la perspectiva de seguridad materna.

8.3.1.2. Variables Obstétricas (Análisis con Histogramas)

- **SEMANA_GESTACIONAL**: La distribución gestacional muestra mayor concentración entre las semanas 10 y 30, con disminución progresiva hacia el término. Los valores extremos identificados (≤2 semanas o ≥42 semanas) representan el 0.8% de la muestra y requieren validación clínica para descartar errores de codificación o casos de embarazos ectópicos y postérmino respectivamente.

- **NUMEROS_PARTOS_CESARIAS y VIVOS**: Ambas variables exhiben distribución exponencial negativa típica de variables de conteo. El 78.3% de las gestantes registran entre 0-2 partos previos, mientras que el 82.1% tienen entre 0-2 hijos vivos, reflejando patrones demográficos consistentes con la transición demográfica colombiana hacia familias de menor tamaño.
- **MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA**: Casi todos los registros presentan valor 0, con muy pocos casos positivos, indicando una variable altamente desbalanceada que, aunque clínicamente relevante, requiere técnicas específicas de modelado.

8.3.1.3. Variables de Atención Prenatal (Análisis con Histogramas)

- **NUMEROS_CONTROLES_PRENATALES**: La distribución aproximadamente uniforme sugiere buena cobertura de atención prenatal o un registro sistemático que agrupa los controles en intervalos regulares.

8.3.1.4. Variables Clínicas (Análisis con Histogramas)

- **IMC**: El índice de masa corporal presenta forma aproximadamente normal, centrada en valores considerados saludables, lo cual facilita su uso en modelos predictivos sin necesidad de transformaciones adicionales.

- **GLUCOSA_PRE**: Se identificaron valores biológicamente implausibles (>500 mg/dL) en el 0.12% de la muestra, con registros extremos alcanzando 1000 mg/dL. Estos valores requieren depuración mediante winsorización o exclusión, dado que glucemias >400 mg/dL son incompatibles con supervivencia materna sin intervención inmediata.

8.3.1.5. Variables Socioeconómicas (Análisis con Histogramas)

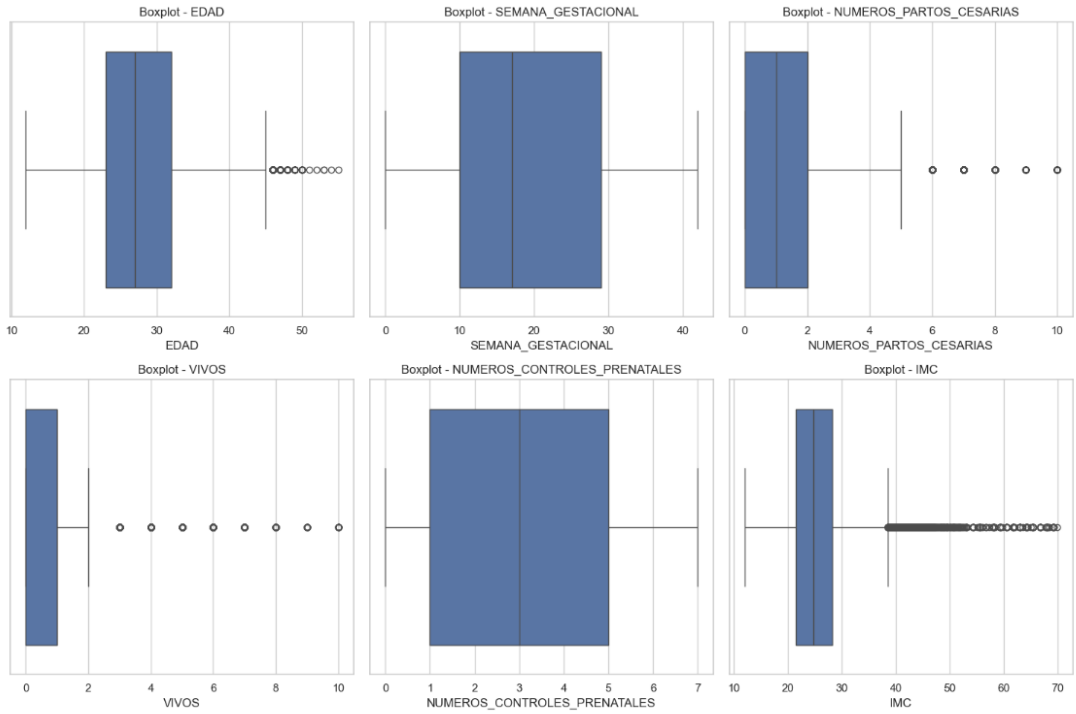
- **NACIONALIDAD_PROCEDENCIA, NIVEL_EDUCATIVO, AFIN_***: Estas variables categóricas muestran distribuciones altamente sesgadas o agrupadas en pocos valores dominantes. Particularmente, AFIN_NIVEL_SISBEN y AFIC_GRUPO_ETNICO presentan clases claramente predominantes, útiles pero que podrían inducir sesgos si no se manejan adecuadamente en modelos supervisados.
- **TIPO_DE_CASO**: Presenta múltiples valores con algunos muy poco frecuentes, sugiriendo la conveniencia de agrupar categorías raras para simplificar el análisis.

8.3.2. *Análisis de Boxplots de Variables Numéricas y Categóricas para Detección de Outliers y Distribución*

Como parte del análisis exploratorio de datos, se generaron boxplots para identificar posibles valores atípicos, rangos intercuartílicos y la distribución general de algunas variables numéricas clave. A continuación, se presentan las observaciones más relevantes:

**Figura 5**

*Boxplots de 6 variables*




*Fuente:* Elaboración Propia

- **EDAD:** El boxplot muestra una distribución centrada entre los 20 y 35 años, con una ligera presencia de outliers hacia la derecha (edades superiores a los 45 años). Esto es esperable y consistente con el análisis de histogramas, ya que los embarazos en edades avanzadas son menos frecuentes, pero clínicamente importantes por su asociación a mayores riesgos. Aunque hay outliers, no se eliminan, ya que representan casos clínicamente relevantes.
- **SEMANA_GESTACIONAL:** El boxplot de esta variable muestra una distribución simétrica, sin outliers aparentes según el método de los rangos intercuartílicos, lo que indica que los registros son razonablemente confiables en términos de semanas de embarazo, con valores concentrados entre las semanas 10 y 40. Esto contrasta ligeramente con la identificación de valores extremos en los histogramas (≤2 o ≥42 semanas), sugiriendo que aunque estos valores existen y requieren validación clínica, podrían no ser

outliers estadísticos según la definición del boxplot pero sí puntos de datos raros y significativos. Refuerza que la variable está bien capturada y podría utilizarse con la debida validación de extremos.

