ANEXO 2

METODOLOGÍA PARA EL DESARROLLO DE MODELOS PREDICTIVOS DE MORBILIDAD MATERNA EXTREMA (MME) EN MUJERES EMBARAZADAS DEL CARIBE COLOMBIANO

CONTENIDO

TABLAS

RESUMEN EJECUTIVO

Esta metodología constituye el producto principal de una investigación que procesó **88.810 registros** de mujeres embarazadas afiliadas a la EPS MUTUAL SER de cinco departamentos del Caribe colombiano (Bolívar, Córdoba, Atlántico, Sucre y Magdalena), identificando **4.395 casos de MME** (4.95% de prevalencia) datos recolectados entre noviembre 2019 a corte de febrero de 2025. A través de técnicas de Machine Learning, análisis exploratorio exhaustivo y procesos de limpieza y selección de características, se desarrolló un framework[1] metodológico validado empíricamente para la predicción de Morbilidad Materna Extrema.

1. INTRODUCCIÓN

1.1 Contexto de la Investigación Realizada

La investigación se ejecutó utilizando una base de datos robusta de **88.810 registros** provenientes de la EPS MUTUAL SER ESS y consultas realizadas a SIVIGILA, abarcando cinco departamentos estratégicos del Caribe colombiano. Esta población representa una muestra significativa y geográficamente diversa que permitió desarrollar una metodología integral para el desarrollo de un aplicativo que logrará mayor predicción de MME

1.2 Caracterización poblacional procesada:

En la Tabla 1 se describe el comportamiento de los casos identificados de MME en la base de datos consultada

[1] Un framework metodológico es un conjunto organizado de principios, métodos y procesos que guía la planificación, ejecución y evaluación de una investigación o proyecto, asegurando coherencia y rigor en el desarrollo.

Tabla 1

Caracterización de Datos Departamentos Bolívar, Córdoba, Atlántico, Sucre y Magdalena

Departamento	Total Registros	Casos MME	Casos No-MME	Prevalencia MME
Bolívar	30.266	1.489	28.777	4,92%
Córdoba	21.112	955	20.157	4,52%
Atlántico	16.561	1.159	15.402	7,00%
Sucre	11.219	345	10.874	3,07%
Magdalena	9.652	447	9.205	4,63%
TOTAL	**88.810**	**4.395**	**84.415**	**4,95%**

Fuente: *Elaboración Propia*

2. FUNDAMENTACIÓN METODOLÓGICA VALIDADA

2.1. Enfoque de Investigación Aplicado

La metodología desarrollada se basa en un enfoque cuantitativo aplicado con validación empírica a gran escala, integrando técnicas avanzadas de machine learning con conocimiento epidemiológico específico del Caribe colombiano.

2.2. Selección y Validación de Algoritmos

Basándose en el análisis de 88.810 registros y considerando las características específicas del dataset (desbalance de clases 4,95% vs 95,05%), se validaron dos algoritmos óptimos:

2.2.1. Random Forest (RF) - Modelo de Alta Precisión

- Validado empíricamente con datos masivos del Caribe colombiano
- Es un Algoritmo que demuestra robustez ante el desbalance natural de clases en este caso con los datos analizados de casos positivos de MME
- Capacidad superior para capturar patrones geográficos complejos

2.2.2. Regresión Logística (RL) - Modelo Interpretativo

- Validado en contexto de alta variabilidad geográfica
- Permite cuantificación de factores de riesgo específicos por departamento
- Facilita interpretación clínica en contextos de atención primaria

Evaluar ambos modelos permite identificar cuál ofrece mejor desempeño predictivo y garantiza que los hallazgos no dependan de un único enfoque, contribuyendo a evitar problemas de sobreajuste o subajuste.

3. METODOLOGÍA DESARROLLADA

3.1. Arquitectura Metodológica Implementada

La metodología se estructura en cuatro fases secuenciales validadas[2]:

- Fase I: Ingeniería de Datos
- Fase II: Desarrollo y Optimización
- Fase III: Validación e Interpretación
- Fase IV: Interpretabilidad y Aplicabilidad Clínica

A continuación, se detallan las variables predictoras seleccionadas con base en técnicas de Chi-cuadrado y RF, aplicadas al conjunto de datos previamente mencionado.

