



Para citaciones: Araújo Zárate, P., Contreras Chávez, J., y Martínez Lobo, D. (2024). Modelos predictivos en la clasificación de nacidos vivos y mortinatos: un estudio comparativo entre técnicas de *machine learning* y regresión logística en función de variables sociodemográficas y clínicas. *Revista Ciencias Biomédicas*, 13(4), 175-189.
<https://doi.org/10.32997/rcb-2024-4940>

Recibido: 14 de agosto de 2024
Aprobado: 10 de octubre de 2024
Publicado: 15 de octubre de 2024

Autor de correspondencia:
 Pedro Araújo Zárate
parauijo@unbosque.edu.co

Editor: Inés Benedetti. Universidad de Cartagena-Colombia.

Copyright: © 2024. Araújo Zárate, P., Contreras Chávez, J., y Martínez Lobo, D. Este es un artículo de acceso abierto, distribuido bajo los términos de la licencia <https://creativecommons.org/licenses/by-nc-nd/4.0/> la cual permite el uso sin restricciones, distribución y reproducción en cualquier medio, siempre y cuando el original, el autor y la fuente sean acreditados.



Modelos predictivos en la clasificación de nacidos vivos y mortinatos: un estudio comparativo entre técnicas de *machine learning* y regresión logística en función de variables sociodemográficas y clínicas

Predictive models in the classification of live births and stillbirths: a comparative study between machine learning and logistic regression techniques as a function of sociodemographic and clinical variables

Pedro Araújo Zárate¹, John Contreras Chávez¹ & Danny Martínez Lobo²

¹ MD, Candidato a Magister, Maestría en Estadística Aplicada y Ciencia de Datos, Universidad El Bosque, Bogotá, Colombia.

² Maestría en Estadística Aplicada y Ciencia de Datos, Universidad El Bosque, Bogotá, Colombia.

RESUMEN

Introducción: La mortalidad fetal continúa siendo un problema de salud pública en Colombia, que afecta significativamente el bienestar familiar y social. Es fundamental identificar los factores asociados y predecir el riesgo para implementar intervenciones efectivas.

Objetivo: Desarrollar un modelo estadístico para clasificar a las gestantes en riesgo de mortalidad fetal en Colombia durante el año 2022.

Métodos: Se realizó un estudio de casos y controles utilizando datos de nacidos vivos y defunciones fetales reportados por el Departamento Administrativo Nacional de Estadística. Se aplicaron técnicas de imputación de datos faltantes y balanceo de clases mediante el método SMOTE. Se evaluaron cuatro modelos de clasificación: regresión logística, K-Nearest Neighbors (KNN), árbol de decisión y máquina de soporte vectorial. El rendimiento de los modelos se comparó utilizando métricas de exactitud, sensibilidad, especificidad, puntaje F1 y precisión.

Resultados: El conjunto de datos final incluyó 566.806 registros, con 562.828 nacidos vivos y 3.978 muertes fetales. El modelo KNN presentó el mejor rendimiento, con una exactitud de 0,988, sensibilidad de 0,989, especificidad de 0,986 y puntaje F1 de 0,988. Los factores asociados significativamente con la probabilidad de nacer vivo incluyeron el número de hijos, el sexo, el área de residencia, el régimen de afiliación, las semanas de gestación, el peso al nacer, la edad y el nivel educativo de la madre.

Conclusión: El modelo KNN demostró ser efectivo en la predicción del riesgo de mortalidad fetal. Los resultados resaltan la importancia de factores socioeconómicos y clínicos en la supervivencia neonatal, sugiriendo la necesidad de intervenciones focalizadas para reducir las muertes fetales en Colombia.

Palabras Clave: Mortalidad fetal; Modelos estadísticos; Salud pública, Machine learning.

ABSTRACT

Introduction: Fetal mortality continues to be a public health problem in Colombia, significantly affecting family and social well-being. It is essential to identify associated factors and predict risk in order to implement effective interventions.

Objective: Develop a statistical model to classify pregnant women at risk of fetal mortality in Colombia during 2022.

Methods: A case-control study was conducted using data on live births and fetal deaths reported by the National Administrative Department of Statistics. Data imputation and class balancing techniques were applied using the smote method. Four classification models were evaluated: logistic regression, K-Nearest Neighbors (KNN), decision tree, and support vector machine. Model performance was compared using accuracy, sensitivity, specificity, F1 score, and precision metrics.

Results: The final dataset included 566,806 records, with 562,828 live births and 3,978 fetal deaths. The KNN model showed the best performance, with an accuracy of 0.988, sensitivity of 0.989, specificity of 0.986, and F1 score of 0.988. Factors significantly associated with the probability of live birth included the number of children, sex, area of residence, affiliation regime, gestational weeks, birth weight, and mother's age and educational level.

Conclusions: The KNN model proved effective in predicting the risk of fetal mortality. The results highlight the importance of socioeconomic and clinical factors in neonatal survival, suggesting the need for targeted interventions to reduce fetal deaths in Colombia.

Keywords: Fetal Death; Statistical Models; Public Health; Machine Learning.

INTRODUCCIÓN

Uno de los mayores avances de la obstetricia moderna ha sido la disminución de la tasa de mortinato, logrando una reducción casi de 10 veces con cuatro décadas atrás (1), no obstante, la mortalidad fetal continua como uno de los grandes problemas de salud pública, estimándose que el 57% de las muertes perinatales son fetales (2).

La defunción fetal es definida como muerte del producto de la concepción antes de su expulsión o extracción completa del cuerpo de la madre, independientemente de la duración del embarazo (3). La muerte está indicada por el hecho de que después de la separación, el feto no respira ni da ninguna otra señal de vida, como latidos del corazón, pulsaciones del cordón umbilical o

movimientos efectivos de los músculos de contracción voluntaria (3,4).

La mortalidad fetal intrauterina antes del parto afecta a 1.4 millones de casos anualmente, mientras que cerca de 1.2 millones de muertes fetales ocurren durante el trabajo de parto y el nacimiento. En América Latina, se reportó en 2015 una tasa de 18.4 muertes fetales por cada 1000 nacidos vivos, lo que representa una reducción notable en comparación con el año 2000, cuando la tasa era de 24.7. En Colombia, según el Departamento Administrativo Nacional de Estadística (DANE), se registraron 119,097 muertes fetales entre 2014 y 2016. Las tasas de mortalidad fetal fueron de 93.1 por cada 1000 nacidos vivos en madres adolescentes tempranas, 60.6 en adolescentes tardías y 68.8 en madres adultas (2), posterior al año 2016 se ha observado un comportamiento a la

disminución de los casos llegando hasta aproximadamente 49,4% menos en el 2023 (4).

