# Determinación de factores de riesgo de bajo peso al nacer en Colombia aplicando análisis exploratorio de datos y modelos de *machine learning* para los años 2017 al 2021

**César Eduardo Aguas Aldana**

**Monografía presentada como requisito para optar al título de:**
**Matemático**

**Director:**
**Oscar Leonardo Acevedo Pabón**

**Institución Universitaria Politécnico Grancolombiano**
**Facultad de Ingeniería, Diseño e Innovación**
**Escuela de Ciencias Básicas**
**Bogotá D.C., Colombia**
**2025**

## Tabla de contenido

## Resumen

El bajo peso al nacer (BPN) es una condición que presenta riesgos significativos para la salud de los recién nacidos. Este proyecto se enfoca en determinar los factores de riesgo asociados con el BPN en Colombia aplicando análisis exploratorio de datos y modelos de *machine learning*. Como fuente de datos, se utilizó una muestra para el periodo 2017-2021 proporcionada por el Departamento Administrativo Nacional de Estadística (DANE,2022).

Dentro de las variables exógenas que podrían constituir un factor de riesgo que se incluyeron en los modelos hay características sociodemográficas de la madre como la edad, el número de partos, el nivel educativo y la región geográfica. Adicionalmente, se discuten otras variables potencialmente importantes como la ruralidad y las condiciones socioeconómicas, aunque estas no se incorporaron en los modelos debido a la incompletitud de los datos.

Un análisis bivariado mostró que las madres adolescentes y de mayor edad, con bajo nivel educativo y residentes en áreas rurales tienen un mayor riesgo de tener bebés con BPN. Varios modelos de *machine learning* fueron evaluados, y el modelo con mejor desempeño fue una regresión logística mejorada con la técnica SMOTE (*Synthetic Minority Over-sampling Technique*). Para evaluar el rendimiento de los modelos, se aplicaron técnicas de validación cruzada y se utilizaron métricas como precisión, *recall* y F1-*score*.

El estudio concluye que las intervenciones para mejorar el acceso a servicios de salud en áreas rurales y las políticas que abordan disparidades educativas y socioeconómicas pueden contribuir a reducir la incidencia de BPN en Colombia. La confidencialidad de los datos de la entidad y sus clientes se ha mantenido en todo momento.

**Abstract**

Low birth weight (LBW) is a condition that poses significant health risks to newborns. This project focuses on determining the risk factors associated with LBW in Colombia by applying exploratory data analysis and machine learning models. As a data source, we used a sample for the period 2017-2021 provided by the Colombian National Administrative Department of Statistics (DANE).

Among the exogenous variables included in the model that could constitute a risk factor, there are sociodemographic characteristics of the mother, such as age, number of births, educational level, and geographic region. Additionally, other potentially important variables such as rurality and socioeconomic conditions are discussed, although these were not incorporated into the models due to the incompleteness of the data.

A bivariate analysis showed that adolescent and older mothers with low educational level and living in rural areas have a higher risk of having LBW babies. Several machine learning models were evaluated, and the best-performing model was a logistic regression improved with SMOTE technique (Synthetic Minority Over-sampling). To evaluate the performance of the models, cross-validation techniques were applied and metrics such as precision, recall and F1-score were used.

Our study concludes that interventions to improve access to health services in rural areas and policies that address educational and socioeconomic disparities can contribute to lower the incidence of LBW in Colombia. Data confidentiality of the entity and its clients have always been maintained.

## 1. Objetivos

Este trabajo de investigación tiene como objetivo principal emplear modelos de *machine learning* y análisis exploratorio de datos para identificar factores de riesgo y determinar la probabilidad de bajo peso al nacer en Colombia. Para lograr este objetivo, se plantean los siguientes objetivos específicos: evaluar la consistencia de los datos del DANE, realizar análisis exploratorio de los datos para identificar factores de riesgo, seleccionar variables predictoras significativas mediante el método del valor-*p* y aplicar modelos de *machine learning* para la predicción del bajo peso al nacer.

## 2. Justificación

La justificación de este estudio se basa en razones técnicas, sociales y económicas. Desde el punto de vista técnico, la aplicación de *machine learning* y análisis de datos permitirá una comprensión más profunda y precisa de los factores implicados en el bajo peso al nacer. Socialmente, el estudio aborda un problema de salud pública significativo, con el potencial de mejorar la calidad de vida de los neonatos y sus familias y de sensibilizar a la sociedad sobre la importancia de la atención prenatal. Económicamente, la prevención del bajo peso al nacer puede reducir costos médicos y mejorar la productividad y el bienestar económico de la sociedad. En conjunto, este trabajo busca contribuir a la mejora de la salud y el bienestar en Colombia mediante una mejor comprensión y abordaje del problema del bajo peso al nacer.

## 3. Introducción

Una de las principales metas de la Organización Mundial de la Salud es reducir un 30% los casos de bajo peso al nacer entre 2012 y 2025, lo que implicaría una disminución anual del 3% en las cifras de este problema de salud. Esto significaría pasar de aproximadamente 20 a 14 millones de neonatos afectados por esta condición de morbilidad cada año. La relevancia de este objetivo radica no solo en los beneficios inmediatos para la salud de los recién nacidos, sino también en la prevención de posibles complicaciones a largo plazo, como el aumento del riesgo de padecer enfermedades crónicas como la diabetes o enfermedades cardiovasculares (OMS, 2017). Este estudio se enmarca en ese objetivo de la OMS para Colombia, donde el bajo peso al nacer es un problema significativo de salud neonatal, por lo que es esencial identificar y comprender los factores que contribuyen a esta problemática para desarrollar e implementar estrategias de prevención e intervención eficazmente.

En los últimos años, los avances en las tecnologías de *machine learning* han abierto nuevas oportunidades para abordar diversos problemas de salud pública. Estos modelos permiten analizar grandes volúmenes de datos con precisión y eficiencia, facilitando la identificación de patrones complejos y la predicción de resultados clínicos (Shailaja 2018). En el contexto de la salud pública, la predicción del bajo peso al nacer es un área crítica donde el *machine learning* puede desempeñar un papel fundamental. Al utilizar algoritmos avanzados, es posible prever riesgos y tomar medidas preventivas que mejoren la salud neonatal y materna (Faruk 2018).

Aplicar modelos de *machine learning* en Colombia es especialmente relevante debido a las características específicas del sistema de salud y las variables socioeconómicas del país. El sistema de salud colombiano enfrenta desafíos significativos, como la desigualdad en el acceso a los servicios de salud y la diversidad geográfica y demográfica. Las variables socioeconómicas, como la pobreza, el nivel educativo y el acceso a servicios básicos, varían considerablemente en diferentes regiones del país, afectando la salud materna e infantil (Salazar Blandón 2023). En este contexto, el uso de modelos predictivos puede ayudar a identificar grupos de alto riesgo y optimizar la asignación de recursos de salud.

Un enfoque predictivo utilizando modelos de *machine learning* proporciona información valiosa para identificar y abordar los riesgos de bajo peso al nacer en etapas tempranas. Este enfoque permite la implementación de intervenciones preventivas oportunas, mejorando así los resultados de salud neonatal. Al predecir el riesgo de bajo peso al nacer, los profesionales de la salud pueden diseñar programas específicos de apoyo y seguimiento para las madres en riesgo, contribuyendo a reducir la incidencia de esta condición.

Esta investigación se posiciona como una contribución innovadora dentro del panorama actual de estudios relacionados con el bajo peso al nacer y el uso de *machine learning* en salud pública. Al integrar un análisis detallado de variables socioeconómicas y aplicar técnicas avanzadas de *machine learning*, este estudio aporta una perspectiva novedosa y práctica para abordar un problema crítico de salud pública en Colombia.

## 4. Estado del Arte

Para predecir el BPN, varios investigadores han utilizado técnicas de *Machine Learning* (ML) aprovechando datos históricos que incluyen factores de salud materna, antecedentes médicos, hábitos de vida y datos socioeconómicos. Entre los métodos más comunes se encuentran los algoritmos de clasificación como árboles de decisión, regresión logística y redes neuronales. A continuación, se resumen los trabajos previos que guardan mayor relación con el presente.

Según Ticona *et al.* (2017), los OR son una medida clave en epidemiología y salud para evaluar la relación entre un factor de riesgo y un resultado. Los OR se definen como el cociente de las probabilidades condicionales de que ocurra un evento entre grupos expuestos y no expuestos, sin medir la incidencia directa. Un OR mayor a 1 indica una asociación positiva, uno menor a 1 una asociación negativa, y un OR igual a 1 significa que no hay asociación. Estas métricas son esenciales en estudios de casos y controles para interpretar asociaciones entre variables cuando no se puede determinar la incidencia directamente.

Pérez *et al.* (2017) estudiaron la relación entre diversas variables sociodemográficas presentes en las madres y familias con el BPN en una clínica universitaria en Chía, Cundinamarca, Colombia. Utilizaron un diseño de estudio transversal con 301 recién nacidos y aplicaron un análisis bivariado seguido de un modelo de regresión logística. Los resultados mostraron que un bajo nivel educativo materno se asocia significativamente con el BPN (OR, *odds ratio*, de 2.65), mientras que asistir a cuatro o más controles prenatales actúa como un factor protector (OR 0.34).

Quiñones (2020) investigó la incidencia de BPN en recién nacidos en Antioquia, Colombia, y su relación con características sociodemográficas de las madres entre 2014 y 2018. Utilizó un estudio descriptivo y de corte transversal con 6430 casos de BPN que significaron una prevalencia del 1.71%. Los riesgos que halló fueron la falta de pareja y la ausencia de educación formal. Por otro lado, encontró que el grupo de edad de 10 a 19 años y los embarazos previos actúan como factores protectores.

Monsreal *et al.* (2018) realizaron una evaluación multivariada de 17 variables en relación con el BPN en un estudio de cohorte retrospectiva en el Hospital Integral José María Morelos, Quintana Roo, México. Analizaron 1,147 nacimientos y encontraron que factores como estado civil no casada, peso materno bajo y menos de cinco consultas prenatales aumentan el riesgo de BPN.

En el estudio que realizaron Planchez *et al.* (2021) en Guanabacoa (Cuba), se aplicó un índice pronóstico para estratificar el riesgo de BPN utilizando un modelo de regresión logística. Se clasificó a las gestantes en categorías de riesgo y se demostró que el índice posee alta efectividad, aunque se destaca la necesidad de perfeccionarlo para mejorar su precisión.

Rivas (2021) analizó la prevalencia de BPN y factores asociados en Colombia durante 2021, utilizando datos del DANE. Encontró que el régimen subsidiado y nacimientos antes de las 36 semanas aumentan significativamente el riesgo de BPN. Además, la falta de control prenatal incrementa el riesgo de BPN.

