FACULTAD DE CIENCIAS ECONÓMICAS

UNIVERSIDAD NACIONAL DE MISIONES

Predicción del bajo peso al nacer en Argentina: comparación de enfoques predictivos y análisis de determinantes socioeconómicos

Tesis de Grado en Economía — Universidad Nacional de Misiones

Alumno: Dilger, Johan Axel

Director: Mgter. Staudt, Agustín

2025

Resumen

El bajo peso al nacer (BPN) es uno de los principales factores de riesgo de mortalidad infantil y repercute en resultados a lo largo del ciclo de la vida de las personas como el rendimiento educativo, el empleo y la longevidad. El mismo puede resultar en costos sustanciales para la sociedad y constituye un mecanismo que compromete la productividad y el desarrollo económico de largo plazo. El presente trabajo se propone predecir el bajo peso al nacer en Argentina utilizando regresión logística y métodos de machine learning lasso y random forest. Los resultados de dichos modelos se comparan para encontrar el estimador de mayor consistencia y poder predictivo. Adicionalmente, se indaga la influencia que las características socioeconómicas de las madres y del hogar tienen sobre dicha probabilidad. Se utilizan datos de la Encuesta Nacional de Niñas, Niños y Adolescentes (MICS) 2019-2020 (UNICEF, 2021). Los resultados indican que la regresión logística mejora significativamente la predicción del bajo peso al nacer comparado con los demás modelos, identificando como determinantes clave la edad de la madre, la región, su estado civil, entre otros factores. Estos resultados permitirían implementar políticas públicas focalizadas para prevenir la prevalencia del bajo peso al nacer.

Palabras clave: Peso al nacer; Machine learning; Predicción; Determinantes socioeconómicos.

1. Introducción

La evidencia muestra que el peso al nacer se relaciona con una variedad de resultados a lo largo del ciclo de la vida de la persona, actuando como un predictor socioeconómico significativo de la educación, el empleo, la salud y la longevidad de las personas (Aizer y Currie, 2014; Gillion, 2017; Akbulut-Yuksel et al., 2020; Conti et al., 2020; Cuestas et al., 2021). A su vez, es considerado el indicador más utilizado de la salud neonatal (Conti et al., 2020). Un canal por el cual las desigualdades en capital humano se transmiten de una generación a otra es por la salud al nacer, que a su vez está condicionada por las circunstancias del entorno previo a la gestación, durante ese período y en el momento del nacimiento (Berniell y De La Mata, 2022). De esta forma, un bajo peso al nacer afecta negativamente el bienestar futuro del recién nacido, siendo este un proceso de difícil reversión (Cruces et al., 2012).

El bajo peso al nacer (BPN) es definido por la Organización Mundial de la Salud [OMS] como un peso al momento del nacimiento inferior o igual a 2.5 kg. Se estima que entre un 15% y 20% de los niños nacidos en todo el mundo presentan BPN, lo que implica más de 20 millones de nacimientos anuales con bajo peso (OMS, 2017). La Asamblea Mundial de la Salud [AMS] propuso reducir el bajo peso al nacer un 30% entre 2012 y 2030, sin embargo, en la mayoría de las regiones del mundo esta reducción aún no pudo lograrse. En Asia y América Latina inclusive se registra un retroceso en este objetivo (UNICEF, 2023). En 2020 la prevalencia del bajo peso al nacer fue mayor en Asia Meridional (24.9%) mientras que en América Latina y el Caribe se estimó en 9.6% (UNICEF, 2023). En Argentina, el 7.9% de los nacidos vivos presentaron bajo peso al nacer según el último dato disponible de 2022 (Ministerio de Salud, 2024). Los casos con esta condición a nivel nacional se mantuvieron en un promedio del 7.2% en el período 2006-2022, dando cuenta que, si bien la cantidad de nacimientos se redujo un 23,9% en este período, la problemática se mantuvo constante.

La literatura que investiga el bajo peso al nacer y su relación con factores socioeconómicos es amplia. Por su parte, se encuentra que las características socioeconómicas de la madre tienen un rol importante en la probabilidad del bajo peso al nacer (Anderson, 2022; Szabó y Boros, 2023). En este sentido, Grytten et al. (2014) encuentra que el incremento en el nivel educativo de las madres disminuye la probabilidad de bajo peso al nacer con datos de Noruega. Hidalgo-Lopezosa et al. (2019) establece que para España la edad materna, el nivel educativo y el estado civil se asocian al bajo peso al nacer, mientras que Rosenwaike (1971) haya una relación entre el nivel educativo, la raza, el nivel socioeconómico y el bajo peso al nacer para Estados Unidos. En el caso de países en desarrollo, Torres-Arreola et al. (2005) encuentra que un nivel socioeconómico bajo es el factor de riesgo más relevante para el bajo

peso al nacer con datos de la Ciudad de México, mientras que Mahumud et al. (2017) menciona la edad materna, los cuidados prenatales, la educación y un estatus económico bajo entre los principales determinantes. En Argentina, Ratowiecki et al. (2018) encuentra un impacto de la desigualdad socioeconómica sobre la prevalencia de bajo peso, especialmente en nacimientos ocurridos en hospitales públicos y madres en edades extremas. De esta forma, la literatura encuentra evidencia de que las condiciones de la madre, del hogar y del entorno afectan la salud del recién nacido, cuyo peso es un predictor importante de sus resultados futuros. Por lo que la prevalencia del BPN tiene consecuencias a largo plazo, afectando no sólo el bienestar humano, sino también la distribución del ingreso y del capital humano (Aizer y Currie, 2014).