A diferencia de la mortalidad materna y la mortalidad neonatal, la muerte fetal ha recibido poca atención (1,3). Esta falta de enfoque no ha permitido identificar las causas subyacentes de la mortalidad fetal con la finalidad de desarrollar intervenciones efectivas (2). La subestimación del problema, combinada con la escasez de visibilidad en las políticas de salud, dificulta la formulación de estrategias adecuadas para su prevención. Además, el estigma y el dolor asociados a la muerte fetal a menudo son invisibilizados, afectando el bienestar emocional de las familias afectadas (2).

Este estudio tiene como objetivo desarrollar un modelo predictivo basado en técnicas estadísticas y de machine learning para identificar a las gestantes en riesgo de mortalidad fetal. Utilizando datos de nacidos vivos y defunciones fetales reportados por el DANE, este modelo buscará ofrecer un enfoque más preciso y temprano para la clasificación del riesgo, lo que permitirá diseñar estrategias de intervención más efectivas y focalizadas. Además, se busca destacar la importancia de los factores sociodemográficos y clínicos en la mortalidad fetal, sugiriendo que intervenciones oportunas, basadas en este tipo de modelos predictivos, pueden reducir significativamente la tasa de muertes fetales en Colombia.

MÉTODOS

Se realizó un estudio de casos y controles para evaluar los factores asociados a la mortalidad fetal en Colombia durante el año 2022. La fuente de datos incluyó registros de nacidos vivos y defunciones fetales reportados oficialmente por el Departamento Administrativo Nacional de Estadística (DANE), y el análisis estadístico se ejecutó en R, versión 4.2.3 (R Foundation for Statistical Computing, Viena, Austria).

Recopilación de datos

Los datos de nacidos vivos y defunciones fetales se obtuvieron de la microdata anonimizada del DANE, la entidad nacional responsable de recopilar, procesar y publicar estadísticas vitales. El registro de nacidos vivos se realiza en instituciones de salud, incluyendo hospitales y clínicas, quienes reportan los nacimientos al DANE tras su inscripción en el registro civil. Las defunciones fetales, que comprenden eventos ocurridos antes o durante el parto, también son registradas y reportadas a los registros civiles a través de estas instituciones. La información publicada por el DANE está disponible en informes anuales que ofrecen estadísticas desglosadas y pueden consultarse en <https://microdatos.dane.gov.co/index.php/catalog/807/get-microdata>.

Para el análisis de este estudio, se seleccionaron datos correspondientes a nacidos vivos y defunciones fetales del año 2022, excluyendo casos de abortos espontáneos mediante la selección de eventos con una edad gestacional superior a 22 semanas. Las bases de datos fueron unificadas mediante variables comunes, y se añadió una variable de clasificación que identificaba cada registro como nacido vivo o defunción fetal.

Procesamiento de datos faltantes

Se identificaron datos faltantes en varias variables, lo que podía introducir sesgos y afectar la precisión del análisis. Para abordar esta limitación, se siguieron varias etapas. Primero, se eliminaron registros que contenían más del 20% de valores faltantes para preservar la integridad del conjunto de datos sin sacrificar un volumen considerable de información. Posteriormente, se aplicó imputación múltiple mediante el paquete mice (Multivariate Imputation by Chained Equations) en R, un método que permite una estimación robusta de valores faltantes al generar múltiples versiones completas del conjunto de datos.

Para la imputación de variables categóricas, se utilizó la regresión polinómica (polyreg), basada en

regresión logística multinomial para predecir la categoría más probable de observaciones faltantes. Para variables continuas o discretas, se empleó el método de Predictive Mean Matching (pmm), que selecciona valores reales observados de casos similares, manteniendo la coherencia con los datos originales. Este proceso se ejecutó en cinco iteraciones, generando cinco conjuntos de datos completos imputados, cada uno reflejando posibles variaciones de los datos faltantes y garantizando así una estimación sólida y representativa.

Balanceo de clases

Dado que la clase de nacidos vivos representaba una mayoría en la variable de resultado, se aplicó el método SMOTE (Synthetic Minority Over-sampling Technique) para equilibrar las clases y evitar sesgos en el análisis. Este procedimiento, ejecutado mediante el paquete *smotefamily* en R, generó observaciones sintéticas para la clase minoritaria a partir de la interpolación de los valores de sus cinco vecinos más cercanos ($K = 5$). La reproducibilidad de los resultados fue garantizada mediante la fijación de una semilla aleatoria (101).

Preparación final del conjunto de datos

La evaluación del balanceo incluyó el uso de métodos estadísticos para asegurar que el procedimiento no introdujera sesgos en las variables predictoras. Se utilizó la prueba de Chi-cuadrado para analizar las distribuciones de las variables categóricas clave y la prueba de Mantel para verificar que las relaciones entre las variables numéricas permanecieran consistentes antes y después del balanceo. Estas evaluaciones garantizan que el balanceo fue realizado de manera adecuada y no comprometió la integridad de las características originales del conjunto de datos.

La aplicación de este conjunto de procedimientos —que incluyó la eliminación de registros con datos faltantes significativos, la imputación múltiple y el

balanceo de clases mediante SMOTE— mejoró la representatividad y calidad del conjunto de datos, reduciendo el sesgo asociado tanto a los datos faltantes como al desequilibrio de clases. Estas medidas aseguran una base de datos robusta y equilibrada, idónea para el desarrollo y evaluación del modelo predictivo utilizado en este estudio.

Análisis estadístico

Se realizó un análisis descriptivo de las variables, separado por clases (nacido vivo y defunción fetal). Para las variables numéricas, se calcularon la media y desviación estándar; las variables categóricas se expresaron en porcentajes. Tomando como variable respuesta la clasificación de nacido vivo o mortinato, se modelaron los datos con Modelo Lineal Generalizado (GLM), Máquina de Soporte Vectorial (SVM), algoritmo K-Nearest Neighbors (KNN) y Random Forest.