3.2. Resultados de Selección Variables

3.2.1. Variables predictoras identificadas

3.2.1.1. Variables biomédicas:

[2] Tenga en cuenta que las fases de análisis exploratorio de datos (EDA), limpieza de datos y selección de características ya han sido adelantadas previamente y no constituyen la etapa de desarrollo

- **IMC**: Índice de masa corporal materno - Validado como predictor robusto en las 5 poblaciones departamentales
- **HEMOGLOBINA**: Niveles de hemoglobina durante el embarazo - Consistente poder predictivo a nivel regional
- **GLUCOSA_PRE**: Niveles de glucosa preprandial - Relevancia confirmada en análisis multivariado

3.2.1.2. Variable demográfica:

- **EDAD**: Edad materna - Factor de riesgo validado con variaciones geográficas específicas

3.2.1.3. Variables temporales y gestacionales:

- **FUM**: Fecha de última menstruación - Fundamental para cálculo de edad gestacional
- **FPP**: Fecha probable de parto - Variable de seguimiento obstétrico
- **SEMANA_GESTACIONAL**: Edad gestacional en semanas - Predictor crítico validado
- **FECHA_HB**: Fecha de toma de hemoglobina - Indicador de seguimiento prenatal

3.2.1.4. Variable geográfica:

- **COD_MUNICIPIO**: Código del municipio - Captura variabilidad geográfica significativa identificada

3.2.2. Exclusiones

3.2.2.1. Variables excluidas del modelo predictivo

- **DOCUMENTO**: Excluida por riesgo de sobreajuste y ausencia de poder predictivo generalizable
- **DIAGNOSTICOS**: Relevante conceptualmente, pero con menor poder predictivo que variables biomédicas

- **HIPERTENSION**: Importante para análisis descriptivos, no incluida en modelo final predictivo
- **NIVEL_EDUCATIVO**: Significativa sociológicamente, menor impacto predictivo individual

3.2.3. Criterios de exclusión aplicados

1. Poder predictivo validado mediante métricas de importancia
2. Generalización a nivel regional (5 departamentos)
3. Disponibilidad y confiabilidad de datos
4. Relevancia clínica versus complejidad operacional

3.3. Fase I: Ingeniería de Datos Específica

Tratamiento optimizado del conjunto de 9 variables seleccionadas

3.3.1. Procesamiento por Tipo de Variable

3.3.1.1. Variables biomédicas (IMC, HEMOGLOBINA, GLUCOSA_PRE):

- Detección de outliers utilizando percentiles 1 y 99 por departamento
- Normalización Z-score[3] preservando variabilidad geográfica natural
- Imputación estratificada por edad materna y departamento

3.3.1.2. Variable demográfica (EDAD):

- Preservación de variable continua para capturar tendencias no lineales
- Creación de categorías de riesgo: <18, 18-19, 20-34, ≥35 años
- Análisis de distribución por departamento para identificar patrones regionales

[3] La normalización Z-score estandariza los datos restando la media y dividiendo por la desviación estándar, dejando los valores con media 0 y desviación 1, lo que facilita comparaciones y mejora modelos.

3.3.1.3. Variables temporales (FUM, FPP, FECHA_HB):

- Transformación a intervalos relativos al momento del parto
- Validación de coherencia temporal entre variables relacionadas
- Manejo de inconsistencias mediante algoritmos de corrección automática

3.3.1.4. Variable gestacional (SEMANA_GESTACIONAL):

- Validación de rangos clínicamente aceptables (12-42 semanas)
- Imputación basada en coherencia con FUM y FPP
- Categorización por trimestres para análisis complementarios

3.3.1.5. Variable geográfica (COD_MUNICIPIO):

- Preservación de granularidad municipal
- Agrupación por niveles de complejidad hospitalaria disponible
- Codificación que mantiene interpretabilidad geográfica

3.3.2. Pipeline de Transformación Validado

3.3.2.1. Secuencia de procesamiento optimizada:

1. Validación de integridad y consistencia
2. Tratamiento de valores faltantes por tipo de variable
3. Detección y corrección de outliers[4]
4. Normalización y estandarización apropiada
5. Validación final de distribuciones resultantes

[4] 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.