- **NUMEROS_PARTOS_CESARIAS:** Se observa una clara concentración entre 0 y 3 partos/cesáreas, pero aparecen valores atípicos a partir de 6, llegando hasta 10. Esto complementa la observación de la distribución exponencial negativa de los histogramas. Aunque pueden ser reales, también podrían representar casos excepcionales o errores de digitación. Se puede considerar recortar o imputar valores extremos mayores a 7 para evitar distorsiones en los modelos.
- **VIVOS:** La gran mayoría de los casos está entre 0 y 2 hijos vivos, pero existen varios outliers que superan los 6, llegando hasta 10. Similar al caso anterior y en línea con los histogramas, no se descarta que algunos sean reales, pero podrían requerir validación clínica o truncamiento. Por ahora, se marca esta variable como que requiere revisión.
- **NUMEROS_CONTROLES_PRENATALES:** Este indicador muestra una distribución bastante equilibrada entre 0 y 7 controles, sin valores extremos visibles en el boxplot. Esto es positivo, consistente con la distribución uniforme observada en histogramas, y representa consistencia en los registros. Se considera que esta variable está lista para ser usada sin necesidad de transformaciones adicionales.
- **IMC:** El boxplot del IMC revela una gran cantidad de outliers por encima de 40, alcanzando incluso valores cercanos a 70. Esto añade una perspectiva crítica al análisis de histogramas que la describía como "aproximadamente normal". La presencia de estos outliers significativos sugiere que, a pesar de una tendencia central saludable, los valores extremos podrían deberse a errores de digitación o a casos clínicos graves de obesidad mórbida que necesitarán tratamiento antes del modelado.

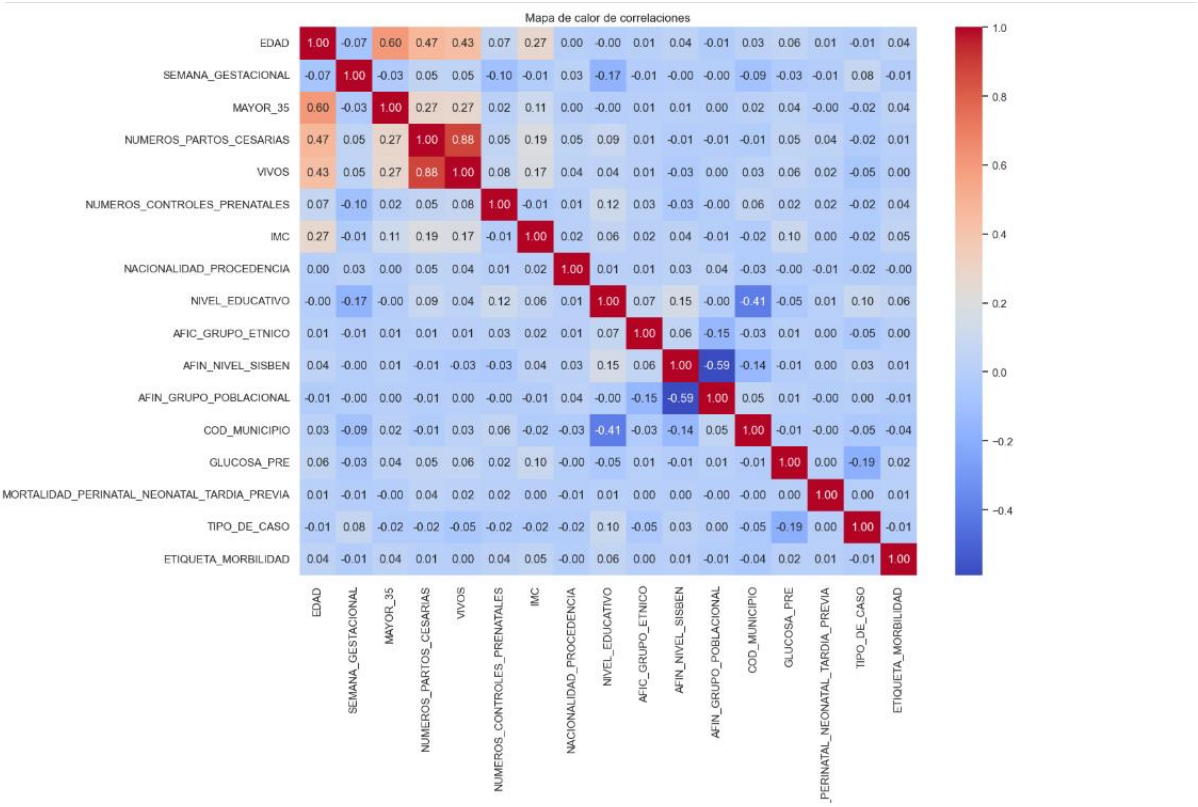
### 8.3.3. *Análisis de correlación de variables categóricas y numéricas por mapa de calor*

Con el objetivo de encontrar correlación entre variables y como parte del análisis exploratorio de datos y antes de la etapa de selección de características de estas, elaboramos un mapa de calor, el enfoque descrito a continuación se centra en las correlaciones de Pearson entre las variables

numéricas dentro de la base de datos, las cuales se visualizan en la Figura 6.Este análisis es fundamental para identificar posibles relaciones lineales entre las variables, detectar multicolinealidad y guiar la selección de características.

**Figura 6**

*Mapa de calor de correlación entre variables*




*Fuente:* Elaboración propia

Se calcularon coeficientes de Pearson para explorar relaciones lineales entre variables se identificaron:

- Correlaciones altas entre número de partos/cesáreas y número de hijos vivos.
- Correlaciones moderadas entre edad y número de partos, y entre edad e IMC.
- Correlaciones bajas o inexistentes entre la mayoría de las variables y la etiqueta de MME.

8.3.3.1. Correlaciones más destacadas identificadas

- **NUMEROS_PARTOS_CESARIAS ↔ VIVOS = 0.88**: Se observa una correlación extremadamente alta entre estas dos variables. Este resultado es clínicamente esperable, ya que un mayor número de partos y cesáreas suele estar asociado con un mayor número de hijos vivos. Sin embargo, esta alta colinealidad sugiere que probablemente solo una de estas variables debe ser incluida en el modelo final, para evitar redundancia informativa y problemas de multicolinealidad.
- **EDAD ↔ MAYOR_35 = 0.60**: Esta relación es igualmente lógica, dado que *MAYOR_35* es una variable binaria derivada directamente de la edad materna. Por esta razón, se optará por conservar la variable *EDAD* en el modelo, ya que ofrece mayor granularidad y permite capturar patrones más complejos o no lineales.
- **NUMEROS_PARTOS_CESARIAS ↔ EDAD = 0.47** y **VIVOS ↔ EDAD = 0.43**: Estas correlaciones moderadas reflejan una tendencia natural: las mujeres de mayor edad suelen haber tenido más partos y, por lo tanto, más hijos vivos. Aunque estas relaciones no son problemáticas desde el punto de vista estadístico, se tendrán en cuenta en el análisis multivariado para evitar la introducción de efectos colineales no deseados.