Para evaluar el rendimiento de los modelos de clasificación utilizados en este estudio, se aplicaron varias métricas de bondad y ajuste, que incluyen **exactitud (accuracy)**, **sensibilidad**, **especificidad** y **F-1 score (18)**. Estas métricas proporcionan una visión integral de la capacidad predictiva de los modelos y permiten una comparación detallada de su desempeño.

La **exactitud (accuracy)** mide la proporción de predicciones correctas (tanto verdaderos positivos como verdaderos negativos) sobre el total de predicciones realizadas (19). La **sensibilidad** evalúa la capacidad del modelo para identificar correctamente los casos positivos reales (20). La **especificidad** mide la capacidad del modelo para identificar correctamente los casos negativos (21). El **F-1 score** proporciona una medida balanceada entre la precisión y la sensibilidad, tomando en cuenta tanto los falsos positivos como los falsos negativos (22).

Cada modelo (GLM, Random Forest, SVM y KNN) fue evaluado utilizando estas métricas para

determinar su rendimiento en la clasificación de datos. Las métricas permitieron una comparación efectiva y una comprensión detallada de las fortalezas y debilidades de cada modelo en el

contexto de los datos analizados. Las decisiones sobre el modelo más adecuado se basaron en estas métricas.

Tabla 1. Descripción de las clases desbalanceadas

	Nacidos Vivos	Muertes fetales
	562.828 (99,30%)	3.978 (0,70%)
Sexo		
Hombres	287.627 (51,10%)	2.122 (53,34%)
Mujeres	275.165 (48,89%)	1.675 (42,11%)
Indeterminado	35 (0,01%)	181 (4,55%)
Tiempo de gestación		
22 – 27 semanas	2.175 (0,39%)	1.927 (48,44%)
28 – 37 semanas	140.155 (24,90%)	1.608 (40,42%)
38 – 41 semanas	418.104 (74,28%)	429 (2,49%)
> 42 semanas	710 (0,13%)	9 (0,23%)
Ignorado	1.684 (1,88%)	5 (0,13%)
Peso		
< 1.000 gr	2.382 (0,43%)	1.964 (49,37%)
1.000 - 1.499	4.453 (0,79%)	584 (14,68%)
1.500 - 1.999	10.608 (1,88%)	406 (10,21%)
2.000 - 2.499	41.301 (7,34%)	380 (9,55%)
2.500 - 2.999	168.867 (30,00%)	317 (7,97%)
3.000 - 3.499	236.338 (41,99%)	215 (5,40%)
3.500 - 3.999	87.265 (15,50%)	69 (1,73%)
> 4.000	11.614 (2,06%)	43 (1,08%)
Edad de la Madre		
10-14	4.068 (0,72%)	57 (1,43%)
15-19	91.657 (16,29%)	716 (18,00%)
20-24	154.040 (27,37%)	1.092 (27,45%)
25-29	144.015 (25,59%)	889 (22,35%)
30-34	101.037 (17,95%)	608 (15,28%)
35-39	52.712 (9,37%)	404 (10,16%)
40-44	14.322 (2,54%)	194 (4,88%)
45-49	888 (0,16%)	17 (0,43%)
50-54	89 (0,02%)	1 (0,03%)
Número de hijos		
Media (DE)	1.93 (1.23)	1.17 (1.48)
Área de residencia		
Cabecera municipal	423.453 (75,24%)	2.98 (75,11%)
Centro poblado	40.544 (7,20%)	263 (6,61%)
Rural disperso	98.831 (17,56%)	727 (18,28%)
Régimen afiliación		
Contributivo	205.395 (36,49%)	1.219 (30,64%)
Subsidiado	309.775 (44,04%)	2.291 (57,59%)
Exceptuado	7.784 (1,38%)	42 (1,06%)
Especial	140 (0,02%)	1 (0,03%)

No asegurado	39.734 (7,06%)	425 (10,68%)
Unión libre con 2 años o mas	317.823 (56,47%)	1.997 (50,20%)
Unión libre con menos 2 años	103.250 (18,34%)	662 (16,64%)
Separada o Divorciada	1.766 (0,31%)	15 (0,38%)
Viuda	642 (0,11%)	6 (0,15%)
Soltera	67.948 (12,07%)	973 (24,46%)
Casada	71.399 (12,69%)	325 (8,17%)
Preescolar	1.290 (0,23%)	7 (0,18%)
Básica primaria	64.623 (11,48%)	602 (15,13%)
Básica secundaria	126.227 (22,43%)	984 (24,74%)
Media académica	195.049 (34,66%)	1.473 (37,03%)
Media técnica	14.784 (2,63%)	124 (3,12%)
Normalista	781 (0,14%)	8 (0,20%)
Técnica profesional	49.912 (8,87%)	238 (5,98%)
Tecnológica	28.219 (5,01%)	141 (3,54%)
Profesional	65.665 (11,67%)	273 (6,86%)
Especialización	7.089 (1,26%)	20 (0,50%)
Maestría	2.435 (0,43%)	4 (0,10%)
Doctorado	144 (0,03%)	1 (0,03%)
Ninguno	6.610 (1,17%)	103(2,59%)

El área de residencia predominante fue la cabecera municipal en ambos grupos, con un 75,24% en nacidos vivos y un 75,11% en muertes fetales. En cuanto al régimen de afiliación, la modalidad subsidiada fue la más frecuente, alcanzando un 44,04% en el grupo de nacidos vivos y un 57,59% en el grupo de muertes fetales.

Para abordar el desbalance en la variable de resultado, se empleó el método SMOTE con el objetivo de ajustar la proporción inicial de nacidos vivos (99,30%) y muertes fetales (0,70%) a una distribución balanceada del 50% para cada clase. Este procedimiento, basado en la generación de observaciones sintéticas mediante interpolación de los valores de los cinco vecinos más cercanos, garantizó una mejor representación de la clase minoritaria en el análisis. La eficacia del balanceo fue evaluada a través de pruebas estadísticas rigurosas. La prueba de Chi-cuadrado identificó cambios significativos en la distribución de varias variables categóricas, lo que evidencia un impacto del proceso de balanceo en la estructura de estas variables. Sin embargo, la prueba de Mantel

demostró que las correlaciones entre las variables originales y balanceadas se mantuvieron altamente consistentes, con un estadístico r de 0,8876 y un p -valor significativo ($< 0,001$). Estos resultados confirman que el balanceo preservó las relaciones fundamentales entre las variables, garantizando un conjunto de datos representativo y equilibrado, ideal para el análisis predictivo y la identificación de factores asociados con los desenlaces perinatales.