Panduro (2022) utilizó el algoritmo de bosque aleatorio para predecir la anemia en niños en Perú, identificando variables clave como la edad del niño, altitud del conglomerado y número de visitas prenatales. Este enfoque demuestra la efectividad de los algoritmos de *machine learning* en la predicción basada en factores de riesgo comunes.

## 5. Metodología

Se empezó extrayendo los datos del Departamento Administrativo Nacional de Estadística para el período de 2017 a 2021 (DANE,2024). Fue crucial verificar la coherencia de los datos y se confirmó que estaban debidamente homogeneizados para mantener la integridad del análisis posterior.

Luego, se realizó un análisis exhaustivo de los datos para identificar y tratar valores atípicos, reconociendo la categoría de 4.000 o más gramos de peso pertenecientes a información sobre recién nacidos fisiológicamente excepcionales como se muestra en la figura 3 y tomando como objeto de estudio los datos pertenecientes a la categoría de 'Menos de 1.000' ya que es poco probable que contengan registros de neonatos con menos de 500 gr, porque sus probabilidades de sobrevivir son realmente bajas (Márquez-Beltrán 2013).

En este estudio se consideraron varias variables socioeconómicas que influyen en la probabilidad de bajo peso al nacer, tales como el estado civil de la madre, el nivel educativo de los padres, el acceso a servicios de salud prenatal, y las condiciones de vivienda. Estas variables son cruciales porque reflejan el entorno socioeconómico en el que se desarrolla el embarazo y pueden tener un impacto significativo en la salud del recién nacido. Sin embargo, se eliminaron las variables que tenían gran porcentaje de valores faltantes o que no contenían información. Otras variables fueron eliminadas por no estar directamente relacionadas con el objetivo del análisis o la pregunta de investigación en cuestión o por tener una de las categorías con un amplio porcentaje predominante y no presentar grandes aportes de variabilidad. Con las variables restantes nunca se usó el método de imputación de datos para reemplazar la categoría 'sin información´, ya que este representaba un pequeño porcentaje del total de datos (Yuan, 2024). Posteriormente, se condujo un análisis exploratorio detallado de las variables para comprender mejor las tendencias y patrones presentes en los datos.

Para tal objeto, se emplearon técnicas estadísticas descriptivas tanto gráficas como numéricas para hallar la información más relevante.

Una vez que se tuvo una comprensión sólida de la data, se implementaron modelos de regresión logística (Dharmaraj *et al*., 2024) y *Random Forest (*Ahmadi *et al*., 2017) para identificar los factores de riesgo que pueden aumentar la probabilidad de bajo peso al nacer en Colombia. También se abordó el desbalanceo en los datos, aplicando la técnica SMOTE (*Synthetic Minority Over-sampling Technique*) a la regresión logística y al *Random Forest* y penalización por compensar solo a la regresión.

Este último modelo, también conocido como regularización de Ridge, se emplea para para evitar el sobreajuste y mejorar la generalización del modelo (van Wieringen 2015).

La penalización se incorpora directamente en la función de pérdida, que en regresión logística busca minimizar la siguiente expresión:

$$\mathcal{L}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] + \lambda \sum_{j=1}^{p} \beta_j^2$$

Donde:

- $y_i$ es la clase real del ejemplo $i$.
- $p_i$ es la probabilidad predicha por el modelo para $y_i = 1$.
- $\beta_i$ son los coeficientes del modelo.
- $\lambda$ es el **parámetro de regularización** que controla el grado de penalización.

En la práctica, la implementación de regresión logística en *sklearn* emplea el parámetro $C = \frac{1}{\lambda}$, por lo que valores más pequeños de $C$ implican una penalización más fuerte sobre los coeficientes del modelo (*Scikit-learn developers*, 2024). Para determinar el valor óptimo de $C$, se aplicó validación cruzada estratificada mediante *GridSearchCV*, utilizando como métrica de evaluación el *recall*, dado que el objetivo principal es minimizar los falsos negativos. Se probaron los valores $C = 0.1,\ 1,\ 10$ , encontrando que el mejor desempeño se logró con $C = 1$ , alcanzando un *recall* promedio de 0.828 en los conjuntos de validación. Este resultado sugiere un buen equilibrio entre ajuste y

capacidad de generalización del modelo, favoreciendo la identificación correcta de los casos positivos de BPN.

Después de obtener los resultados de cada modelo, la eficacia de cada uno fue evaluada utilizando las métricas como precisión, *recall* y F1-*score*. Para la discusión sobre factores de riesgo y la interpretación más detallada de resultados, se seleccionó el modelo con el mejor rendimiento que fue penalización por compensar, utilizando como criterio la métrica *recall* debido a que Jafarigol y Trafalis (2023) la definen como la proporción de instancias positivas en los datos de prueba que se etiquetaron correctamente, es decir, tiene una mayor sensibilidad al detectar los casos de bajo peso al nacer correspondientes a la clase minoritaria. Todos los análisis se hicieron en lenguaje *Python* versión 3.9.12 y con el paquete *scikitlearn,* versión 1.2.2.

Para garantizar la integridad del proceso de entrenamiento y evaluación, se dividieron los datos en conjuntos de entrenamiento y prueba, utilizando el 70% de los datos para el entrenamiento y el 30% para la prueba. Es importante destacar que el conjunto de prueba no se utilizó en ningún momento para entrenar el modelo ni para la optimización de hiperparámetros, asegurando que las métricas de ajuste se calcularon exclusivamente en el conjunto de prueba. Las predicciones realizadas en el conjunto de prueba permitieron evaluar el rendimiento del modelo mediante el informe de clasificación y la precisión. Adicionalmente, se aplicó validación cruzada con cinco pliegues en todo el conjunto de datos para evaluar la consistencia del modelo, manteniendo el conjunto de prueba separado para la evaluación final.

# 6. Resultados

## 6.1 Descripción de la muestra y de las variables

El conjunto de datos de análisis consta de 2.946.829 elementos con información de 13 variables diferentes, lo que indica un tamaño considerable. A continuación, se explicará n todas las variables analizadas en este estudio:

I. **Código Departamento**: Esta variable almacena el código del departamento do nde ocurrieron los eventos relacionados con la salud materno-infantil. Su esc ala de medición es nominal.

II. **Sexo del Recién Nacido**: Indica el sexo del recién nacido, que puede ser "Masculino," "Femenino", o "Indeterminado." Su escala de medición es nominal.

III. **Peso del Recién Nacido**: Representa el peso del recién nacido al nacer y se clasifica en rangos como "Menos de 1000", "1000 - 1499", "1500 - 1999" y otros. Funciona como ordinal, aunque proviene de datos cuantitativos continuos. Se discretizó por intervalos. Esta variable se transformó en binaria, siendo 1 (positivo) cuando el rango es menor a 2500 gramos y constituye nuestra variable endógena o de respuesta correspondiente al bajo peso al nacer y 0 para peso normal, correspondiente al rango de 2500 gramos o más. En la figura 1 se muestra la distribución de dicha variable.

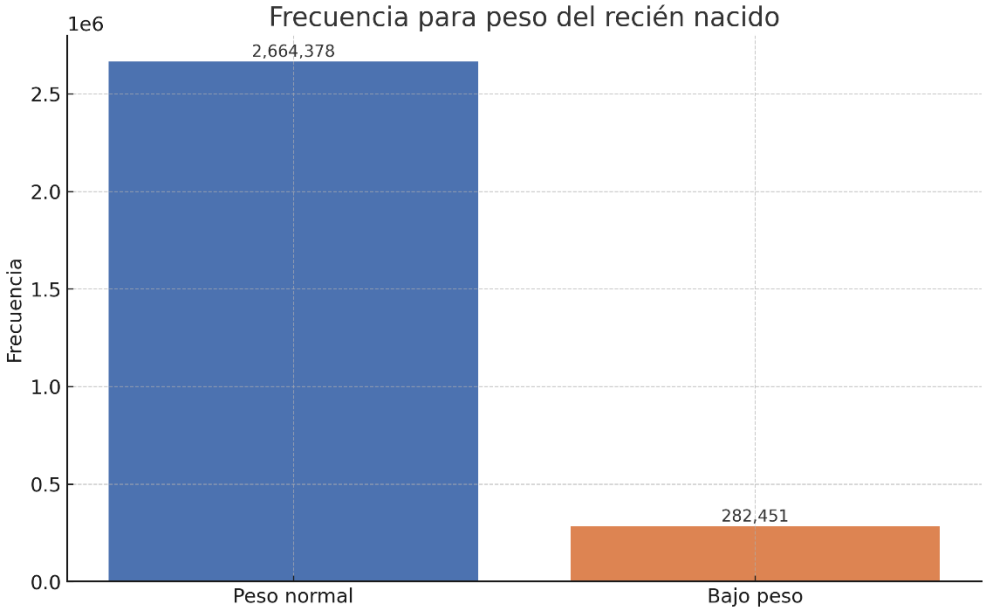


***Figura 1***

*Diagrama de barras para peso del recién nacido. Elaboración propia*

IV. **Año de Registro:** Contiene el año en que se registraron los eventos relacionados con la salud materno-infantil. Su escala y tipo es de intervalo y cuantitativa discreta.

V. **Tiempo de Gestación:** Representa la duración de la gestación en semanas y se agrupa en categorías como "Menos de 22," "De 22 a 27," "De 28 a 37," y otros. Funciona como ordinal, aunque proviene de datos cuantitativos continuos. Se discretizó por intervalos.

VI. **Número de Consultas:** Indica el número de consultas médicas durante el período de embarazo. Su tipo es cuantitativa discreta, escala de razón.

VII. **Multiplicidad del Parto:** Indica la multiplicidad del parto, que puede ser "Simple," "Doble," "Triple," "Cuádruple o más," o "Sin información." Su tipo es cuantitativa discreta, escala de razón.

VIII. **Edad de la Madre:** Representa la edad de la madre en categorías como "De 10-14 Años," "De 15-19 Años," "De 20-24 Años," y otras. Funciona como ordinal, aunque proviene de datos cuantitativos continuos. Se discretizó por intervalos.

IX. **Estado Civil de la Madre:** Describe el estado civil de la madre, incluyendo "No está casada y lleva dos o más años viviendo con su pareja," "Está casada," "Está soltera," y otros. Su escala de medición es nominal.

X. **Nivel Educativo de la Madre:** Indica el nivel educativo de la madre, con categorías que incluyen "Preescolar," "Profesional," "Maestría" y “Doctorado". Su escala es ordinal.

XI. **Número de Hijos Vivos:** Indica el número de hijos vivos de la madre. Su tipo es cuantitativa discreta, escala de razón.