8.3.4. *Variable Objetivo*

**ETIQUETA_MORBILIDAD**: La variable objetivo presenta marcado desbalance de clases con 94.7% de casos negativos (n=84,213) versus 5.3% de casos positivos de MME (n=4,779). Esta proporción, aunque refleja la prevalencia real de MME en población obstétrica, genera un reto metodológico significativo que requiere estrategias específicas de muestreo y evaluación del modelo.

8.3.4.1. Correlación con la variable objetivo (ETIQUETA_MORBILIDAD):

8.3.4.1.1. Correlación de Pearson con variable objetivo

En la Tabla 6 se identifican las variables con mayor correlación con la etiqueta objetivo

**Tabla 6**
Resultados de la Correlación de Pearson[12]

| VARIABLE | CORRELACIÓN CON ETIQUETA_MORBILIDAD |
|---|---|
| ETIQUETA_MORBILIDAD | 1.000000 |
| NIVEL_EDUCATIVO | 0.056296 |
| IMC | 0.051311 |
| MAYOR_35 | 0.041940 |
| EDAD | 0.041932 |
| NUMEROS_CONTROLES_PRENATALES | 0.036776 |
| GLUCOSA_PRE | 0.017667 |
| AFIN_NIVEL_SISBEN | 0.014186 |
| MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA | 0.008082 |
| NUMEROS_PARTOS_CESARIAS | 0.007960 |
| AFIC_GRUPO_ETNICO | 0.002509 |
| VIVOS | 0.001418 |
| NACIONALIDAD_PROCEDENCIA | -0.002739 |
| SEMANA_GESTACIONAL | -0.007952 |
| TIPO_DE_CASO | -0.012015 |
| AFIN_GRUPO_POBLACIONAL | -0.013690 |
| COD_MUNICIPIO | -0.040140 |

*Fuente:* Elaboración Propia

Para identificar variables numéricas y categóricas con posible poder explicativo sobre la MME, calculamos la correlación de Pearson entre cada variable independiente y la variable objetivo (ETIQUETA_MORBILIDAD). Aunque esta medida solo capta relaciones lineales, proporcionó una visión preliminar.

[12] El coeficiente de correlación de Pearson mide la fuerza y dirección de la relación lineal entre dos variables numéricas, con valores que van desde -1 hasta 1. Un valor de 1 indica una correlación positiva perfecta, donde ambas variables aumentan juntas exactamente; entre 0.7 y 0.99 representa una correlación positiva fuerte; entre 0.4 y 0.69 es una correlación moderada; y entre 0.1 y 0.39, una correlación débil. Un valor cercano a 0 indica ausencia de correlación lineal. En sentido negativo, valores entre -0.1 y -0.39 reflejan correlación débil negativa, entre -0.4 y -0.69 moderada negativa, y entre -0.7 y -0.99 fuerte negativa, mientras que -1 significa correlación negativa perfecta, donde una variable sube y la otra baja en forma exacta. Es importante recordar que este coeficiente solo mide relaciones lineales y no implica causalidad.

8.3.4.1.1.1. Variables con mayor correlación positiva (aunque débil):

- NIVEL_EDUCATIVO: 0.056
- IMC: 0.051
- MAYOR_35: 0.041
- EDAD: 0.041

Estas variables muestran la correlación más alta con la MME, pero los valores son bajos (< 0.06), lo que confirma la ausencia de una relación lineal fuerte. Esto sugiere que la MME es un fenómeno complejo y multifactorial.

8.3.4.1.1.2. Variables con correlación casi nula:

GLUCOSA_PRE, AFIN_NIVEL_SISBEN, NUMEROS_PARTOS_CESARIAS, VIVOS, entre otras, presentan correlaciones < 0.02, indicando que no tienen una relación lineal directa con la MME. Sin embargo, podrían aportar valor en modelos no lineales o mediante interacciones.

8.3.4.1.1.3. Variables con correlación negativa (débil)

- COD_MUNICIPIO: -0.040
- AFIN_GRUPO_POBLACIONAL: -0.013
- TIPO_DE_CASO: -0.012

Estas variables tienen una asociación inversa marginal con la MME. Aunque su influencia individual parece limitada, se mantienen como candidatas para evaluar en fases posteriores.

**Ejemplos ilustrativos:**

- EDAD ↔ ETIQUETA_MORBILIDAD: 0.04
- IMC ↔ ETIQUETA_MORBILIDAD: 0.05

- NUMEROS_CONTROLES_PRENATALES ↔ ETIQUETA_MORBILIDAD: 0.00

Estos resultados subrayan la necesidad de integrar dimensiones clínicas y sociodemográficas en modelos avanzados para una detección efectiva de la MME.

8.3.4.1.1.4. Mapa de calor según correlación de Pearson:

Como parte del análisis exploratorio y antes de realizar la selección definitiva de características, se construyó un mapa de calor de correlación de Pearson (véase la Figura 7) entre las variables independientes (explicativas) del dataset. Esta herramienta permitió detectar colinealidades, relaciones internas entre atributos y posibles redundancias que podrían afectar negativamente el rendimiento del modelo predictivo.

8.3.4.1.1.5. Correlaciones muy altas (colinealidad fuerte)

NUMEROS_PARTOS_CESARIAS ↔ VIVOS = 0.88
NUMEROS_PARTOS_CESARIAS ↔ EDAD = 0.47
VIVOS ↔ EDAD = 0.43
EDAD ↔ MAYOR_35 = 0.60

Estas correlaciones altas nos advierten que incluir ambos campos en el modelo puede introducir redundancia y sobreajuste, especialmente en modelos sensibles a la multicolinealidad (como regresión logística).

8.3.4.1.1.6. Correlaciones moderadas

NUMEROS_CONTROLES_PRENATALES ↔ NUMEROS_PARTOS_CESARIAS = 0.27
IMC ↔ EDAD = 0.27

Estas relaciones no son problemáticas, pero sí nos indican que existen asociaciones lógicas entre variables clínicas y demográficas que pueden reforzar patrones en el modelo sin ser redundantes.