Se seleccionaron las variables del modelo más parsimonioso. En el análisis de regresión logística presentado se evaluaron los factores asociados con la probabilidad de nacer vivo en comparación con nacer muerto. Todos los factores incluidos en el modelo mostraron asociaciones estadísticamente significativas, con sus respectivos Odds Ratios (OR) y valores p igual a 0,00 (Tabla 2). Se observó que variables como, el número de hijos, el sexo, el área de residencia, el régimen de seguridad social, las semanas de gestación, el peso al nacer, la edad y el estado civil de la madre, así como su nivel educativo, influyen de manera significativa en la probabilidad de un nacimiento vivo. Estos

resultados indican que cada una de estas variables es un predictor importante del desenlace perinatal (Tabla 2).

El número de hijos se asoció positivamente con la probabilidad de nacer vivo, donde cada hijo

adicional incrementa la probabilidad de nacer vivo en aproximadamente un 143% (OR: 2,43; IC95%: 2,41-2,45). Este intervalo estrecho refuerza la confianza en esta estimación, indicando un efecto práctico relevante.

Tabla 2: Odds de las variables del modelo

Variable		Odds Ratio (OR)	Coefficiente	IC95% Inferior	IC95% Superior	Valor p
	Intercepto	0	-17,66	0,00	0,00	0,00
	Número de hijos	2,43	0,889	2,41	2,45	0,00
Sexo	Masculino	19844,86	9,896	9225,12	45290,53	0,00
	Femenino	26095,23	10,17	12130,66	59555,53	0,00
Área de residencia	Cabecera Municipal	1,31	0,267	1,28	1,33	0,00
	Centro Poblado	1,34	0,291	1,30	1,38	0,00
Régimen de afiliación en salud	Contributivo	2,53	0,927	2,45	2,61	0,00
	Subsidiado	1,31	0,268	1,27	1,34	0,00
	Excepción	2,50	0,915	2,30	2,71	0,00
	Especial	5,13	1,635	2,88	9,12	0,00
Semanas de gestación	22-27	0,60	-0,517	0,51	0,70	0,00
	28-37	1,88	0,632	1,63	2,17	0,00
	38-41	5,20	1,649	4,51	5,99	0,00
Peso al nacer	<1.000	0,03	-3,682	0,02	0,03	0,00
	1.000 - 1.499	0,09	-2,466	0,08	0,09	0,00
	1.500 - 1.999	0,23	-1,474	0,22	0,24	0,00
	2.000 - 2.499	0,91	-0,093	0,87	0,95	0,00
	2.500 - 2.999	3,31	1,197	3,17	3,46	0,00
	3.000 - 3.499	5,45	1,695	5,21	5,69	0,00
	3.500 - 3.999	6,31	1,842	6,00	6,63	0,00
Edad de la madre	10-14	25,34	3,233	23,18	27,72	0,00
	15-19	11,35	2,429	10,97	11,74	0,00
	20-24	5,42	1,69	5,26	5,58	0,00
	25-29	3,05	1,116	2,97	3,14	0,00
	30-34	1,88	0,628	1,82	1,93	0,00
	40-44	0,51	-0,676	0,49	0,53	0,00
	45-49	0,24	-1,409	0,21	0,28	0,00
Estado civil de la madre	No Casada y 2 Años o más	1,06	0,055	1,03	1,08	0,00
	No Casada y Menos 2 Años	1,49	0,398	1,45	1,53	0,00
	Separada o Divorciada	1,79	0,58	1,55	2,07	0,00

	Variable	Odds Ratio (OR)	Coefficiente	IC95% Inferior	IC95% Superior	Valor p
	Viuda	0,77	-0,257	0,63	0,96	0,02
	Casada	1,43	0,354	1,38	1,47	0,00
Nivel Educativo de la madre	Preescolar	37,75	3,631	31,09	46,05	0,00
	Básica Primaria	10,34	2,336	9,84	10,86	0,00
	Básica Secundaria	20,94	3,042	19,93	22,00	0,00
	Media Académica	32,89	3,493	31,29	34,57	0,00
	Media Técnica	57,90	4,059	53,91	62,21	0,00
	Normalista	21,61	3,073	17,86	26,25	0,00
	Técnica Profesional	96,07	4,565	90,52	101,97	0,00
	Tecnológica	70,05	4,249	65,71	74,68	0,00
	Profesional	143,07	4,963	134,72	151,96	0,00
	Especialización	328,19	5,794	291,39	370,04	0,00
	Maestría	459,81	6,131	369,54	576,64	0,00
	Doctorado	1334,89	7,197	462,51	4041,47	0,00

En cuanto al sexo, los nacidos masculinos y femeninos tienen probabilidades mucho mayores de nacer vivos en comparación con la categoría indeterminada, que corresponde a la intersexualidad o indeterminado, una condición en la cual una persona presenta características sexuales físicas que no encajan completamente en las definiciones típicas de masculino o femenino.

Los OR asociados al sexo masculino (OR ≈ 19844,86; IC95%: 9225,12-45290,53) y femenino (OR ≈ 26095,23; IC95%: 12130,66-59555,53) reflejan un fuerte impacto, aunque la amplitud de sus intervalos sugiere cierta variabilidad que podría atribuirse a la baja frecuencia de casos de sexo indeterminado en la muestra.

Para la residencia en cabecera municipal y centro poblado, los resultados muestran ligeros aumentos en la probabilidad de nacer vivo (Cabecera municipal: OR ≈ 1,31; IC95%: 1,28-1,33; Centro poblado: OR ≈ 1,34; IC95%: 1,30-1,38), lo cual podría reflejar las posibles ventajas de accesibilidad o calidad de atención en estos entornos. Los IC

estrechos respaldan la consistencia de estos efectos.