XII. **Número de Embarazos:** Indica el número de embarazos. Su tipo es cuantitativa discreta, escala de razón.

XIII. **Seguro Social:** Describe el tipo de seguro social, que puede ser "Contributivo," "Subsidiado," "Vinculado," y otros. Su escala es nominal.

### 6.2 Análisis descriptivo

El análisis de los registros de nacimientos en Colombia nos da una visión clara de diversas características de los nacimientos en el país. La distribución de nacimientos con bajo peso al nacer en Colombia revela que Bogotá presenta la mayor frecuencia relativa (14,39%), seguida por Nariño (10,42%), Sucre (9,96%) y Antioquia (9,76%). Otros departamentos como Atlántico (9,45%) y Valle del Cauca (9,51%) también registran proporciones significativas. En contraste, departamentos como Guaviare (3,72%), Putumayo (4,67%) y Amazonas (5,35%) presentan frecuencias relativas muy bajas, lo que indica un número reducido de nacimientos con bajo peso en comparación con el resto del país. Estos resultados reflejan diferencias regionales que pueden estar asociadas a factores socioeconómicos y de acceso a servicios de salud.

Para ilustrar mejor estas variaciones, en la Figura 2 se presenta un mapa de calor que muestra la distribución de la tasa de bajo peso al nacer (BPN) por departamento en Colombia durante el periodo 2017-2021.

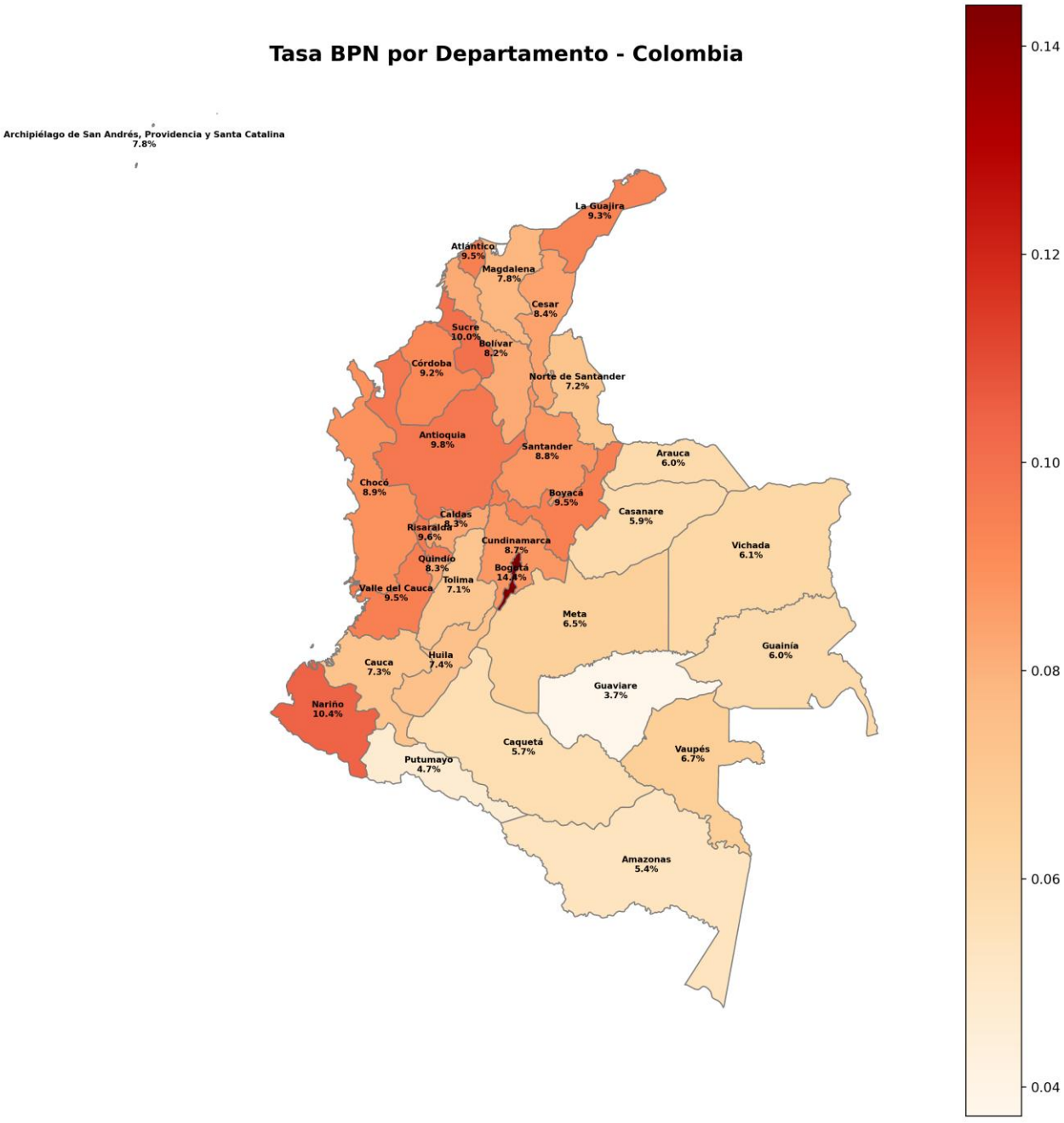


***Figura 2***

*Mapa de calor: BPN al nacer en Colombia por departamento de 2017 a 2021. Elaboración propia.*

La distribución de nacimientos con bajo peso al nacer (BPN) según el sexo en Colombia muestra que los nacimientos de sexo femenino presentan una mayor frecuencia relativa de BPN (10,30%) en comparación con los nacimientos de sexo masculino (8,89%). Esto indica que el sexo femenino está asociado con un mayor riesgo de bajo peso al nacer en el país al representar el 53,7% de los nacimientos con bajo peso. En la figura 3 se aprecia su distribución.

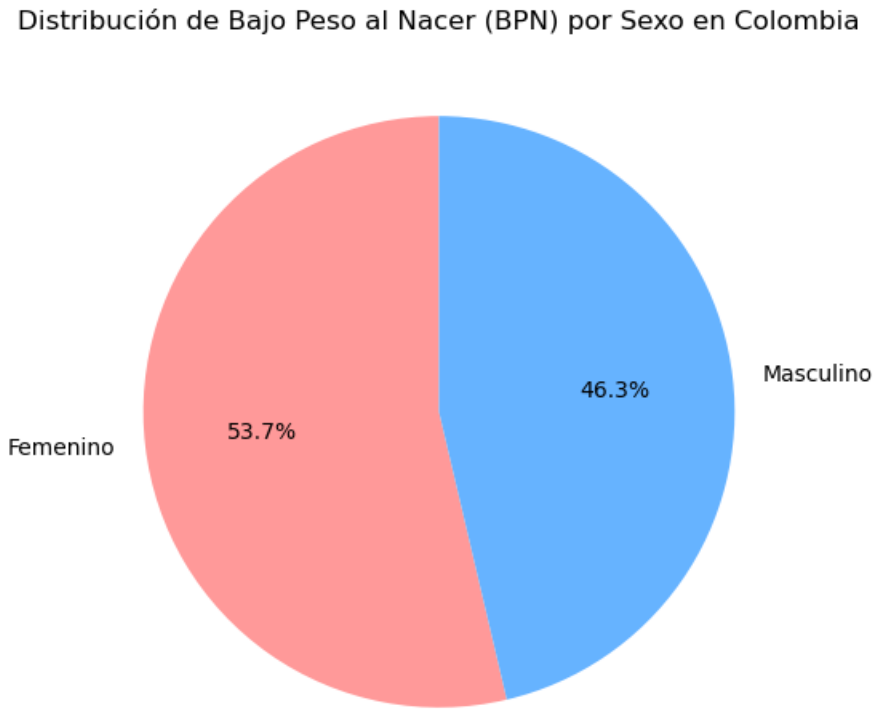


***Figura 3***

*Diagrama de pastel para la categoría sexo. Elaboración propia*

La mayoría de los bebés nacen con un peso entre 3.000 y 3.499 gramos (1,293,436 registros). Le siguen los bebés que pesan entre 2.500 y 2.999 gramos (840,527 registros) y entre 3.500 y 3.999 gramos (530,415 registros). Esta distribución se aprecia en la figura 4.

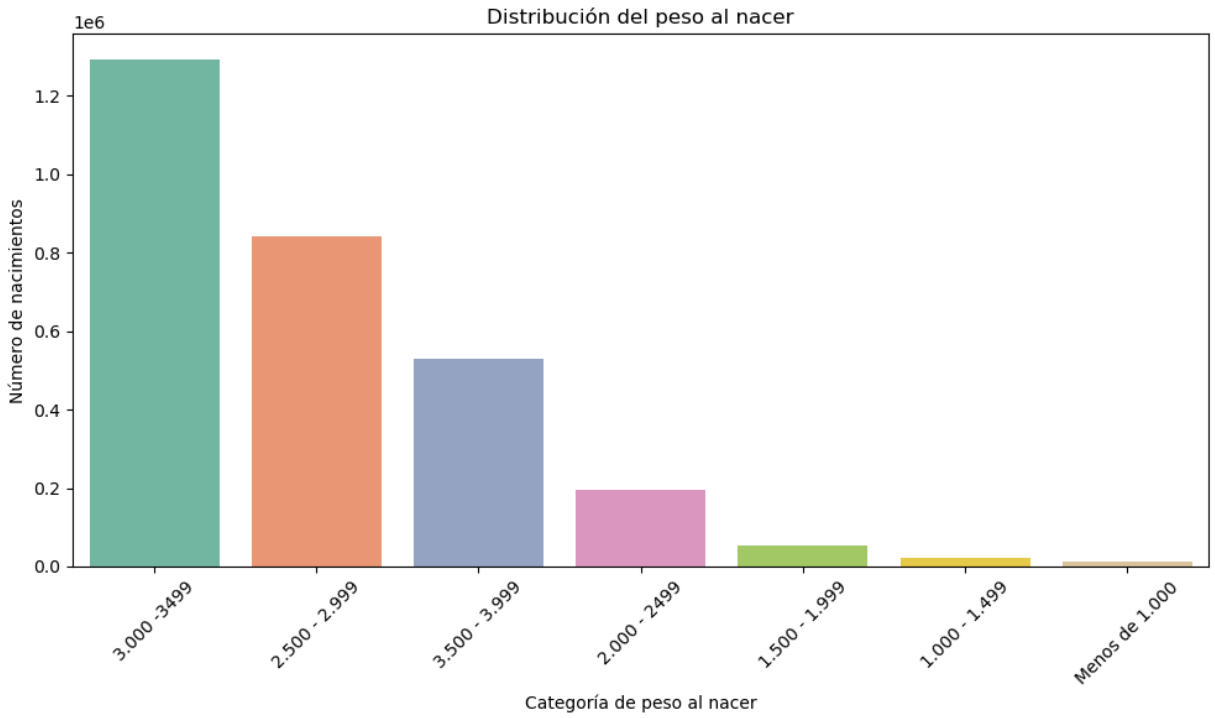


***Figura 4***

*Diagrama de barras para el peso del nacido en gramos. Elaboración propia.*

La distribución de nacimientos con bajo peso al nacer (BPN) en Colombia muestra que en el año 2021 se registró la mayor frecuencia relativa de casos (10,08%), seguido por 2019 con una frecuencia relativa de 9,66%. Otros años como 2018 y 2020 también presentan proporciones significativas, con 9,44% y 9,40% respectivamente. En contraste, el año 2017 reportó la frecuencia relativa más baja de BPN (9,33%), lo que indica una ligera variación en la prevalencia de BPN a lo largo de los años considerados como se aprecia en la siguiente figura:

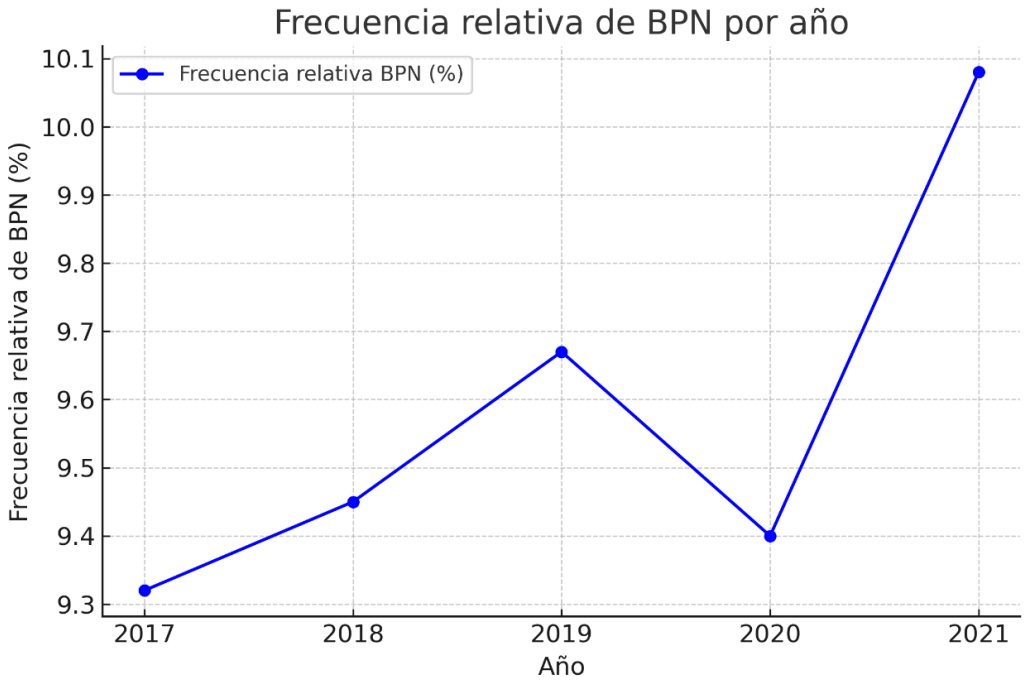


***Figura 5***

*Gráfico de línea para la categoría año de BPN. Elaboración propia.*

La distribución de bajo peso al nacer (BPN) según el tiempo de gestación muestra que la mayor proporción de casos ocurre entre 22 y 27 semanas, con una frecuencia relativa del 99,45 %, lo que evidencia un riesgo muy alto en nacimientos muy prematuros. Le sigue el grupo de menos de 22 semanas, con una frecuencia relativa de 98,82%, también indicando un riesgo significativo. Por otro lado, el grupo de 28 a 37 semanas presenta una frecuencia relativa de 34,79%, mientras que los grupos de 38 a 41 semanas y 42 semanas o más muestran frecuencias relativas mucho menores, de 2,17% y 1,60% respectivamente, lo que sugiere que el riesgo de BPN disminuye considerablemente a medida que aumenta el tiempo de gestación como se muestra en la figura 6.

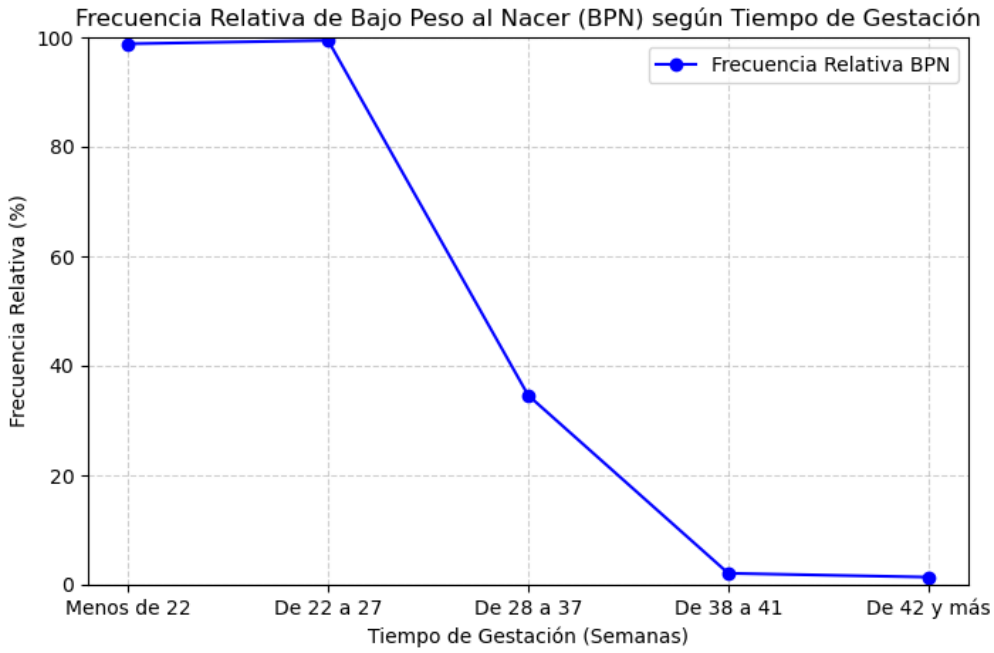


***Figura 6***

*Gráfico de línea para la categoría tiempo de gestación. Elaboración propia.*

El análisis de la frecuencia relativa del número de consultas muestra que la mayor proporción de casos de bajo peso al nacer (BPN) se concentra en el grupo con 24 consultas (22.85%). Otros grupos con frecuencias relativas significativas incluyen aquellos con 25 consultas (22.37%) y 20 consultas (22%). En contraste, los grupos con menor frecuencia relativa corresponden a aquellos con 9 consultas (4.90%) y 11 consultas (5.94%). Esto sugiere que el número de consultas podría estar asociado con variaciones en la prevalencia de BPN, destacando la importancia de un seguimiento continuo durante el embarazo.

El análisis de la frecuencia relativa del tipo de parto muestra que la mayor proporción de casos de bajo peso al nacer (BPN) se presenta en los partos triples, con una frecuencia relativa del 81.81%. Le siguen los partos dobles, con una proporción del 68.68%. En contraste, los partos cuádruples o más presentan una frecuencia relativa significativamente menor del 9.09%, mientras que los partos simples tienen la frecuencia relativa más baja, con un 8.51%. Esto sugiere que los partos múltiples están asociados con mayor prevalencia de BPN que los simples.

El análisis de la frecuencia relativa de bajo peso al nacer (BPN) según la edad de la madre muestra que la mayor proporción de casos se presenta en el grupo de madres de 45 a 49 años, con una frecuencia relativa de 17.05%. Le siguen las madres de 40 a 44 años, con un 13.13%, y de 10 a 14 años, con un 12.75%. En contraste, el grupo con la menor frecuencia relativa corresponde a las madres de 50 a 54 años, con un 6.58%. Esto sugiere que las madres adolescentes y las mayores tienen una mayor prevalencia de bajo peso al nacer que las madres jóvenes y adultas.

El análisis de la frecuencia relativa según el estado civil muestra que la mayor proporción de casos se presenta en mujeres que están solteras, con una frecuencia relativa de 11.02% y que representa un 19.1% del total de casos de BPN. Le siguen aquellas que no están casadas y llevan menos de dos años viviendo con su pareja, con una frecuencia relativa del 9.97%, y las que están separadas o divorciadas, con un 9.84%. En contraste, el grupo con la menor frecuencia relativa corresponde a las mujeres que están viudas, con un 8.68% y que representa un 14.9% del total de casos de Bajo Peso al Nacer. Esto indica que el estado civil puede estar relacionado con la distribución de los casos, siendo más frecuente en mujeres solteras, así como se aprecia en la figura 7.

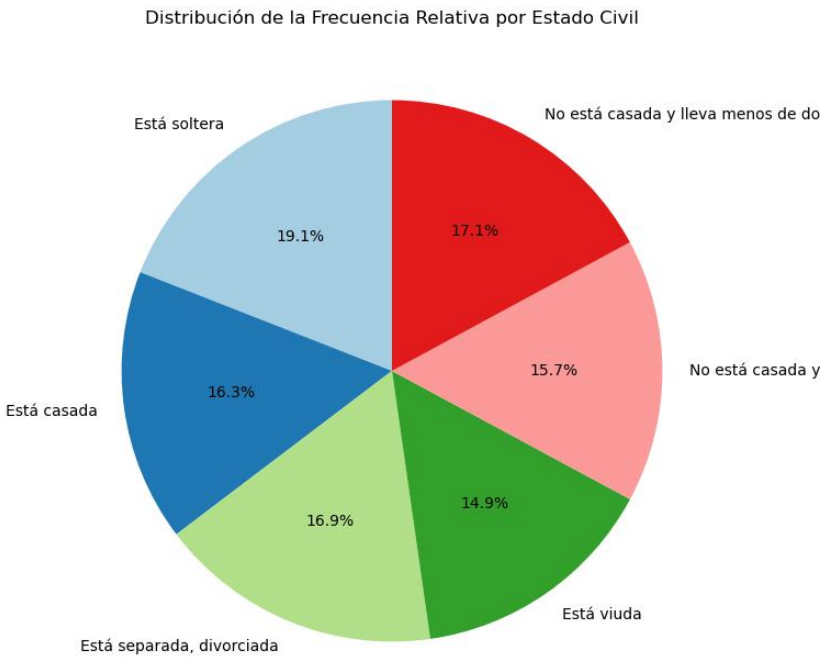


***Figura 7***

*Gráfico de línea para la categoría Estado civil. Elaboración propia.*

El análisis de las frecuencias relativas para el nivel educativo muestra que el mayor porcentaje de BPN se observa en el nivel educativo de Doctorado con un 11.09%, seguido de Maestría con un 10.64% y Ninguno con un 10.55%. A continuación, se encuentran los niveles de, Especialización con un 9.93%, Media académica o clásica con un 9.58% y Profesional con un 9.29%. Por otro lado, los niveles educativos con menor proporción de BPN son Normalista con un 6.08%, Media técnica con un 8.42% y Básica secundaria con un 8.94%. Esto sugiere que los niveles educativos más altos o ninguno podrían asociarse con más casos de BPN por la posible edad alta de la madre o el desconocimiento sobre cuidados prenatales.