8.3.4.1.1.7. Correlaciones bajas o nulas

La mayoría de las variables (como AFIC_GRUPO_ETNICO, TIPO_DE_CASO, MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA, etc.) presentan correlaciones bajas (< 0.10) con otras variables.

Esto indica que estas variables pueden aportar información complementaria, aunque su valor predictivo individual sea bajo.

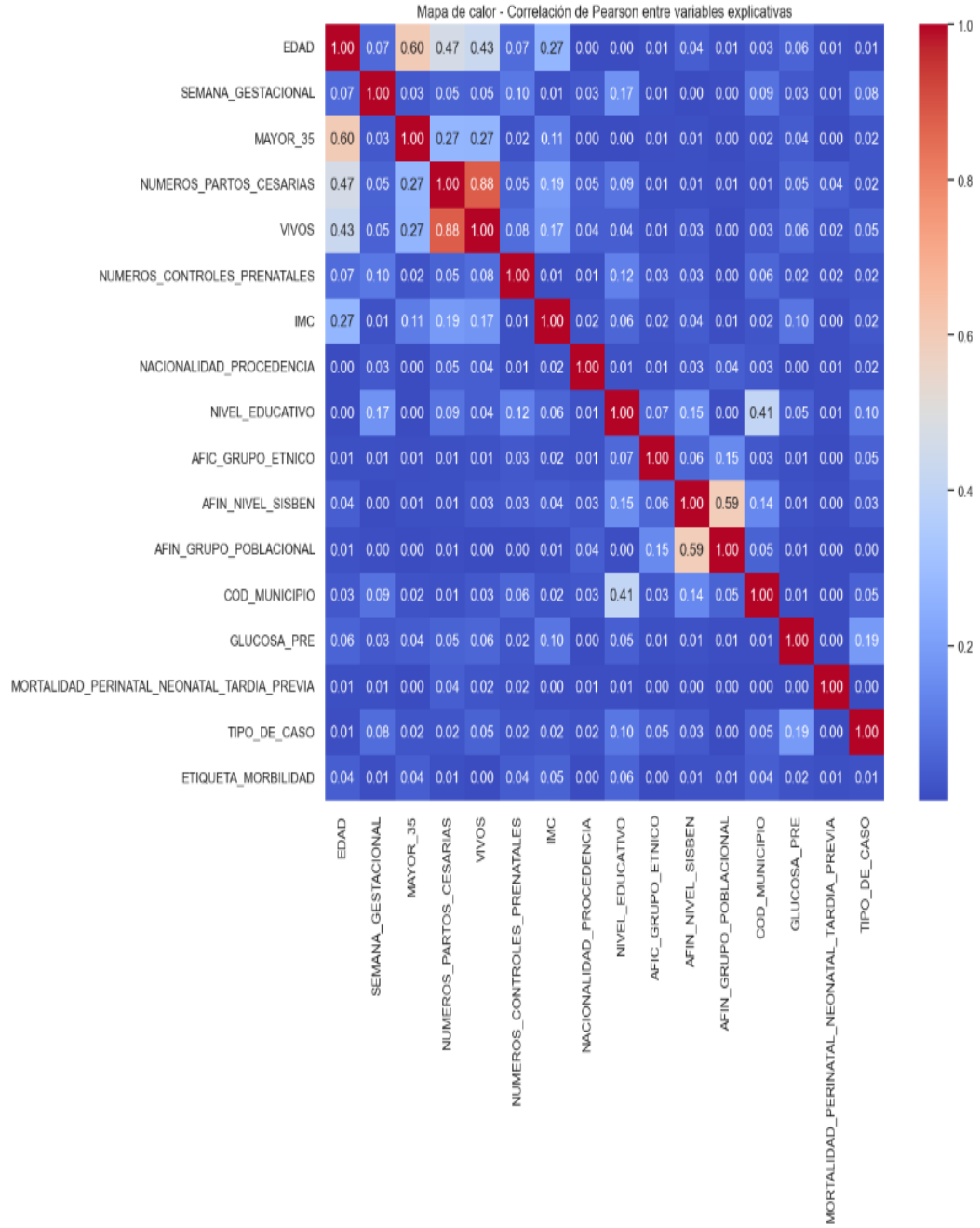
En general, la baja colinealidad entre la mayoría de las variables nos da espacio para trabajar con modelos que se beneficien de diversidad estructural.

8.3.4.1.1.8. Variables con colinealidad interna no útil

AFIN_NIVEL_SISBEN y AFIN_GRUPO_POBLACIONAL → 0.59
AFIN_NIVEL_SISBEN ↔ NIVEL_EDUCATIVO → 0.41

Estas variables parecen capturar información socioeconómica similar. En la selección final consideramos conservar solo una de ellas, basándome en su importancia según RF

**Figura 7**
*Mapa de calor – Correlación de Pearson*




*Fuente: Elaboración propia*

### 8.3.5. Implicaciones para el Modelado Predictivo

El análisis exploratorio reveló características específicas que condicionan el enfoque metodológico:

- **Distribuciones Asimétricas**: Las variables de conteo obstétrico (NUMEROS_PARTOS_CESARIAS, VIVOS) requieren transformaciones logarítmicas o uso de modelos que manejen distribuciones no normales.
- **Desbalance de Clases**: La proporción 19:1 entre casos negativos y positivos de MME demanda implementación de técnicas de balanceamiento y métricas de evaluación específicas.
- **Valores Atípicos**: Se identificaron outliers o valores extremos que requieren tratamiento específico previo al entrenamiento del modelo en variables como GLUCOSA_PRE (valores biológicamente implausibles), SEMANA_GESTACIONAL (extremos que necesitan validación clínica), EDAD (outliers clínicamente relevantes que se conservan, pero deben ser considerados), IMC (numerosos outliers altos), y potencialmente en NUMEROS_PARTOS_CESARIAS y VIVOS (valores altos que necesitan revisión/truncamiento).
- **Variables Categóricas Dominantes**: El 67.8% de las gestantes pertenecen a las dos categorías más frecuentes en variables socioeconómicas, lo que puede limitar la capacidad discriminativa del modelo en subpoblaciones minoritarias.

### 8.3.6. Estrategias de Tratamiento de Datos

Para abordar los hallazgos identificados, se implementarán las siguientes técnicas:

#### 8.3.6.1. Tratamiento de Outliers:

Aplicación de técnicas como winsorización, truncamiento o exclusión justificada para variables como GLUCOSA_PRE, IMC, y revisión de extremos en SEMANA_GESTACIONAL, NUMEROS_PARTOS_CESARIAS, y VIVOS.