3.4. Fase II: Desarrollo y Optimización de Modelos

3.4.1. Protocolo de Entrenamiento de RL

3.4.1.1. Configuración optimizada para las 9 variables:

- Evaluación de multicolinealidad específica entre variables seleccionadas
- Regularización L2 optimizada para prevenir overfitting en dataset masivo
- Validación cruzada estratificada por departamento

3.4.1.2. Proceso de optimización

1. División estratificada manteniendo proporción de MME por departamento
2. Grid Search para optimización de hiperparámetros de regularización
3. Validación de estabilidad de coeficientes mediante bootstrap[5]
4. Análisis de significancia estadística con corrección por múltiples comparaciones

3.4.2. Protocolo de Entrenamiento de RF

3.4.2.1. Configuración específica para el conjunto de variables:

- Optimización de número de árboles considerando tamaño del dataset (88,810 registros)
- Ajuste de parámetros de profundidad y muestreo para prevenir overfitting
- Evaluación de importancia específica de las 9 variables seleccionadas

3.4.2.2. Proceso de entrenamiento:

1. Partición estratificada preservando distribución departamental
2. RandomizedSearchCV para optimización eficiente de hiperparámetros

[5] Bootstrap es una técnica de remuestreo que consiste en generar múltiples muestras con reemplazo a partir de un conjunto de datos original, para estimar la variabilidad de un modelo o estadístico sin necesidad de asumir distribuciones teóricas.

3. Evaluación de consenso entre árboles para cada una de las 9 variables
4. Validación de robustez mediante perturbaciones controladas del dataset

3.5. Fase III: Validación y Evaluación Integral

3.5.1. Framework de Evaluación para Datos Desbalanceados

3.5.1.1. Métricas adaptadas al contexto epidemiológico (4,95% prevalencia)

se insta a usar las siguientes métricas para evaluar la presión del modelo:

3.5.1.1.1. Métricas de discriminación:

- **AUC-ROC**: Evaluación de capacidad discriminativa general
- **AUC-PR**: Optimizada para clase minoritaria (MME)
- **Balanced Accuracy**: Ajustada por desbalance natural

3.5.1.1.2. Métricas clínicamente orientadas:

- **Sensibilidad**: Prioritaria para detección de MME (minimizar falsos negativos)
- **Especificidad**: Relevante para eficiencia operacional
- **Valor Predictivo Positivo**: Ajustado por prevalencia departamental
- **Número Necesario para Detectar (NND)**: Métrica de utilidad clínica

3.5.1.2. Validación Geográfica y Temporal

3.5.1.2.1. Estrategia de validación multicapa

1. **Validación por departamento**: Evaluación de desempeño en cada una de las 5 poblaciones
2. **Validación cruzada geográfica**: Entrenamiento en 4 departamentos, validación en 1

3. **Validación de robustez**: Evaluación ante variaciones en tamaño muestral
4. **Validación de generalización**: Análisis de transferibilidad entre departamentos

3.6. Fase IV: Interpretabilidad y Aplicabilidad Clínica

3.6.1. Interpretación de Modelos con Variables Específicas

3.6.1.1. Para R L

- Cálculo de Odds Ratios[6] para cada una de las 9 variables
- Análisis de intervalos de confianza por variable y departamento
- Interpretación clínica de magnitudes de efecto

3.6.1.2. Para RF:

- Ranking de importancia de las 9 variables seleccionadas
- SHAP values[7] para explicación de predicciones individuales
- Partial Dependence Plots[8] para variables biomédicas clave

3.7. CONSIDERACIONES ADICIONALES

3.7.1. Recomendaciones de Implementación Para Implementación Institucional En EPS MUTUAL SER ESS:

1. Piloto inicial en departamento con mayor prevalencia (Atlántico)
2. Integración gradual en sistemas de información existentes
3. Capacitación de personal en interpretación de las 9 variables clave

[6] Los Odds Ratios (OR) son una medida estadística que indica la fuerza de asociación entre una exposición (o variable) y un resultado. Un OR > 1 sugiere mayor riesgo, OR < 1 indica un efecto protector, y OR = 1 implica ausencia de asociación.
[7] SHAP values son una técnica que explica la contribución de cada variable a la predicción de un modelo, asignando un valor que indica cuánto influye esa variable en el resultado, facilitando la interpretabilidad de modelos complejos.
[8] Los Partial Dependence Plots (PDP) muestran cómo afecta una o dos variables específicas a la predicción promedio de un modelo, ayudando a entender la relación entre esas variables y el resultado.

4. Monitoreo continuo de desempeño y actualización periódica