La distribución de bajo peso al nacer (BPN) en relación con el número de hijos vivos y el número de embarazos muestra patrones interesantes. En el primer caso, las categorías de entre 1 y 12 hijos presentan frecuencias relativas de BPN entre el 8% y el 10%, lo que indica que, en estos grupos, la proporción de casos es estable. No obstante, en las categorías con números extremos de hijos, por ejemplo, el grupo de 16 hijos presenta una frecuencia relativa inusualmente alta del 25%, mientras que las categorías de 17 a 20 hijos muestran 0% de BPN, lo que sugiere que estos valores atípicos pueden estar influenciados por un tamaño muestral muy reducido. Por otro lado, en el análisis del número de embarazos, la frecuencia relativa de BPN varía de forma moderada en las categorías principales (de 1 a 16 embarazos), ubicándose generalmente entre el 8% y el 11%, con un pico del 11.48% en el grupo de 12 embarazos y un valor mínimo del 6.72% en la categoría de 13 embarazos. Las categorías con muy pocos casos, como la de 19 embarazos (con un 100% de BPN) o aquellas que registran 0%, subrayan la importancia de considerar el tamaño muestral al interpretar estos porcentajes. En conjunto, estos resultados indican que, en los grupos con tamaños muestrales representativos, la prevalencia de BPN se mantiene en un rango moderado, mientras que las variaciones observadas en los extremos deben tomarse con cautela.

La distribución de nacimientos con bajo peso al nacer (BPN) según el régimen de seguridad social en Colombia muestra que la mayor frecuencia relativa de BPN se presenta en el régimen de No asegurado con un 10.16%, representando un 21.34% del total de casos registrados de BPN; seguido por el régimen Especial con un 10.04%, que representa un 21.10% del total de casos de BPN. El régimen Contributivo presenta una frecuencia relativa considerable del 10.07%. En contraste, los regímenes Subsidiado y Excepción tienen frecuencias relativas menores, del 8.71% y 8.63% respectivamente, lo que sugiere una menor proporción de nacimientos con bajo peso en estos grupos en comparación con los demás, así como se demuestra en la figura 8.

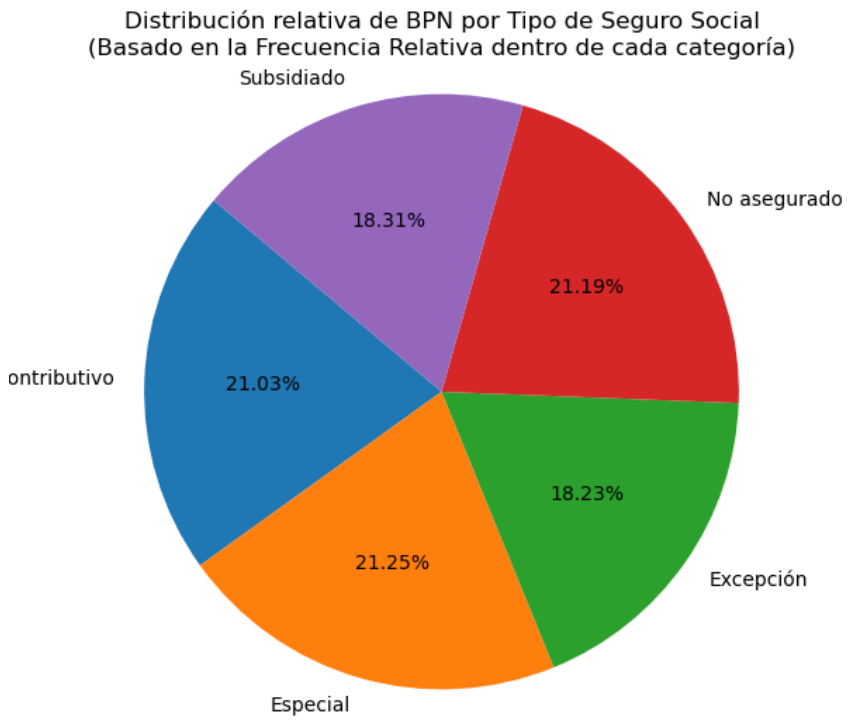


***Figura 8***

*Gráfico de línea para la categoría Tipo de Seguro Social. Elaboración propia.*

## 6.3 Resultados de los modelos

### 6.3.1 Comparación de matrices de confusión.

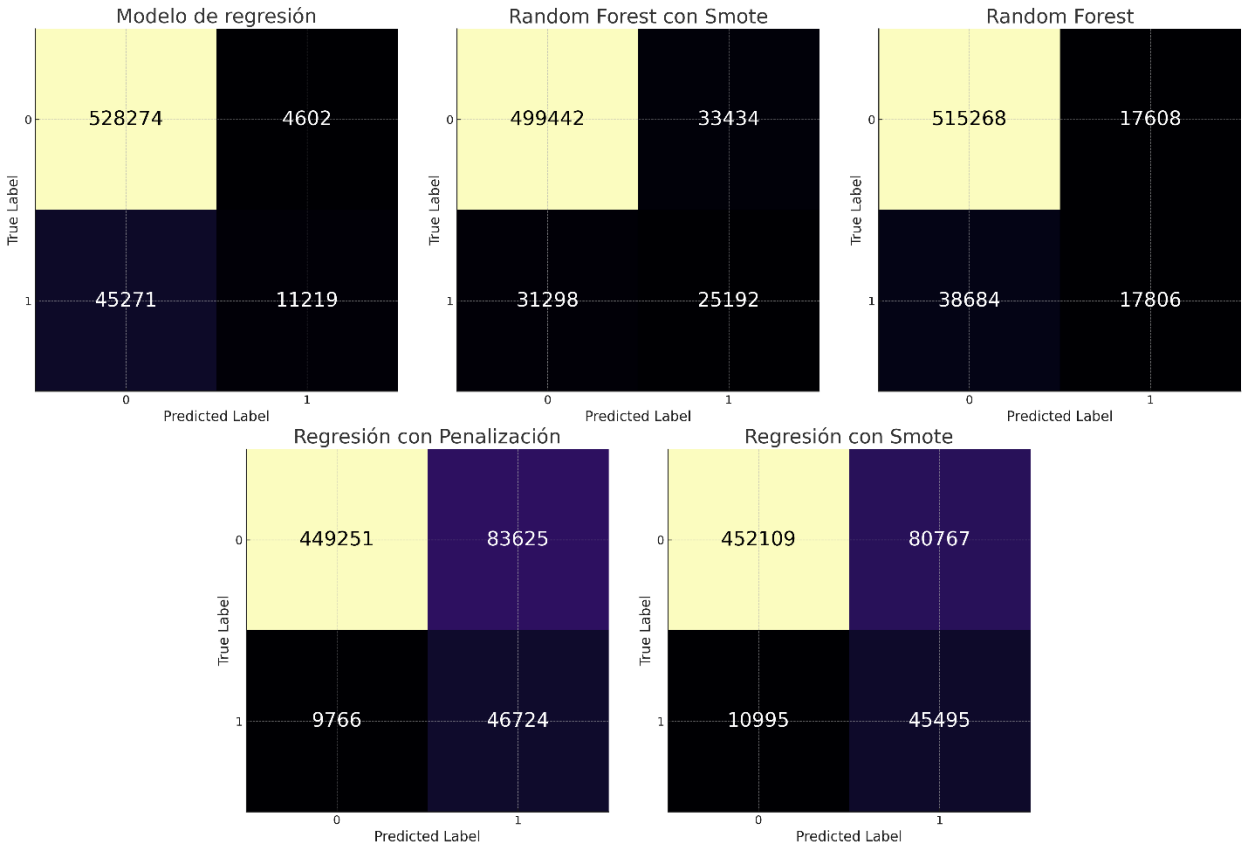


***Figura 9***

*Matrices de confusión para los modelos usados. Elaboración propia.*

#### 6.3.1.1 Análisis comparativo de los modelos

En el contexto de la predicción de bajo peso al nacer (BPN), es esencial minimizar los falsos negativos, es decir, los casos en los que un recién nacido con bajo peso no es identificado correctamente, dado que las consecuencias para su salud pueden ser graves. En la Figura 9 se presentan las matrices de confusión correspondientes a los cinco modelos evaluados.

El **modelo de regresión simple** concentra un alto número de verdaderos negativos (528,274) y falsos negativos (45,271), mientras que logra identificar relativamente pocos verdaderos positivos (11,219). Esto refleja la dificultad de un modelo lineal para capturar patrones complejos, lo que lo lleva a subestimar la clase minoritaria y a no detectar un gran número de casos de bajo peso.

El **modelo de *Random Forest*** muestra una mejora respecto a la regresión simple: aumenta el número de verdaderos positivos (17,806) y reduce los falsos negativos (38,684). Sin embargo, este avance viene acompañado de un incremento de falsos positivos (17,608), lo que evidencia que, aunque la técnica es más robusta, sigue limitada por el desbalance de clases.

La **regresión con *SMOTE*** incrementa de manera considerable los verdaderos positivos (45,495) y reduce los falsos negativos (10,995). Este resultado se debe a la generación de ejemplos sintéticos de la clase minoritaria, lo que permite al modelo aprender mejor sus características. No obstante, se observa un aumento notable en falsos positivos (80,767).

De forma similar, el ***Random Forest* con *SMOTE*** alcanza 25,192 verdaderos positivos, con una reducción de falsos negativos (31,298), aunque también con un número importante de falsos positivos (33,434). La combinación de un algoritmo más robusto con datos balanceados potencia la detección de casos de bajo peso, pero también incrementa la probabilidad de clasificar erróneamente a bebés con peso normal como de bajo peso.

Finalmente, la **regresión con penalización** muestra un desempeño más equilibrado. Presenta 46,724 verdaderos positivos y solo 9,766 falsos negativos, al tiempo que mantiene un nivel de falsos positivos (83,625) comparable al de la regresión con *SMOTE*. El ajuste de los costos en las predicciones incorrectas aumenta la sensibilidad hacia la clase minoritaria, permitiendo reducir los casos no detectados de bajo peso, aspecto clave dado su impacto en la salud neonatal.