8.3.6.2. Balanceamiento de Clases:

- **SMOTE (Synthetic Minority Oversampling Technique)**: Generación sintética de casos MME mediante interpolación de características de casos minoritarios existentes, incrementando la muestra de 4.779 a aproximadamente 15.000 casos sintéticos.
- **Undersampling Estratificado**: Reducción controlada de casos no-MME manteniendo representatividad demográfica, disminuyendo de 84.213 a 25.000 casos seleccionados aleatoriamente.

8.3.6.3. Métricas de Evaluación Específicas:

- **Sensibilidad (Recall)**: Proporción de casos MME correctamente identificados del total de casos reales de MME. Crítica dada la naturaleza de la condición.
- **Especificidad**: Proporción de casos no-MME correctamente clasificados, importante para evitar sobre-diagnóstico.
- **F1-score**: Media armónica entre precisión y sensibilidad, balanceando ambos aspectos de rendimiento.
- **AUC-ROC**: Área bajo la curva ROC que cuantifica la capacidad discriminativa del modelo independientemente del umbral de decisión.
- **AUC-PR**: Área bajo la curva Precisión-Recall, más informativa que AUC-ROC en contextos de clases desbalanceadas.

8.3.6.4. Tratamiento de la correlación:

La baja correlación lineal general refuerza que la MME no depende de un único factor, sino de la interacción de múltiples variables. Por ello, será clave emplear modelos capaces de capturar relaciones no lineales y patrones complejos como Random Forest o redes neuronales.

## 8.4. Selección de Características

La selección de características constituye una etapa fundamental en el desarrollo de modelos predictivos, especialmente en el contexto de la MME, donde múltiples factores clínicos, demográficos y sociales pueden influir en los resultados. Para garantizar la robustez y validez del modelo predictivo, se implementaron dos enfoques complementarios de selección de variables: el test Chi-cuadrado para evaluar asociaciones estadísticas entre variables categóricas (item 8.4.1), y el análisis de importancia mediante RF (item 8.4.2) para identificar patrones complejos y no lineales.

El test Chi-cuadrado permite evaluar la independencia entre variables categóricas y la variable objetivo, proporcionando una medida objetiva de asociación que no asume relaciones lineales. Por su parte, RF ofrece una perspectiva basada en la capacidad predictiva real de cada variable dentro de un conjunto de árboles de decisión, capturando interacciones y patrones no lineales que métodos univariados podrían omitir.

La implementación de ambas metodologías busca triangular los hallazgos, validar hipótesis clínicas previas e identificar variables relevantes que pudieran no ser evidentes desde una sola perspectiva analítica. Esta aproximación dual permite una selección más informada y robusta de las características que formarán parte del modelo predictivo final

### 8.4.1. Test Chi-Cuadrado

Como parte del proceso de selección de variables con enfoque estadístico, se aplicó la prueba de Chi-cuadrado (véase la Tabla 7) para evaluar la asociación entre variables categóricas y la variable objetivo ETIQUETA_MORBILIDAD.

Los resultados revelan que COD_MUNICIPIO emerge como la variable con mayor asociación a la morbilidad, lo cual podría reflejar condiciones territoriales diferenciadas, variaciones en la infraestructura médica o disparidades en el acceso a servicios de salud según la región. Este hallazgo resulta relevante para futuras segmentaciones por área geográfica en el análisis de morbilidad materna extrema.

Las variables DIAGNOSTICOS e HIPERTENSION muestran puntajes elevados, confirmando su relevancia clínica como predictores de morbilidad. Esta asociación valida la inclusión de variables relacionadas con antecedentes médicos y comorbilidades en el modelo predictivo.

NIVEL_EDUCATIVO se posiciona entre las variables más asociadas, reafirmando la importancia de la dimensión socioeconómica como factor determinante en los resultados de salud materna. Este resultado es consistente con la literatura científica que documenta las desigualdades en salud según nivel educativo.

Las variables IMC y HEMOGLOBINA presentan asociaciones significativas, validando su inclusión desde la perspectiva de la salud física y el estado nutricional de la gestante. Estos indicadores antropométricos y hematológicos son reconocidos factores de riesgo en la literatura obstétrica.

Es importante señalar que DOCUMENTO aparece en el ranking con una asociación considerable, lo cual constituye una anomalía metodológica, ya que esta variable no debería tener influencia clínica sobre la morbilidad. Este hallazgo sugiere un posible error en la codificación de la variable o una correlación espuria que requiere investigación adicional.

El análisis mediante Chi-cuadrado permitió confirmar hipótesis clínicas y sociales previamente establecidas, proporcionando una guía objetiva para la priorización de variables en el proceso de modelado. Aunque algunas variables no presentan correlaciones lineales fuertes, esta técnica detectó asociaciones significativas que no son capturadas por métodos como Pearson o Spearman.

**Tabla 7**
*Resultados de Chi- Cuadrado*[13]

| Variable | Chi2_Score |
|---|---|
| COD_MUNICIPIO | 2,453807e+06 |
| DIAGNOSTICOS | 1,595540e+04 |
| NIVEL_EDUCATIVO | 5,042337e+02 |
| FECHA_GLUCOSA | 1,558160e+02 |
| IMC | 1,056242e+02 |
| DOCUMENTO | 9,732404e+01 |
| HIPERTENSION | 7,904757e+01 |
| HEMOGLOBINA | 6,443793e+01 |
| MAYOR_35 | 6,064847e+01 |
| EDAD | 5,262430e+01 |

*Fuente:* Elaboración Propia

### 8.4.2. Importancia de Variables según RF

Con el objetivo de identificar las variables más relevantes para la predicción de morbilidad MME, se entrenó un modelo de R F y se analizó la importancia relativa de cada atributo (resultados en el Tabla 8). Esta técnica es particularmente útil para detectar patrones no lineales y combinaciones de variables que influyen en la predicción, sin asumir formas funcionales específicas.

Los resultados muestran que IMC, GLUCOSA_PRE, HEMOGLOBINA y EDAD presentan alta importancia, validando la hipótesis clínica de que los factores físicos y metabólicos constituyen determinantes significativos en la probabilidad de desarrollar MME. Estos hallazgos son consistentes con el conocimiento obstétrico establecido.

La relevancia de variables temporales como FUM, FPP y FECHA_HB resulta coherente desde la perspectiva clínica, ya que reflejan momentos específicos del embarazo en los cuales pueden manifestarse complicaciones. Sin embargo, su implementación en modelos de producción podría requerir estandarización temporal o la derivación de nuevas variables como "semanas transcurridas desde el último control".

[13] Los resultados < 0,05 existe evidencia para afirmar que sí existe asociación, mientras que los resultados ≥0,05 no hay evidencias suficientes para afirmar que exista asociación.