El régimen de seguridad social mostró diferencias notables. Los nacidos en regímenes contributivo (OR ≈ 2,53; IC95%: 2,45-2,61) y de excepción (OR ≈ 2,50; IC95%: 2,30-2,71) tuvieron mayores probabilidades de nacer vivos en comparación con los no asegurados. Este efecto es aún más marcado para el régimen especial (OR ≈ 5,13; IC95%: 2,88-9,12), aunque un IC más amplio indica una mayor incertidumbre para este subgrupo.

Respecto a las semanas de gestación, los nacimientos entre 22-27 semanas y 28-37 semanas presentaron diferencias significativas. La probabilidad de nacer vivo es menor entre los nacidos en el rango de 22-27 semanas (OR ≈ 0,60; IC95%: 0,51-0,70), mientras que los nacidos entre 28-37 semanas (OR ≈ 1,88; IC95%: 1,63-2,17) y 38-41 semanas (OR ≈ 5,20; IC95%: 4,51-5,99) tienen probabilidades crecientes de supervivencia, lo cual es consistente con el riesgo asociado de nacer prematuro.

El peso al nacer se asoció fuertemente con la probabilidad de nacer vivo. Los recién nacidos con peso menor a 1.000 gramos tuvieron una probabilidad menor de nacer vivos (OR $\approx$ 0,03; IC95%: 0,02-0,03), mientras que aquellos con pesos superiores a 2.500 gramos presentaron incrementos significativos (por ejemplo, 3.500-3.999 gramos: OR $\approx$ 6,31; IC95%: 6,00-6,63). La precisión de los IC en estos rangos destaca la importancia crítica del peso al nacer.

La edad materna mostró patrones claros. Las madres más jóvenes, especialmente de 10-14 años (OR $\approx$ 25,34; IC95%: 23,18-27,72), tuvieron una probabilidad significativamente mayor de nacer vivos en comparación con madres de mayor edad, como aquellas de 45-49 años (OR $\approx$ 0,24; IC95%: 0,21-0,28), donde el efecto negativo es consistente con el riesgo elevado de complicaciones obstétricas en edades extremas.

El estado civil de la madre mostró asociaciones relevantes con la probabilidad de un nacimiento vivo. Las madres casadas tuvieron un OR de 1,43 (IC95%: 1,38-1,47), lo que sugiere que este grupo tiene una mayor probabilidad de nacer vivo en comparación con las madres solteras, establecidas como la categoría de referencia. Por su parte, las madres separadas o divorciadas también presentaron una probabilidad superior de un nacimiento vivo (OR: 1,79; IC95%: 1,55-2,07). Sin embargo, las madres viudas mostraron una menor probabilidad relativa, con un OR de 0,77 (IC95%: 0,63-0,96), lo que podría indicar una situación de vulnerabilidad social y económica que afecta el desenlace perinatal. Este hallazgo resalta la importancia del apoyo social y económico para las madres en situaciones de mayor riesgo.

Finalmente, el nivel educativo de la madre se asoció con mayores probabilidades de nacimiento vivo. Las madres con niveles educativos avanzados, como maestría (OR $\approx$ 459,81; IC95%: 369,54-576,64) y doctorado (OR $\approx$ 1334,89; IC95%: 462,51-4.041,47), mostraron los mayores efectos positivos, aunque los IC amplios reflejan mayor incertidumbre en estos casos debido al tamaño reducido de estos subgrupos. Por el contrario, niveles educativos básicos también estuvieron significativamente asociados con un aumento en la probabilidad de nacer vivo (e.g., secundaria: OR $\approx$ 20.94; IC95%: 19.93-22.00).

En conjunto, estos resultados sugieren que factores como el sexo, el régimen de seguridad social, el peso al nacer, la edad materna y el nivel educativo de la madre están significativamente asociados con la probabilidad de nacer vivo.

En la Tabla 3, se muestran los resultados de las medidas de bondad y ajustes para los modelos tratados, se observa que el modelo KNN presentó la mejor exactitud (0,988), lo que indica que este modelo fue el más efectivo en clasificar correctamente la mayor cantidad de casos. Esto fue respaldado por su alta sensibilidad (0,989), reflejando su capacidad para detectar correctamente los casos positivos, y por su especificidad (0,986), que mostró una alta precisión en la identificación de los casos negativos. El valor del F1 Score (0,988) en KNN, una medida que balancea precisión y sensibilidad, también fue el más alto entre los modelos evaluados, reforzando su capacidad para lograr un equilibrio adecuado entre la detección y la precisión en sus predicciones.

Tabla 3: Comparación de las medidas de bondad y ajuste de los modelos

Modelo	Exactitud (Accuracy)	Sensibilidad	Especificidad	F1 Score	Precisión
Regresión Logística	0,907	0,879	0,936	0,905	0,932
KNN	0,988	0,989	0,986	0,988	0,986
Árbol de Decisión	0,929	0,888	0,970	0,926	0,967
Máquina de Soporte Vectorial	0,902	0,869	0,935	0,898	0,930

En contraste, aunque el Árbol de Decisión mostró un buen desempeño en términos de especificidad (0,970) y precisión (0,967), su sensibilidad fue menor (0,888) en comparación con KNN, lo que indica que el Árbol de Decisión podría estar pasando por alto algunos casos positivos. Aun así, su exactitud general (0,929) y su F1 Score (0,926) fueron altos, lo que señala que, si bien es un modelo robusto, podría ser menos eficiente en la detección de ciertos casos en comparación con KNN.

Los modelos de Regresión Logística y SVM presentaron rendimientos similares, aunque menores en todas las métricas clave. La Regresión Logística mostró una exactitud de 0,907, con una sensibilidad de 0,879 y una especificidad de 0,936. Por su parte, el modelo SVM obtuvo una exactitud de 0,902, una sensibilidad de 0,869 y una especificidad de 0,935. Ambos modelos, aunque efectivos en términos generales, se quedaron por detrás de KNN y el Árbol de Decisión en la mayoría de las métricas, lo que sugiere que podrían ser menos efectivos en este contexto específico.

Estos resultados permiten concluir que el modelo KNN es el que mejor rendimiento ofrece para predecir el desenlace de interés en este estudio, ya que no solo demostró una alta exactitud, sino que también destacó en sensibilidad y especificidad, lo cual sugiere que es confiable tanto en la detección de casos positivos como en la clasificación de casos negativos. Esto implica que KNN podría ser altamente efectivo en aplicaciones prácticas, donde el balance entre sensibilidad y especificidad es crucial para asegurar la precisión del modelo en un entorno clínico o de salud pública.