### 6.3.2 Comparación de las métricas de desempeño.

| Modelo | Accuracy | Recall | Precision | F1-Score |
|---|---|---|---|---|
| Regresión simple | 0.92 | 0.20 | 0.71 | 0.31 |
| Random Forest | 0.90 | 0.32 | 0.50 | 0.39 |
| Regresión con SMOTE | 0.84 | 0.81 | 0.36 | 0.50 |
| Random Forest con SMOTE | 0.89 | 0.45 | 0.43 | 0.44 |
| Regresión con penalización | 0.84 | 0.83 | 0.36 | 0.50 |

*Tabla 1: Desempeño de los modelos para detectar el bajo peso al nacer. Fuente: Elaboración propia*

#### 6.3.2.1. Regresión simple

El modelo de regresión simple presenta una alta *accuracy* (0.92), lo que refleja aciertos en la mayoría de las predicciones. Sin embargo, su recall para bajo peso al nacer es bajo (0.20), indicando que detecta pocos casos de esta condición. La precisión para la clase minoritaria es moderada (0.71), es decir, cuando predice bajo peso, suele acertar, aunque son pocos los casos detectados. El *F1-Score* (0.31) confirma un pobre equilibrio entre precisión y *recall*. En conclusión, aunque el modelo es globalmente preciso y consistente, su baja sensibilidad hacia los casos de bajo peso limita su efectividad.

#### 6.3.2.2. *Random Forest*

El modelo *Random Forest* presenta una *accuracy* de 0.90. Su *recall* para bajo peso es de 0.32, lo que indica que mejora levemente en la detección respecto a la regresión simple, aunque sigue siendo bajo. La precisión es de 0.50, de modo que acierta en la mitad de las veces que predice bajo peso. El F1-Score (0.39) muestra un balance limitado entre precisión y *recall.* En conclusión, *Random Forest* ofrece mayor sensibilidad que la regresión simple, pero sigue sin detectar un número suficiente de casos.

#### 6.3.2.3. Regresión con *Smote*

El modelo de regresión con SMOTE logra una *accuracy* de 0.84 y un *recall* muy alto (0.81), lo que significa que identifica la mayoría de los casos de bajo peso al nacer. No obstante, su precisión (0.36) es baja, indicando que muchas predicciones positivas son incorrectas. El *F1-Score* (0.50) refleja un balance moderado entre precisión y recall. La validación cruzada (0.84) muestra consistencia aceptable. En conclusión, este modelo es altamente sensible y permite detectar la mayoría de los casos, aunque a costa de un mayor número de falsos positivos.

#### 6.3.2.4. *Random Forest* con *Smote*

Este modelo presenta una *accuracy* de 0.89 y un *recall* moderado (0.45), detectando algunos casos de bajo peso al nacer, aunque sin llegar al nivel de la regresión con *SMOTE*. La precisión es de 0.43, y el *F1-Score* (0.44) refleja un balance moderado. En conclusión, este modelo mantiene un balance razonable entre precisión y recall, pero su consistencia es más débil.

#### 6.3.2.5. Regresión con penalización

El modelo de regresión con penalización alcanza una *accuracy* de 0.84 y un *recall* alto (0.83), lo que significa que detecta la mayoría de los casos de bajo peso al nacer. Sin embargo, su precisión es baja (0.36), indicando un número elevado de falsos positivos. El *F1-Score* (0.50) refleja un balance moderado. En conclusión, este modelo es una de las mejores opciones para detectar bajo peso al nacer, gracias a su alta sensibilidad y consistencia, aunque sacrifica precisión.

La Figura 10 muestra la curva ROC del modelo, con un AUC de 0.88. Según guías clínicas, un valor por encima de 0.80 se considera clínicamente útil, y en algunos esquemas se califica como bueno o excelente (Çorbacıoğlu, 2023). Este resultado indica que el modelo discrimina eficazmente entre recién nacidos con y sin bajo peso al nacer. Además, estudios en el ámbito del bajo peso neonatal han reportado valores similares (AUC ≈ 0.90), lo que respalda la robustez discriminativa que sugiere esta métrica (Çorbacıoğlu, 2023). Como la métrica AUC es independiente del umbral, permite una evaluación global del rendimiento del modelo, útil para la comparación con otros modelos y su interpretación en escenarios clínicos variados.

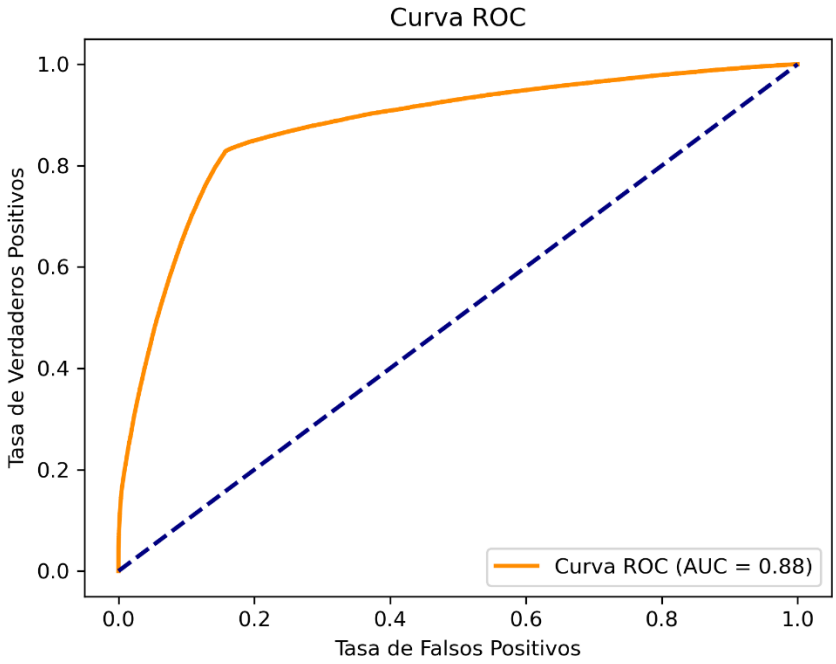


***Figura 10***

*Curva Roc para el modelo de Regresión con penalización. Elaboración propia.*

De manera complementaria, la Figura 11 presenta la curva *Precision–Recall*, con una Precisión Promedio (AP) de 0.52. Esta métrica adquiere especial relevancia en situaciones de marcado desbalance de clases, como en el presente análisis, donde los casos de bajo peso al nacer representan una proporción reducida. La trayectoria de la curva evidencia que el modelo conserva un nivel elevado de recall, asegurando la detección de la mayoría de los casos positivos; sin embargo, la precisión disminuye de forma progresiva conforme aumenta la sensibilidad, lo que refleja el equilibrio inevitable entre maximizar la cobertura de detección y asumir un mayor número de falsos positivos. En este contexto, la curva *Precision–Recall* ofrece una apreciación más clara y reveladora del comportamiento del modelo que la proporcionada exclusivamente por la curva ROC en escenarios con clases desbalanceadas (Tissera, 2025).

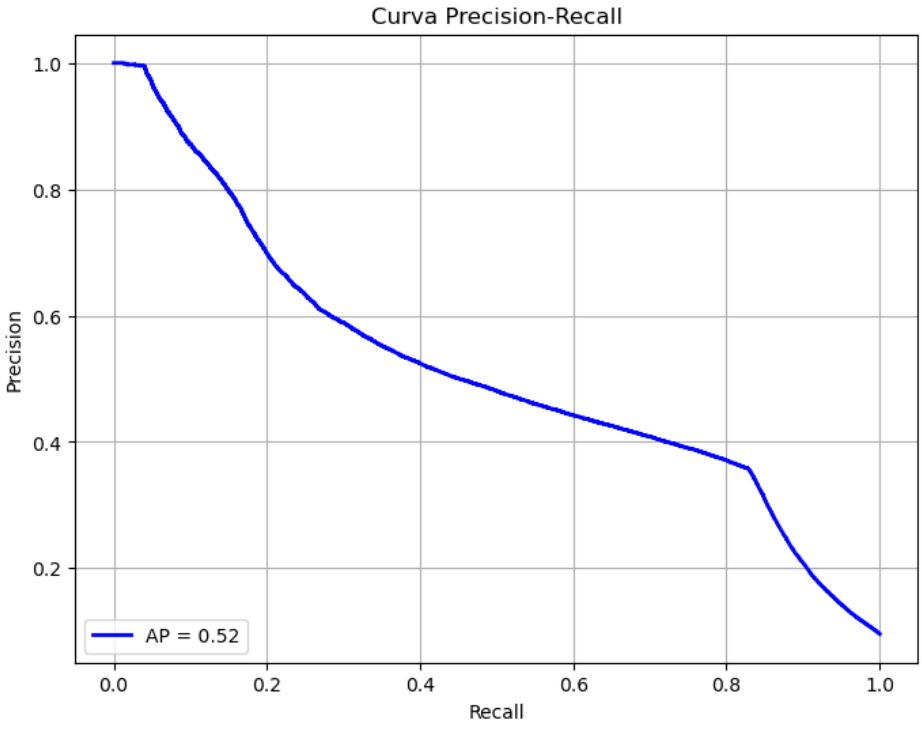


***Figura 11***

*Curva Precision–Recall para el modelo de Regresión con penalización. Elaboración propia.*

En conjunto, tanto las métricas como las curvas ROC y *Precision–Recall* evidencian que la regresión con penalización ofrece un desempeño sólido para la detección de bajo peso al nacer. Es una de las mejores opciones cuando el objetivo es minimizar falsos negativos, ya que combina alta sensibilidad con estabilidad, aun sacrificando precisión.