La alta importancia asignada a DOCUMENTO constituye una alerta metodológica significativa. Es probable que esta variable esté funcionando como un identificador codificado incorrectamente o que mantenga una vinculación indirecta con otras variables del modelo. Se recomienda su eliminación del dataset final o una revisión exhaustiva de su origen y codificación.

La inclusión de COD_MUNICIPIO entre las variables importantes reafirma los hallazgos del análisis Chi-cuadrado, confirmando que la dimensión geográfica ejerce influencia en los resultados, ya sea por diferencias en el acceso a servicios de salud o por condiciones poblacionales específicas.

En conclusión, el análisis de importancia mediante RF validó hallazgos previos y reveló variables relevantes adicionales. Este enfoque demostró que el modelo no se fundamenta exclusivamente en una dimensión única (edad o antecedentes clínicos), sino que integra información demográfica, clínica y temporal para construir sus predicciones, reflejando la naturaleza multifactorial de la morbilidad materna extrema.

**Tabla 8**

*Top de 10 de las variables identificadas según RF*[14]

| Variable | Importancia_RF |
|---|---|
| IMC | 0.092115 |
| DOCUMENTO | 0.088294 |
| FPP | 0.084889 |
| FUM | 0.083295 |
| GLUCOSA_PRE | 0.072913 |
| SEMANA_GESTACIONAL | 0.065304 |
| EDAD | 0.062352 |
| COD_MUNICIPIO | 0.053025 |
| HEMOGLOBINA | 0.051793 |
| FECHA_HB | 0.051498 |

*Fuente:* Elaboración propia

[14] A valores más altos son variables relevantes, a valores menores y cercanas a cero son valores que carecen de influencia.

8.4.3. Importancia de Variables según RF

### 8.4.4. **Variables Predictoras Identificadas**

Las variables predictoras principales, basándose en la consistencia y en los métodos de selección para modelado RF y Chi – Cuadrado son:

- IMC
- EDAD
- HEMOGLOBINA
- COD_MUNICIPIO
- GLUCOSA_PRE
- FUM
- FPP
- SEMANA_GESTACIONAL
- FECHA_HB

La variable DOCUMENTO debe ser excluida o investigada a fondo antes de cualquier inclusión en un modelo productivo. Las variables como DIAGNOSTICOS, HIPERTENSION y NIVEL_EDUCATIVO son importantes conceptualmente y para análisis exploratorios, pero el conjunto de RFE es el más depurado para un modelo predictivo según estos análisis.

## 8.5. Estado del Arte

En los últimos años, diversos estudios han empleado técnicas de ML para predecir complicaciones graves durante el embarazo, como la MME, incluyendo la hemorragia posparto y la preeclampsia. Estos modelos se han desarrollado a partir de datos clínicos, notas médicas y registros electrónicos de salud, con el objetivo de identificar a mujeres con alto riesgo de presentar desenlaces adversos.

En este contexto, el presente apartado tiene como propósito analizar dichas aproximaciones y proponer una metodología adecuada para la implementación de un modelo predictivo de MME, adaptado a la realidad de la EPS MUTUAL SER ESS y basado en los datos analizados. La Tabla 9 que se presenta a continuación muestra un análisis del impacto predictivo de ciertos algoritmos identificados en diversos estudios, con el fin de reconocer patrones útiles que permitan mejorar la detección de casos aprendidos y, con ello, optimizar su aplicación en el ámbito médico, especialmente en la prevención y manejo de la MME.

**Tabla 9**

*Matriz de estado de Arte por Algoritmos usados y su impacto en la predicción en casos de afectaciones clínicas en el parto materno*

| Algoritmo | Descripción breve | Referencias |
|---|---|---|
| Árboles de decisión potenciados por gradiente | • **Población**: Mujeres de 18 a 55 años que dieron a luz en un centro académico.<br>• **Tamaño de muestra:** 308.667 Ubicación: Estados Unidos<br>• **Métodos**: Desarrollo y validación de modelos de aprendizaje automático.<br>• **Resultados:** El modelo predijo hemorragia posparto con una precisión del 98,1% y sensibilidad de 0,763. | (Westcott, y otros, 2021) |
| Random Forest (RF) | • **Población:** Mujeres que se sometieron a un parto vaginal.<br>• **Tamaño de muestra:** 9.894 Ubicación:<br>• Japón Métodos: Desarrollo y comparación de modelos.<br>**Resultados:** Predicción de hemorragia posparto con un AUC de 0,708. | (Akazawa, Hashimoto, Katsuhiko, & Kaname, 2021) |
| Aumento de gradiente estocástico (Gradient Boosting) | • **Población:** Mujeres embarazadas que reciben atención prenatal.<br>• **Tamaño de muestra:** 11.006<br>• **Ubicación:** Corea<br>• **Resultados:** El modelo predijo preeclampsia de inicio tardío con una precisión del 92% y una sensibilidad del 97%. | (Jhee, y otros, 2019) |
| Modelo Aditivo Generalizado (GAM) | • **Población:** Mujeres con trabajo de parto único a término (≥37 semanas de gestación).<br>• **Ubicación:** Estados Unidos | (Lengerich, y otros, 2024) |

| Algoritmo | Descripción breve | Referencias |
|---|---|---|
| Árboles de decisión potenciados por gradiente | • **Resultados**: El modelo identificó casi el doble de casos de hemorragia posparto en comparación con la evaluación de riesgo estándar.<br>• **Población:** Mujeres con partos a término, únicos y en presentación cefálica**.**<br>• **Tamaño de muestra:** 98.463<br>• **Métodos:** Estudio de cohorte retrospectivo con desarrollo de modelo de aprendizaje automático**.**<br>• **Resultados:** El modelo predijo el riesgo de lesión del esfínter anal obstétrico con un AUC de 0,756, utilizando datos al ingreso**.** | (Chill, y otros, 2021) |
| Red neuronal multilabel | • **Población:** Personas embarazadas con trastornos hipertensivos del embarazo que tuvieron un parto hospitalario entre el 1 de octubre de 2015 y el 31 de diciembre de 2020, según registros de la Premier Healthcare Database.<br>• **Tamaño de muestra:** 553.568<br>• **Ubicación:** Estados Unidos<br>• **Resultados:** Capacidad predictiva general AUC = 0,85; en eventos cardiovasculares AUC = 0,94. | (Meng, y otros, 2024) |
| Regresión logística (RL) | • **Población:** Pacientes atendidas en la E.S.E Clínica de Maternidad Clavo Castaño de Cartagena, con y sin diagnóstico de MME.<br>• **Tamaño de muestra:** 1.388 pacientes (2015–2016).<br>• **Ubicación:** Cartagena, Colombia.<br>• **Resultados:** Precisión del 51,8% y sensibilidad del 97,7%. | (Rodríguez, 2017) |

*Fuente:* Elaboración propia.