DISCUSIÓN

El presente estudio analizó los factores asociados con la probabilidad de nacer vivo en Colombia durante el año 2022, utilizando una base de datos nacional proporcionada por el DANE. Los hallazgos revelan asociaciones significativas entre diversas

variables maternas, neonatales y socioeconómicas con el desenlace de interés. Estos resultados aportan evidencia actualizada y relevante al campo de la epidemiología perinatal en Colombia, ofreciendo perspectivas para mejorar las estrategias de salud pública orientadas a reducir las muertes fetales.

La alta prevalencia de nacidos vivos (99,30%) en comparación con las muertes fetales (0,70%) es consistente con estadísticas nacionales previas y refleja los avances en la atención prenatal y perinatal en el país (25–27). Sin embargo, las muertes fetales continúan siendo un indicador crítico de la salud materna y perinatal, y su análisis es esencial para identificar áreas de mejora.

Uno de los hallazgos más destacados es la asociación positiva entre el número de hijos y la probabilidad de nacer vivo. Este resultado contrasta con algunos estudios que señalan que existe una alta paridad puede asociarse con mayores riesgos obstétricos (28). Sin embargo, en contextos donde las madres con más hijos podrían tener mayor experiencia y acceso a cuidados prenatales, este factor podría actuar como protector. Es posible que las madres con múltiples partos en Colombia estén más familiarizadas con los servicios de salud y las señales de alarma durante el embarazo, lo que podría actuar como un factor protector en el momento del parto, contribuyendo al éxito del nacimiento del recién nacido.

El sexo del recién nacido mostró una fuerte asociación con la probabilidad de nacer vivo, siendo el sexo masculino y femenino factores que aumentan significativamente dicha probabilidad en comparación con la categoría de referencia. Dado que la categoría de referencia es 'indeterminada' o 'intersexual', que se refiere a aquellos casos en los que las características sexuales del recién nacido no encajan claramente en las definiciones típicas de masculino o femenino,

y considerando que los casos de sexo indeterminado representaron un pequeño porcentaje, este resultado podría reflejar limitaciones en la calidad o completitud de los datos registrados. Es importante abordar este aspecto en futuros estudios para asegurar la precisión de las variables analizadas.

La residencia en cabeceras municipales y centros poblados se asoció con un aumento en la probabilidad de nacer vivo. Esto podría deberse a un mejor acceso a servicios de salud materno-perinatal, mayor disponibilidad de personal calificado y mejor infraestructura sanitaria en áreas urbanas en comparación con zonas rurales (29–31). Este hallazgo resalta la necesidad de fortalecer la atención prenatal y perinatal en zonas rurales y dispersas, donde la falta de recursos puede afectar negativamente los desenlaces neonatales. Por lo tanto, proporciona información clave para el diseño de políticas públicas más focalizadas, enfocadas en la construcción de centros de atención o estrategias que garanticen el acceso en las regiones rurales del país y en poblaciones de escasos recursos.

El régimen de afiliación en salud también mostró asociaciones significativas. Los nacidos en regímenes contributivo, de excepción y especial presentaron mayores probabilidades de nacer vivos. Estos regímenes suelen ofrecer mayor cobertura y calidad en la atención médica (30–32), lo que podría explicar los mejores desenlaces observados. Por el contrario, el régimen subsidiado y la no afiliación, asociados con poblaciones de menores recursos, mostraron una menor probabilidad de nacer vivo, evidenciando inequidades en el sistema de salud que requieren ser abordadas. Esto implica que el problema de las defunciones recae en parte en el Estado, al no garantizar una equidad desde el momento del control prenatal pasando por nacimiento de los niños. Este hecho refleja las fuertes inequidades presentes en Colombia.

La edad gestacional y el peso al nacer fueron factores críticos en la probabilidad de nacer vivo, alineándose con la literatura existente (33). Los recién nacidos con mayor edad gestacional y peso al nacer superior a 2.500 gramos tuvieron mayores probabilidades de nacer vivos. Los nacidos prematuros y el bajo peso al nacer son conocidos factores de riesgo para la mortalidad neonatal, asociados con inmadurez orgánica y mayor susceptibilidad a complicaciones (34,35). Estos resultados subrayan la importancia de intervenciones orientadas a prevenir el parto prematuro y mejorar la nutrición materna.

La edad materna mostró una asociación significativa con la probabilidad de nacer vivo. Nuestros resultados indican que las madres muy jóvenes (10-14 años) presentan una mayor probabilidad de nacimientos vivos, mientras que las madres de mayor edad (45-49 años) tienen un mayor riesgo de muertes fetales después de la semana 22. Estos resultados contrastan con evidencia que manifiesta que el embarazo en adolescentes se asocia con riesgos biológicos y sociales que pueden afectar el desarrollo fetal (36). Por otro lado, concuerda con las gestantes en edad avanzadas aumentan las complicaciones obstétricas y enfermedades crónicas (37). Este hallazgo indica la necesidad de programas de educación sexual y reproductiva, así como de cuidados especializados para embarazos en edades extremas.

El nivel educativo de la madre emergió como un factor determinante. Un mayor nivel educativo se asoció con una mayor probabilidad de nacer vivo, lo que concuerda con estudios previos (38–42) que relacionan la educación materna con mejores prácticas de salud, mayor uso de servicios prenatales y mejores condiciones socioeconómicas (42). La educación empodera a las mujeres para tomar decisiones informadas sobre su salud y la de sus hijos, lo cual es fundamental para mejorar los desenlaces perinatales.

El modelo de K-Nearest Neighbors (KNN) demostró ser el más efectivo en la predicción del desenlace de nacer vivo, superando a la regresión logística, los árboles de decisión y las máquinas de soporte vectorial. La alta exactitud, sensibilidad y especificidad del KNN indican su capacidad para manejar grandes conjuntos de datos y captar patrones complejos (43). Sin embargo, su aplicación en entornos clínicos puede ser limitada debido a su menor interpretabilidad en comparación con modelos como la regresión logística. Es esencial equilibrar la precisión predictiva con la capacidad de extraer conclusiones accionables para la práctica médica. En investigaciones futuras, sería valioso explorar enfoques combinados de técnicas, que permitan aprovechar la potencia predictiva del KNN y la interpretabilidad de otros modelos, mejorando así la comprensión de los resultados.