### 6.3.3 Discusión de relevancia de variables

| Variable | Coef. | Odds Ratio | IC 2.5% | IC 97.5% | Error est. | z | p-valor |
|---|---|---|---|---|---|---|---|
| Número de embarazos | 0.0263 | 1.0266 | 1.0176 | 1.0358 | 0.0045 | 5.82 | 5.7e-09 |
| Departamento de Antioquia | 0.0940 | 1.0986 | 1.0691 | 1.1289 | 0.0139 | 6.76 | 1.3e-11 |
| Departamento del Atlántico | 0.0281 | 1.0285 | 1.0072 | 1.0504 | 0.0107 | 2.61 | 9.0e-03 |
| Distrito Capital Bogotá | 0.2631 | 1.3007 | 1.2611 | 1.3422 | 0.0159 | 16.54 | 0 |
| Departamento de Boyacá | 0.0806 | 1.0839 | 1.0700 | 1.0980 | 0.0066 | 12.24 | 0 |
| Departamento de Caldas | 0.0190 | 1.0192 | 1.0088 | 1.0298 | 0.0053 | 3.62 | 2.9e-04 |
| Departamento del Chocó | 0.0168 | 1.0169 | 1.0089 | 1.0250 | 0.0040 | 4.14 | 3.4e-05 |
| Departamento de Cundinamarca | 0.0682 | 1.0705 | 1.0539 | 1.0875 | 0.0080 | 8.55 | 0 |
| Departamento de La Guajira | 0.0521 | 1.0535 | 1.0391 | 1.0681 | 0.0070 | 7.43 | 1.6e-13 |
| Departamento de Nariño | 0.0411 | 1.0419 | 1.0271 | 1.0569 | 0.0073 | 5.63 | 1.8e-08 |
| Departamento del Quindío | 0.0239 | 1.0242 | 1.0153 | 1.0332 | 0.0045 | 5.38 | 7.8e-08 |
| Departamento de Risaralda | 0.0241 | 1.0244 | 1.0128 | 1.0361 | 0.0058 | 4.16 | 3.2e-05 |
| Departamento de Sucre | 0.0296 | 1.0301 | 1.0186 | 1.0443 | 0.0070 | 4.24 | 2.5e-05 |
| Departamento del Valle del Cauca | 0.0646 | 1.0665 | 1.0231 | 1.0714 | 0.0117 | 3.91 | 9.1e-05 |
| Edad de la madre 30–34 años | 0.0380 | 1.0387 | 1.0235 | 1.0541 | 0.0075 | 5.05 | 4.4e-07 |
| Edad de la madre 35–39 años | 0.0776 | 1.0807 | 1.0683 | 1.0932 | 0.0059 | 13.16 | 0 |
| Edad de la madre 40–44 años | 0.0789 | 1.0812 | 1.0741 | 1.0884 | 0.0039 | 22.99 | 0 |
| Edad de la madre 45–49 años | 0.1321 | 1.1413 | 1.0741 | 1.0884 | 0.0019 | 17.23 | 0 |
| Parto múltiple doble | 0.2017 | 1.2234 | 1.1542 | 1.2969 | 0.0297 | 6.78 | 1.1e-11 |
| Parto múltiple triple | 0.0379 | 1.0386 | 1.0261 | 1.0492 | 0.0051 | 7.34 | 2.1e-13 |
| Nivel educativo ninguno | 0.0172 | 1.0174 | 1.0138 | 1.0209 | 0.0025 | 6.99 | 3.0e-12 |
| Nivel educativo preescolar | 0.0049 | 1.0049 | 1.0015 | 1.0084 | 0.0018 | 2.81 | 4.8e-03 |
| Seguridad social no asegurado | 0.0169 | 1.0171 | 1.0130 | 1.0212 | 0.0021 | 8.24 | 2.2e-16 |
| Seguridad social subsidiado | 0.0389 | 1.0396 | 1.0349 | 1.0444 | 0.0023 | 16.59 | 0 |

*Tabla 2: Resultados del análisis de regresión con penalización. Fuente: Elaboración propia.*

En este análisis, se agrupan las variables más relevantes en cuatro categorías principales: edad de la madre, número de partos, nivel educativo y región geográfica. Además, se discuten algunas variables adicionales que, aunque potencialmente importantes, no se pudieron incorporar debido a la limitación de los datos.

#### 6.3.3.1 Edad de la Madre

La edad materna constituye un factor determinante en los resultados perinatales, con riesgos que varían según el grupo de edad. Las madres adolescentes presentan vulnerabilidades biológicas y socioeconómicas que incrementan la probabilidad de bajo peso al nacer (BPN), partos prematuros y restricción del crecimiento fetal, debido a inmadurez física y menor acceso a cuidados prenatales (Bendezú, 2016). Por otro lado, los resultados de este análisis muestran que las madres de 30–49 años tienen *odds ratio* mayores a 1 frente al grupo de referencia de 10–14 años, lo que indica un aumento progresivo en el riesgo de BPN. El riesgo se vuelve más notorio a partir de los 30 años y alcanza su punto máximo en las de 45–49 años (OR = 1.14), lo que significa un 14% más de probabilidad de tener un hijo con BPN en comparación con las más jóvenes. Estos hallazgos concuerdan con la literatura, que documenta una mayor prevalencia de complicaciones como hipertensión y diabetes gestacional en mujeres mayores de 35 años, condiciones que afectan el desarrollo fetal y aumentan los resultados adversos neonatales (Pérez, 2011).

Según (Ticona, 2017) El OR se calcula comparando la probabilidad de BPN en los distintos grupos de edad, a través de la fórmula:

$$OR = \frac{Odds\ de\ tener\ la\ enfermedad}{Odds\ de\ tener\ no\ la\ enfermedad}$$

Para este caso en particular, el valor de OR=1,14 para el grupo de edad materna de 45 – 49 años se obtuvo calculado:

$$OR = \frac{(casos\ con\ BPN\ en\ el\ grupo\ de\ 45-49\ años)}{(casos\ \text{sin}\quad BPN\ en\ el\ grupo\ de\ 45-49\ años)} \Big/ \frac{(casos\ con\ BPN\ en\ el\ grupo\ de\ 10-14\ años)}{(casos\ \text{sin}\quad BPN\ en\ el\ grupo\ de\ 10-14\ años)}$$

#### 6.3.3.2 Número de Partos

La experiencia maternal puede influir significativamente en los resultados del embarazo. El primer hijo tiene madres primíparas, que sufren mayor estrés emocional y físico por falta de experiencia, lo que puede afectar negativamente el resultado del embarazo. La literatura sugiere que los primeros embarazos son más propensos a complicaciones que pueden llevar al bajo peso al nacer. Por otro lado, las madres con múltiples partos, o multíparas, pueden presentar riesgos aumentados si han tenido complicaciones en embarazos anteriores. Sin embargo, cuando los embarazos previos han sido saludables y bien gestionados, la experiencia adquirida puede contribuir a un mejor manejo del embarazo actual. (Estrada-Chiroque 2022).

El análisis de los resultados muestra que el Odds Ratio (OR) para el número de embarazos es de 1.0266, lo que indica que por cada embarazo adicional la probabilidad de que ocurra bajo peso al nacer (BPN) aumenta en un **2.6%** en comparación con mujeres con menos embarazos, que fueron tomadas como grupo de referencia. Este incremento, aunque moderado, es estadísticamente significativo y puede explicarse por el desgaste físico y las posibles complicaciones acumuladas a lo largo de los embarazos previos, lo que contribuye a un mayor riesgo de resultados perinatales adversos, entre ellos el BPN.

De igual forma, los embarazos múltiples evidencian un impacto importante sobre el BPN. En los partos dobles, el OR fue de 1.2234 (IC95%: 1.1542–1.2969; p=0.0297), lo que se traduce en un aumento cercano al **22%** del riesgo frente a partos únicos. En el caso de los partos triples, el OR alcanzó 1.0386 (IC95%: 1.0261–1.0492; p=0.0051), asociado a un incremento aproximado del **3,9%** en la probabilidad de BPN. Estos resultados confirman que la gestación múltiple constituye un factor de riesgo relevante, probablemente relacionado con la mayor demanda nutricional, la restricción del crecimiento intrauterino y la mayor frecuencia de partos prematuros.

#### 6.3.3.3 Nivel Educativo

El análisis de los resultados muestra que el nivel educativo de la madre se asocia con el riesgo de bajo peso al nacer (BPN). En comparación con las madres con educación básica primaria, aquellas sin ningún nivel educativo presentan un Odds Ratio (OR) de 1.0174, lo que indica que tienen un **1.7% más de probabilidad** de tener un hijo con BPN. De forma similar, las madres con educación preescolar muestran un OR de 1.0049, lo que refleja un riesgo apenas mayor, aunque estadísticamente significativo. Estos hallazgos sugieren que un nivel educativo inferior al de la básica primaria puede limitar el acceso a información y servicios de salud, dificultar la adopción de prácticas de cuidado prenatal y aumentar la vulnerabilidad a resultados adversos. En contraste, un mayor nivel educativo se asocia con mejores condiciones socioeconómicas y un mejor aprovechamiento de los servicios de salud, factores que favorecen el desarrollo fetal y reducen el riesgo de complicaciones (Estrada-Chiroque, 2022).

#### 6.3.3.4 Región Geográfica

El análisis de los resultados muestra que la región geográfica influye de manera significativa en la probabilidad de bajo peso al nacer (BPN). En comparación con la Amazonía, las madres del Distrito Capital Bogotá presentan un OR de 1.3007, lo que equivale a un **30% más de probabilidad** de BPN. De manera similar, en Boyacá el OR es de 1.0839 (**8.4% más de riesgo**) y en Cundinamarca alcanza 1.0705 (**7% más de riesgo**). Otros departamentos también muestran incrementos, aunque más moderados: Antioquia (**9.9%**), Valle del Cauca (**6.6%**), La Guajira (**5.3%**), Nariño (**4.2%**), Sucre (**3%**), Atlántico (**2.8%**), Quindío (**2.4%**), Risaralda (**2.4%**), Caldas (**1.9%**) y Chocó (**1.7%**).

Estas diferencias evidencian las disparidades regionales que existen en el acceso y la calidad de los servicios de salud, así como en las condiciones socioeconómicas de cada territorio. En general, las áreas urbanas ofrecen mejor acceso a servicios especializados y recursos médicos, lo que debería traducirse en mejores resultados perinatales. Sin embargo, la concentración de población y la presión sobre los sistemas de salud también pueden explicar los riesgos elevados en algunas capitales, como Bogotá. En contraste, en zonas rurales persisten barreras relacionadas con la distancia a los centros médicos, la falta de personal de salud y la escasez de infraestructura, factores que limitan la calidad del cuidado prenatal. Estos hallazgos ponen de relieve la necesidad de diseñar políticas públicas que reduzcan las brechas regionales y garanticen igualdad en el acceso a servicios de salud materno-infantil, contribuyendo así a mejorar los resultados perinatales en todo el país (Aguado Quintero, 2007).

#### 6.3.3.5 Variables Adicionales y su Impacto

##### 6.3.3.5.1 Ruralidad

Aunque la ruralidad no pudo ser incluida debido a la incompletitud de los datos, su impacto potencial en los resultados de salud es significativo. Las áreas rurales a menudo enfrentan desafíos como la distancia a centros de salud, menos profesionales de salud disponibles, y una menor infraestructura médica, lo que puede llevar a peores resultados perinatales. La literatura sugiere que las madres en áreas rurales tienen un mayor riesgo de bajo peso al nacer debido a estos factores. En el análisis bivariado, se observó que el 18.3% de las madres en áreas rurales tenían bebés con bajo peso al nacer, en comparación con el 12.5% en áreas urbanas.