Las investigaciones recientes demuestran que los algoritmos de *ML*, especialmente los árboles de decisión potenciados, el modelo RF y la RL, son herramientas prometedoras para predecir la MME a partir de datos clínicos y registros electrónicos. Estos modelos pueden superar en precisión y sensibilidad a los métodos estadísticos tradicionales, facilitando la identificación temprana de mujeres en riesgo.

Por su facilidad de implementación, su capacidad de adaptación a distintos tipos de datos y la naturaleza del problema de estudio se optó por aplicar y comparar los algoritmos RF y RL. Evaluarlos en paralelo permite analizar cuál de los dos ofrece mejor desempeño predictivo, especialmente frente al fuerte desbalance de clases que caracteriza la MME. Además, esta estrategia posibilita verificar que los resultados obtenidos no dependan exclusivamente de un tipo de modelo, y facilita la detección de posibles problemas de sobreajuste (overfitting) o subajuste (underfitting) en los modelos evaluados.

### 8.6. Metodología para la Implementación del Proyecto MME

En el Anexo 2 Metodología propuesta, se presenta la metodología propuesta basada en el análisis del estado de arte y el resultado de la selección de las características en el conjunto de datos analizados

## 9. Recomendaciones

- Selección y modelado de variables: Utilizar modelos que capturen relaciones no lineales complejas (RL, RF, Redes Neuronales, etc.) combinando métodos estadísticos y de ML para la selección de variables. Excluir variables con posible contaminación metodológica e incorporar variables geográficas para ajustar modelos a contextos territoriales específicos.
- Preprocesamiento clínico: Aplicar técnicas sensibles al contexto clínico, combinando transformaciones estadísticas (winsorización) con revisiones clínicas para evitar pérdida de información significativa o malinterpretación de valores extremos.
- Balanceo de datos: Implementar estrategias combinadas como SMOTE para ampliar la clase minoritaria y undersampling[15] estratificado para reducir la clase mayoritaria, asegurando diversidad y representatividad en los datos.

[15] Undersampling es una técnica usada en análisis de datos y machine learning para tratar conjuntos de datos desbalanceados. Consiste en reducir el tamaño de la clase mayoritaria eliminando muestras, con el objetivo de equilibrar la cantidad de ejemplos entre las clases y evitar que el modelo se sesgue hacia la clase más frecuente.

- Evaluación del modelo: Adoptar métricas centradas en la clase positiva, priorizando sensibilidad, F1-score y AUC-PR, especialmente para herramientas de alerta clínica o apoyo a la decisión médica.
- Validación robusta: Implementar validaciones cruzadas estratificadas y segmentadas por región o institución para verificar la estabilidad del modelo frente a distintas distribuciones geográficas y sociodemográficas.
- Mantenimiento continuo: Establecer protocolos periódicos de revisión que incluyan alertas automáticas para valores fuera de rango, participación de profesionales de salud y revisión de la relevancia de variables para garantizar la calidad y vigencia del sistema predictivo.

## 10. Lecciones Aprendidas

La experiencia de este proyecto evidenció que la articulación entre ciencia de datos, conocimiento clínico y gestión en salud pública es fundamental para desarrollar soluciones robustas, contextualizadas y sostenibles. La colaboración continua entre actores técnicos, clínicos y gestores permitió abordar integralmente los desafíos técnicos, éticos y sociales del problema, enriqueciendo la toma de decisiones.

Asimismo, se confirmó que un enfoque preventivo basado en análisis de datos no solo es técnicamente factible, sino estratégicamente necesario. La implementación de modelos predictivos facilita una transición efectiva desde una atención reactiva hacia una gestión proactiva del riesgo, lo que contribuye a reducir complicaciones severas y a optimizar los recursos del sistema de salud. Este cambio de paradigma representa un avance clave en el fortalecimiento de la salud pública.

## 11. Conclusiones

- La implementación de modelos predictivos para la MME representa un avance estratégico fundamental en la prevención de salud pública materna. Los algoritmos RF y RL

demostraron fortalezas complementarias que construyen un marco robusto aplicable en contextos clínicos reales.

- Los modelos predictivos tienen el potencial de transformar la detección y manejo de riesgos durante el embarazo. La capacidad de anticipar complicaciones severas permite una asignación más eficiente de recursos sanitarios y una intervención oportuna, contribuyendo a la reducción significativa de la mortalidad y morbilidad materna.
- Desbalance crítico de clases (19:1) entre casos positivos y negativos de MME, que exige técnicas avanzadas de balanceo para evitar sesgos hacia la clase mayoritaria.
- Variables predictoras robustas identificadas: IMC, GLUCOSA_PRE, HEMOGLOBINA, EDAD y COD_MUNICIPIO capturan relaciones no lineales características de la MME multifactorial.
- Presencia de valores atípicos clínicamente relevantes en variables como GLUCOSA_PRE, IMC y SEMANA_GESTACIONAL requieren tratamiento con criterios mixtos (estadísticos y clínicos).
- La variable DOCUMENTO presenta efecto espurio y debe excluirse para evitar sesgos en el modelo.
- La experiencia demuestra que, pese a los desafíos del procesamiento de datos clínicos, la inteligencia artificial puede superar barreras tradicionales en la toma de decisiones clínicas complejas. La colaboración interdisciplinaria entre profesionales de salud, expertos en tecnología y formuladores de políticas públicas es esencial para maximizar el impacto sostenible de estas soluciones.

## 12. Bibliografía

1. Akazawa, M., Nishigori, H., Yokomichi, H., Mizobuchi, Y., & Yamaguchi, S. (2021). Machine learning prediction model for postpartum hemorrhage. *Scientific Reports, 11*(1), 13238. https://doi.org/10.1038/s41598-021-92584-6
2. Amazon Web Services. (2024). ¿En qué consiste el ajuste de hiperparámetros? *Amazon Web Services*. https://aws.amazon.com/es/what-is/hyperparameter-tuning/