Este estudio presenta algunas limitaciones que deben ser consideradas. La naturaleza retrospectiva y observacional del análisis impide establecer relaciones causales. Además, la dependencia de datos secundarios plantea problemas potenciales de calidad, como registros incompletos o errores en la codificación de variables, especialmente en aspectos como el sexo indeterminado o el peso al nacer. La categorización de variables y la elección de referencias pueden influir en la interpretación de los odds ratios, por lo que se recomienda cautela al extrapolar estos resultados. En futuras investigaciones, es fundamental contar con un sistema de registro más detallado por parte de las instituciones de salud, que permita depurar y manejar la información de manera más limpia. Esto facilitaría la captura de la variabilidad presente en los datos, reduciría el riesgo de perder información y, en consecuencia, mejoraría la precisión de los modelos analíticos.

Asimismo, aunque el modelo KNN mostró un excelente rendimiento predictivo, su aplicabilidad práctica requiere una validación externa y una

evaluación de su utilidad clínica real. La falta de datos sobre factores socioeconómicos más detallados y variables clínicas específicas limita la comprensión completa de los determinantes de la mortalidad fetal.

Los hallazgos de este estudio tienen importantes implicaciones para la salud pública en Colombia. La identificación de los factores asociados con la probabilidad de nacer vivo proporciona una base sólida para orientar intervenciones dirigidas a las poblaciones más vulnerables. En particular, mejorar el acceso y la calidad de la atención prenatal en zonas rurales es prioritario para reducir riesgos perinatales en estas áreas. Asimismo, los resultados destacan la necesidad de implementar políticas de apoyo específicas para grupos con necesidades particulares, como las madres no afiliadas al sistema de salud y las de edad avanzada, quienes requieren atención especializada para optimizar los desenlaces perinatales. Estas políticas deben centrarse en ofrecer una atención integral y fortalecer la equidad en el sistema de salud, garantizando que todos los grupos tengan acceso a cuidados prenatales de calidad.

Las madres que se identificaron como solteras o viudas presentan una mayor probabilidad de que sus hijos no nazcan vivos, lo cual resalta la necesidad de implementar programas de acompañamiento psicológico específicos para esta población. Ofrecer un apoyo emocional adecuado podría mejorar significativamente las probabilidades de un nacimiento vivo, al proporcionar a estas madres el soporte necesario durante el embarazo. En un contexto donde es cada vez más común que las madres enfrenten la maternidad sin pareja, la implementación de estos programas es esencial para optimizar los resultados de salud materno-infantil y reducir los riesgos asociados a la falta de acompañamiento.

Futuros estudios deberían profundizar en los determinantes sociales y clínicos de la mortalidad

fetal, incorporando variables clave como las condiciones maternas preexistentes, la calidad de la atención durante el parto, y los factores ambientales. En este sentido, sería valioso incluir variables adicionales relacionadas con la atención brindada por el sistema de salud, así como otros factores clínicos y ambientales que puedan influir en los desenlaces perinatales. Estas variables permitirían una visión más integral de las causas y condicionantes de la mortalidad fetal, aportando información valiosa para el diseño de intervenciones efectivas.

Además, resulta pertinente realizar validaciones del modelo y los hallazgos en diferentes regiones del país o períodos temporales, con el objetivo de evaluar la consistencia de los resultados y su aplicabilidad en distintos contextos.

En Colombia, es urgente diseñar políticas públicas que reduzcan las brechas económicas entre los diferentes regímenes de salud, garantizando así un acceso equitativo a la atención prenatal. Es esencial fortalecer los programas de salud materno-infantil en las zonas rurales y en las poblaciones marginadas del país, donde los desafíos en el acceso a servicios de salud son más evidentes. Además, se debe mejorar la calidad de los registros de salud, asegurando la fiabilidad de los datos utilizados en investigaciones epidemiológicas. Para ello, es necesario implementar políticas de recolección de datos más detalladas, que incluyan información específica sobre las condiciones maternas, lo que permitirá obtener una visión más precisa y completa de los factores que inciden en los desenlaces perinatales.

En resumen, este estudio aporta evidencia valiosa sobre los factores que influyen en la supervivencia neonatal en Colombia, resaltando la importancia de políticas integrales que aborden las inequidades en salud y fortalezcan los servicios materno-infantiles.

CONTRIBUCIONES DE LOS AUTORES: PLAZ, JJCC, DSML: concepción y diseño del estudio. PLAZ, JJCC: recolección, análisis e interpretación de datos. PLAZ, JJCC: redacción del borrador del artículo. PLAZ, JJCC, DSML: revisión crítica y aprobación de la versión final, responsable de la veracidad e integridad del artículo.

CONFLICTOS DE INTERÉS: los autores declaran no presentar ningún tipo de interés para la realización y publicación del presente reporte de caso.

FINANCIACIÓN: la presente investigación no ha recibido ayudas específicas provenientes de agencias del sector público, sector comercial o entidades sin ánimo de lucro.

REFERENCIAS

1. Goldenberg RL, McClure EM, Belizán JM. Commentary: Reducing the world's stillbirths. Vol. 9, BMC Pregnancy and Childbirth. 2009.
2. García MV, Cuberos MA, Vera MÁ, Vivas García M, Cuberos MA, Vera MÁ. Mortalidad fetal en madres adolescentes y adultas, un problema de salud pública TT - Fetal mortality in adolescent and adult mothers: a public health problem. Salud(i)Ciencia. 2019;23 (5).
3. Lawn JE, Blencowe H, Pattinson R, Cousens S, Kumar R, Ibiebele I, et al. Stillbirths: Where? When? Why? How to make the data count? Vol. 377, The Lancet. 2011.
4. Departamento Administrativo Nacional de Estadísticas. Boletín técnico Estadísticas Vitales (EEVV) [Internet]. Bogotá D.C.; 2024 Mar [cited 2024 Sep 20]. Available from: <https://www.dane.gov.co/files/operaciones/EEVV/bol-EEVV-Defunciones-IVtrim2023.pdf>
5. Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE. Permutation inference for the general linear model. Neuroimage. 2014 May 15;92:381–97.
6. Casella G, Fienberg S, Olkin I. Springer Texts in Statistics [Internet]. Available from: www.springer.com/series/417
7. Chartier S, Faulkner A. General Linear Models: An Integrated Approach to Statistics. Tutor Quant Methods Psychol. 2008 Sep 1;4(2):65–78.