#### 6.3.3.5.1 Condiciones Socioeconómicas

Las condiciones socioeconómicas, incluyendo el ingreso familiar y el empleo, son variables cruciales que influyen en la capacidad de las madres para acceder a cuidados de salud y nutrición adecuada. La falta de datos completos sobre estas variables limita el análisis, pero se reconoce que las familias con menores ingresos enfrentan mayores desafíos en términos de salud prenatal y neonatal. Estudios previos indican que el estatus socioeconómico bajo está fuertemente asociado con peores resultados perinatales debido a la falta de recursos y acceso a servicios de salud.

## 7. Conclusiones

El bajo peso al nacer es un problema significativo de salud pública en Colombia, ya que los neonatos con bajo peso al nacer tienen un mayor riesgo de morir durante el primer mes de vida comparado con aquellos con peso adecuado (UNICEF 2022). Además, estos neonatos son más propensos a desarrollar problemas de salud crónicos a lo largo de su vida, como discapacidades del desarrollo y enfermedades cardiovasculares (Velázquez Quintana 2004).

El cuidado de neonatos con bajo peso al nacer también representa una carga económica significativa para los sistemas de salud y las familias, debido a las hospitalizaciones prolongadas y tratamientos médicos costosos. Altas tasas de bajo peso al nacer son un indicador crítico de deficiencias en la atención prenatal y condiciones socioeconómicas desfavorables, reflejando la salud materna y neonatal en general. Finalmente, la detección temprana de bajo peso al nacer permite intervenciones oportunas, como programas de nutrición y cuidados neonatales especializados, que pueden mejorar significativamente los resultados de salud de los neonatos.

En este estudio, mediante la aplicación de análisis exploratorio de datos y modelos de *machine learning* a los datos proporcionados por el DANE para los años 2017 al 2021, se han identificado varios factores de riesgo significativos asociados con el bajo peso al nacer. Los resultados han destacado la importancia de factores como la calidad de la atención prenatal, el nivel socioeconómico, y las condiciones de salud maternas, como determinantes cruciales en la prevalencia de bajo peso al nacer.

La elección de utilizar modelos de *machine learning* en este estudio se alinea con los objetivos específicos de identificar y mitigar los riesgos de bajo peso al nacer en Colombia. Esta metodología permite una exploración detallada de las relaciones entre múltiples variables y la predicción de resultados con una alta precisión. Justificamos el uso de *machine learning* por su capacidad de manejar grandes conjuntos de datos y descubrir patrones que no son evidentes mediante análisis tradicionales. Este enfoque puede llenar las brechas actuales en la identificación temprana de riesgos y la implementación de estrategias preventivas.

Los modelos predictivos desarrollados en este estudio han mostrado una alta capacidad para anticipar casos de bajo peso al nacer. El modelo de regresión penalizada, con un *recall* de 0.83 y una precisión de 0.84, fue el que demostró mejor rendimiento, particularmente al considerar variables como el "Número de embarazos" (coeficiente: 0.02, p-valor: 5.7e-4) y la "Edad de la madre 45-49 años" (coeficiente: 0.13, p-valor: 0), que tuvieron un impacto significativo en la predicción.

La implementación de estas herramientas en la práctica clínica y de salud pública podría facilitar intervenciones preventivas más precisas y eficaces, optimizando la asignación de recursos y la planificación de programas de salud materna e infantil. Estos resultados subrayan la necesidad de fortalecer los servicios de atención prenatal y de abordar las disparidades socioeconómicas para reducir la incidencia de bajo peso al nacer en Colombia. Además, la integración de tecnologías de machine *learning* en la salud pública podría mejorar los resultados de salud neonatal mediante la detección temprana y las intervenciones preventivas.

Es pertinente un último comentario sobre consideraciones éticas. Estas son fundamentales al usar datos referentes a la salud personal. En este estudio se aseguró la privacidad de los datos y se procuró garantizar la equidad en el análisis y la implementación de las intervenciones. Se tomaron medidas para proteger la confidencialidad de la información y se siguió un enfoque justo y equitativo en la aplicación de los resultados del estudio.

## 8. Bibliografía


Organización Mundial de la Salud, Metas mundiales de nutrición 2025: Documento normativo sobre bajo peso al nacer, vol. 3. 2017, p. 8. doi: Licencia: CC BY-NC-SA 3.0 IGO. 5.

Shailaja, K., Seetharamulu, B., & Jabbar, M. A. (2018, March). Machine learning in healthcare: A review. In 2018 Second international conference on electronics, communication and aerospace technology (ICECA) (pp. 910-914). IEEE.

Faruk, A., Cahyono, E. S., Eliyati, N., & Arifieni, I. (2018). Prediction and classification of low-birth-weight data using machine learning techniques. Indonesian Journal of Science and Technology, 3(1), 18-28.

Salazar Blandon, D. A. (2023). Procesos gaussianos heterogéneos de múltiple salida para la predicción del bajo peso al nacer en Medellín.

Ahmadi, P., Alavimajd, H., Khodakarim, S., Tapak, L., Kariman, N., Amini, P., & Pazhuheian, F. (2017). Prediction of low birth weight using Random Forest: A comparison with Logistic Regression. *Archives of Advances in Biosciences*, *8*(3), 36–43.

Dharmaraj, A., Ghimire, A. & Chinnaiyan, S. (2024). Factors Associated with Low Birth Weight: Analysis from National Family Health Survey-4, India. *Indian J. Pediatr.* **91**, 421.

Jafarigol, E., & Trafalis, T. (2023). A review of machine learning techniques in Imbalanced Data and Future trends. *arXiv preprint arXiv:2310.07917*. Disponible en https://arxiv.org/abs/2310.07917.

Scikit-learn developers. (2024). *sklearn.linear_model.LogisticRegression*. Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Çorbacıoğlu, Ş. K., & Aksel, G. (2023). Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turkish journal of emergency medicine*, *23*(4), 195–198. https://doi.org/10.4103/tjem.tjem_182_23

Tissera, A. P., & Couriel, N. I302-Aprendizaje Automático y Aprendizaje Profundo.

van Wieringen, W. N. (2015). Lecture notes on ridge regression. arXiv preprint arXiv:1509.09169.

Monsreal, J. F., Cobos, M. D. R. T., Gómez, J. R. H., & Peraza, L. E. D. S. S. (2018). Factores de riesgo de bajo peso al nacer según el modelo de regresión logística múltiple. Estudio de cohorte retrospectiva en el municipio José María Morelos, Quintana Roo, México. Medwave, 18(1).

Panduro, B. C. (2022). Aplicación del algoritmo del bosque aleatorio a un modelo de clasificación de la anemia en niños peruanos. *Mediciego*, *28*(1), 3471.

Departamento Administrativo Nacional de Estadística – DANE. (2024). *Microdatos de nacimientos 2017–2021*. DANE. https://microdatos.dane.gov.co

Márquez-Beltrán, M. F., Vargas-Hernández, J. E., Quiroga-Villalobos, E. F., & Pinzón-Villate, G. Y. (2013). Análisis del bajo peso al nacer en Colombia 2005-2009. Revista de Salud Pública, 15, 626-637.

Yuan, Y., Du, J., Luo, J., Zhu, Y., Huang, Q., & Zhang, M. (2024). Discrimination of missing data types in metabolomics data based on particle swarm optimization algorithm and XGBoost model. *Scientific Reports*, *14*(1), 152.

Pérez, S. A., Calderón, M. M., Vargas, M. P., Soto, I. G., Gómez, Á., & Quijano, D. D. (2017). Relación entre factores sociodemográficos y el bajo peso al nacer en una clínica universitaria en Cundinamarca (Colombia). Revista Salud Uninorte, 33(2), 86-97.

Planchez, L. C., Garcia, Y. E. N., de Pedro, N. M., & Sánchez, M. L. (2021). Índice pronóstico de bajo peso al nacer. *Revista Médica Electrónica*, *43*(1).

Quiñones Montes, M. M. (2020). Caracterización del bajo peso al nacer a término en Antioquia, Trabajo de Grado de Pregrado, Universidad de Antioquia. Disponible en https://bibliotecadigital.udea.edu.co/bitstream/10495/16822/5/Qui%c3%b1onesMonica_2020_BajoPesoTermino.pdf.

Rivas Pérez, M. T. Factores Maternos Relacionados con Bajo Peso al Nacer en Colombia en el 2021. Trabajo de Grado de Maestría, Universidad Santo Tomás. Disponible en https://repository.usta.edu.co/handle/11634/51826?show=full.

Ticona, J. P. A., Gutiérrez, M. B. A., Rivas, D. R. Z., & Torres, N. M. C. (2017). Entendiendo la odds ratio. *Revista SCientifica*, *15*(1).

UNICEF. (2022). Levels & Trends in estimates developed by the United Nations inter-agency Group for Child Mortality Estimation 2022. *New York, NY: UNICEF*.

Bendezú, G., Espinoza, D., Bendezú-Quispe, G., Torres-Román, J. S., & Huamán-Gutiérrez, R. M. (2016). Características y riesgos de gestantes adolescentes. *Revista peruana de Ginecología y Obstetricia*, *62*(1), 13-18.

Pérez, B. H., Tejedor, J. G., Cepeda, P. M., & Gómez, A. A. (2011). La edad materna como factor de riesgo obstétrico. Resultados perinatales en gestantes de edad avanzada. *Progresos de Obstetricia y ginecología*, *54*(11), 575-580.

Aguado Quintero, Luis Fernando, Girón Cruz, Luis Eduardo, Osorio Mejía, Ana María, Tovar Cuevas, Luis Miguel, & Ahumada Castro, Jaime Rodrigo. (2007). Determinantes del uso de los servicios de salud materna en el Litoral Pacífico Colombiano. Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud, 5(1), 233-281. Retrieved August 23, 2024, from http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1692-715X2007000100008&lng=en&tlng=es.

Estrada-Chiroque, L. M.., Orostegui-Arenas, M., Burgos-Guanilo , M. del P.., & Amau-Chiroque, J. M. (2022). Características clínicas y resultado materno perinatal en mujeres con diagnóstico confirmado por COVID-19 en un hospital de Perú. Estudio de cohorte retrospectivo. *Revista Colombiana De Obstetricia Y Ginecología*, *73*(1), 28–38. https://doi.org/10.18597/rcog.3776

Velázquez Quintana, N. I., Masud Yunes Zárraga, J. L., & Ávila Reyes, R. (2004). Recién nacidos con bajo peso; causas, problemas y perspectivas a futuro. Boletín médico del Hospital Infantil de México, 61(1), 73-86.