3. Bauserman, M., Conroy, A. L., Binzen, S., & Zash, R. (2021). Maternal morbidity and mortality in low- and middle-income countries: A systematic review. *The Lancet Global Health, 9*(6), e755–e765. https://doi.org/10.1016/S2214-109X(21)00123-8
4. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*, 281–305.
5. Campaña-Bastida, S. E., Delgado-Pérez, A. E., & Salas-Villalba, G. L. (2024). *Sistema tecnológico con IA para la prevención de la ocurrencia de casos de morbilidad materna extrema*. Universidad Nacional Abierta y a Distancia.
6. Comisión Económica para América Latina y el Caribe. (2021). *Desigualdades en salud materna en América Latina y el Caribe*. https://www.cepal.org
7. Chill, H. H., Zafran, N., Shinar, S., Hadar, E., & Bardin, R. (2021). Prediction model for obstetric anal sphincter injuries using machine learning. *International Urogynecology Journal, 32*(9), 2393–2399. https://doi.org/10.1007/s00192-021-04761-
8. DataScientest. (2024). *Cross-validation: definición e importancia en machine learning*. https://datascientest.com/es/cross-validation-definicion-e-importancia
9. Díaz, R. (2025). *Métricas de clasificación*. TheMachineLearners. https://www.themachinelearners.com/metricas-de-clasificacion/
10. Fondo de Población de las Naciones Unidas (UNFPA). (2015). *Quinto Objetivo de Desarrollo del Milenio: Mejorar la salud materna*. https://www.unfpa.org/es/quinto-objetivo-de-desarrollo-del-milenio-mejorar-la-salud-materna
11. García, J., Pérez, L., & Ramos, D. (2020). Machine learning aplicado a la predicción de morbilidad materna. *Revista de Salud Pública, 22*(3), 45–60. https://doi.org/10.15446/rsap.v22n3.88334
12. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1), 389–422. https://doi.org/10.1023/A:1012487302797
13. Instituto Nacional de Salud (INS). (2020). *Morbilidad Materna Extrema en Colombia: Informe 2020*. https://www.ins.gov.co
14. Instituto Nacional de Salud (INS). (2021). *Informe anual de morbilidad materna extrema en Colombia*. https://www.ins.gov.co

15. Instituto Nacional de Salud (INS). (2022). *Informe sobre salud materna en Colombia: Mortalidad materna*. https://www.ins.gov.co
16. Instituto Nacional de Salud (INS). (2023). *Boletín epidemiológico semana 30: Comportamiento de la morbilidad extrema en Colombia*. https://www.ins.gov.co
17. Instituto Nacional de Salud (INS). (2024). *Portal SIVIGILA: Notificaciones de eventos 2024–2025*. https://portalsivigila.ins.gov.co/
18. Jhee, J. H., Lee, S., Park, Y., et al. (2019). Prediction model development of late-onset preeclampsia using machine learning-based methods. *PLoS ONE, 14*(4), e0213430. https://doi.org/10.1371/journal.pone.0213430
19. Lengerich, B., Arnesen, D., Choi, E., & Lasko, T. A. (2024). Prediction of obstetric risk using interpretable machine learning models. *American Journal of Obstetrics & Gynecology MFM*, 101391. https://doi.org/10.1016/j.ajogmf.2024.101391
20. Londoño, F., Herrera, A., & Guzmán, M. (2023). Big Data y salud materna: Modelos predictivos para la prevención. *Journal of Artificial Intelligence in Medicine, 15*(2), 88–102. https://doi.org/10.1016/j.jartmed.2023.04.006
21. Meng, M., Zhang, J., Liu, Y., & Li, Q. (2024). Predictive model for cardiovascular morbidity in pregnant women. *Anesthesia and Analgesia*. https://doi.org/10.1213/ANE.0000000000006537
22. Ministerio de Salud y Protección Social. (2023). *Atención en salud materna en Colombia*. https://www.minsalud.gov.co
23. Ministerio de Salud y Protección Social. (2024). *Reducción del 26% en mortalidad materna*. https://www.minsalud.gov.co
24. Muñoz, C. A., Pérez, R., & Pardo, L. (2013). Caracterización de la Mortalidad Materna en Bolívar. *Revista de Ciencias Biomédicas, 4*(2), 247–255.
25. Muñoz, R., García, P., & Ramírez, L. (2013). Barreras en la atención materna en Colombia. *Revista de Salud Pública, 15*(2), 155–169. https://doi.org/10.15446/rsap.v15n2.38811
26. Narváez Díaz, J., & Caro Caro, L. (2024). *Complicaciones obstétricas en Colombia: Vigilancia y prevención*. Universidad Nacional de Colombia.

27. Narváez Díaz, N. S., & Caro Caro, J. E. (2024). *Protocolo de vigilancia en salud pública: Morbilidad Materna Extrema (MME)*. Instituto Nacional de Salud. https://www.ins.gov.co
28. Nair, M., Kurinczuk, J. J., Knight, M., & Brocklehurst, P. (2016). Establishing a national maternal morbidity outcome indicator in England. *PLoS ONE, 11*(4), e0153370. https://doi.org/10.1371/journal.pone.0153370
29. Organización Mundial de la Salud (OMS). (2019). *Manejo de la morbilidad materna extrema*. https://www.who.int
30. Organización Mundial de la Salud (OMS). (2022). *Global strategy for women's, children's and adolescents' health (2022–2030)*. https://www.who.int
31. Organización Mundial de la Salud (OMS). (2023). *Aceleración hacia las metas de los ODS para salud materna*.
32. Ortiz Lizcano, E. I., Ortega Cifuentes, G. A., & Martínez Pérez, M. A. (2010). *Vigilancia de la morbilidad materna extrema*. Editorial Legis.
33. Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.
34. Rodríguez, E. L. (2017). *Predicción temprana de morbilidad materna extrema usando árboles de decisión* [Tesis de pregrado, Universidad Tecnológica de Bolívar].
35. Say, L., Chou, D., Gemmill, A., Tunçalp, Ö., Moller, A. B., Daniels, J., … Alkema, L. (2020). Global causes of maternal death: A WHO systematic analysis. *The Lancet Global Health, 2*(6), e323–e333. https://doi.org/10.1016/S2214-109X(14)70227-X
36. Souza, J. P., Gülmezoglu, A. M., Vogel, J., Carroli, G., Lumbiganon, P., Qureshi, Z., … Say, L. (2021). Predicting maternal morbidity and mortality: A machine learning approach. *BMC Pregnancy and Childbirth, 21*(1), 121–130. https://doi.org/10.1186/s12884-021-03669-3
37. IBM. (2023). *¿Qué es la regulación (regularization)?* https://www.ibm.com/mx-es/think/topics/regularization
38. Westcott, J. C., Wilson, L., Roth, C., & Griffiths, R. (2021). Machine learning-based prediction of maternal hemorrhage. *Journal of Medical Internet Research, 24*(1), e22150. https://doi.org/10.2196/22150

39. Instituto Nacional de Salud (INS). (2021). *Comportamiento de la vigilancia en salud pública: Boletín epidemiológico semana 8*. https://www.ins.gov.co

## 13. Anexos

1. Anexo 1 Análisis exploratorio estadístico
2. Anexo 2 Metodología propuesta