8. Helwig NE. Robust nonparametric tests of general linear model coefficients: A comparison of permutation methods and test statistics. *Neuroimage*. 2019 Nov 1;201.
9. Aizerman MA, Braverman EM, Rozonoer LI. Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning. *Telemekhanika*. 1964;25(6):971–936.
10. Huang S, Nianguang CAI, Penzuti Pacheco P, Narandes S, Wang Y, Wayne XU. Applications of support vector machine (SVM) learning in cancer genomics. Vol. 15, *Cancer Genomics and Proteomics*. International Institute of Anticancer Research; 2018. p. 41–51.
11. Noble WS. What is a support vector machine? [Internet]. Vol. 24, *NATURE BIOTECHNOLOGY*. 2006. Available from: <http://www.nature.com/naturebiotechnology>
12. Hu LY, Huang MW, Ke SW, Tsai CF. The distance function effect on k-nearest neighbor classification for medical datasets. *Springerplus*. 2016;5(1).
13. Ehsani R, Drabløs F. Robust Distance Measures for kNN Classification of Cancer Data. *Cancer Inform*. 2020;19.
14. Hu J, Szymczak S. A review on longitudinal data analysis with random forest. Vol. 24, *Briefings in Bioinformatics*. 2023.
15. Breiman L. Random Forests. Vol. 45. 2001.
16. Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random forests for classification in ecology. *Ecology*. 2007;88 (11).
17. Graham SE, Clarke SL, Wu KHH, Kanoni S, Zajac GJM, Ramdas S, et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature*. 2021;600 (7890).
18. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology*. 2011;2 (1):37–63.
19. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006 Jun;27 (8):861–74.
20. He H, Garcia EA. Learning from Imbalanced Data. *IEEE Trans Knowl Data Eng*. 2009;21 (9):1263–84.
21. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag*. 2009 Jul;45 (4):427–37.
22. Van Rijsbergen CJ. *Information retrieval*. 1979.
23. Manning C, Raghavan P, Schütze H. *An introduction to information retrieval*. Cambridge; 2009.
24. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. *Intell Data Anal*. 2002 Oct;6(5):429–49.
25. Departamento Administrativo Nacional de Estadística DANE. Boletín Técnico Estadísticas Vitales (EEVV) Nacidos Vivos. Bogotá, DC; 2023 Mar.
26. Departamento Administrativo Nacional de Estadística. Boletín técnico Estadísticas Vitales (EEVV) Defunciones. Bogotá, DC; 2023 Mar.
27. McClure EM, Saleem S, Goudar SS, Garces A, Whitworth R, Esamai F, et al. Stillbirth 2010–2018: a prospective, population-based, multi-country study from the Global Network. *Reprod Health*. 2020 Nov 30;17 (S2):146.
28. Mgaya AH, Massawe SN, Kidanto HL, Mgaya HN. Grand multiparity: is it still a risk in pregnancy? *BMC Pregnancy Childbirth*. 2013 Dec 23;13(1):241.
29. Galvis-Aponte L, Rico J. Desigualdades regionales en la salud en Colombia. 2023 Jun.
30. Ministerio S y PS. Informe Nacional de Calidad de la Atención en Salud 2015 [Internet]. Bogotá, DC; 2015 Dec [cited 2024 Nov 10]. Available from: <https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/informe-nal-calidad-atencion-salud-2015.pdf>
31. Cáceres-Manrique F de M, Molina-Marín G. Inequidades sociales en atención materna. *Rev Colomb Obstet Ginecol*. 2010 Sep 30;61 (3):231–8.

32. Merlano Porras C, Gorbanev I. Health system in Colombia: A systematic review of literature. *Revista Gerencia y Políticas de Salud*. 2013 Jun 1;12:74–86.
33. Kramer MS. Determinants of low birth weight: methodological assessment and meta-analysis. *Bull World Health Organ*. 1987;65 (5):663–737.
34. Blencowe H, Cousens S, Chou D, Oestergaard M, Say L, Moller AB, et al. Born Too Soon: The global epidemiology of 15 million preterm births. *Reprod Health*. 2013 Nov 15;10 (S1):S2.
35. Liu L, Oza S, Hogan D, Chu Y, Perin J, Zhu J, et al. Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. *The Lancet*. 2016 Dec;388 (10063):3027–35.
36. Ganchimeg T, Ota E, Morisaki N, Laopaiboon M, Lumbiganon P, Zhang J, et al. Pregnancy and childbirth outcomes among adolescent mothers: a World Health Organization multicountry study. *BJOG*. 2014 Mar 18;121 (s1):40–8.
37. Jolly M, Sebire N, Harris J, Robinson S, Regan L. The risks associated with pregnancy in women aged 35 years or older. *Human Reproduction*. 2000 Nov 1;15 (11):2433–7.
38. D'Ascoli PT, Alexander GR, Petersen DJ, Kogan MD. Parental factors influencing patterns of prenatal care utilization. *J Perinatol*. 1997;17(4):283–7.
39. Victora CG, Aquino EM, do Carmo Leal M, Monteiro CA, Barros FC, Szwarcwald CL. Maternal and child health in Brazil: progress and challenges. *The Lancet*. 2011 May;377 (9780):1863–76.
40. Gakidou E, Cowling K, Lozano R, Murray CJ. Increased educational attainment and its effect on child mortality in 175 countries between 1970 and 2009: a systematic analysis. *The Lancet*. 2010 Sep;376 (9745):959–74.
41. Fuchs R, Pamuk E, Lutz W. Education or wealth: Which matters more for reducing child mortality in developing countries? *Vienna Yearb Popul Res*. 2010 Nov;8:175–99.
42. Güneş PM. The role of maternal education in child health: Evidence from a compulsory schooling law. *Econ Educ Rev*. 2015 Aug;47:1–16.
43. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory*. 1967 Jan;13 (1):21–7.