El BPN se presenta como uno de los principales factores de riesgos de mortalidad infantil, actuando como el mediador biológico central de la relación entre la clase social y las condiciones económicas con la mortalidad (Paneth, 1995). Mientras que aquellos que sobreviven a este fenómeno presentan dificultades a lo largo de la vida como menor rendimiento educativo, peor estado de salud y reducción en el empleo e ingresos (Almond et al., 2005). Esto último limita los retornos de las inversiones en capital humano realizadas en ellos (Behrman y Rosenzweig, 2004) ya que este fenómeno puede resultar en costos sustanciales para el sector sanitario, los servicios sociales, sobre las familias y la sociedad en general (Petrou et al., 2001; Almond et al., 2005). Por estas razones, el BPN presenta no solo elevados costos económicos (Alderman y Behrman, 2006), sino que constituye un mecanismo que compromete la productividad y el desarrollo económico de largo plazo.

En Argentina poco se ha estudiado a la problemática desde un enfoque predictivo, siendo que predecir el BPN permitiría identificar los factores más influyentes y anticiparse a los casos de mayor probabilidad de ocurrencia. Dicho análisis contribuiría tanto a la literatura como a la generación de políticas públicas, posibilitando el diseño e implementación de medidas que permitan combatir esta problemática de la manera más efectiva. En este sentido, surge la pregunta de si sería posible predecir los casos de bajo peso al nacer a través de variables socioeconómicas y así establecer la influencia que estos factores tienen sobre la probabilidad de BPN.

En este contexto, el presente trabajo se propone predecir los casos de bajo peso al nacer en Argentina, a partir de las condiciones socioeconómicas de la madre y del hogar. Adicionalmente, se indaga la influencia que estos factores tienen sobre la probabilidad de BPN. Para ello, se busca el modelo de mayor consistencia y poder predictivo, y se indaga acerca de la influencia que las características individuales y del hogar tienen sobre dicha probabilidad. Para llevar a cabo el análisis predictivo se utilizan dos métodos. En primer lugar,

se realiza la estimación a partir del modelo de regresión logística, considerado en este trabajo como una técnica tradicional de predicción, y se agregan al análisis dos modelos de clasificación de aprendizaje automático: el estimador lasso y el modelo random forest. Para ello, se utilizan los datos disponibles a través de la encuesta UNICEF MICS para Argentina, que incluye un cuestionario donde se relevan características de las madres, los hogares y los niños.

Los resultados encontrados indican una diferencia estadísticamente significativa del modelo de regresión logística a la hora de predecir el bajo peso al nacer respecto a los demás modelos, en línea con algunos hallazgos encontrados anteriormente (Borson et al., 2020; Islam Pollob et al., 2022; Khan et al., 2022; Arayeshgari et al., 2023; Patterson et al., 2023; Mathew y Thinakaran, 2025). El modelo tradicional (logit) aumentó la performance predictiva en casi 14 puntos porcentuales. A su vez, haciendo uso de una técnica de selección de variables, la edad de la madre se encuentra entre las variables más influyentes de la predicción. En el caso de logit es la variable más importante, mientras que random forest y en la regresión lasso se ubican en el segundo y sexto lugar, respectivamente. En la mayoría de los modelos, la misma está por delante o detrás de los deciles de riqueza, por lo que considerar a la edad es tan importante como considerar el estatus económico del hogar.

Además, se encuentran otros factores relevantes que coinciden con los hallazgos de la literatura de bajo peso al nacer. La región de pertenencia, el estado civil de la madre, la descendencia indígena y la educación tanto de la madre como del jefe/a de hogar se relacionan con la probabilidad de bajo peso al nacer, siendo estos resultados compatibles con lo encontrado por investigaciones previas.

A través del análisis predictivo el trabajo contribuye a la literatura de bajo peso al nacer en distintas direcciones. En primer lugar, al ser un estudio con un enfoque poco explorado aún de análisis predictivo del bajo peso al nacer utilizando variables socioeconómicas con datos de Argentina. En segundo lugar, al brindar un aporte metodológico al comparar metodologías tradicionales con otras más avanzadas. A su vez, utiliza datos MICS, escasamente explorados a nivel general para Argentina. Por su parte, los hallazgos de este trabajo serían de utilidad para la formulación de políticas públicas, por el hecho de que los resultados permitirían identificar los potenciales casos de bajo peso al nacer, prestando atención a las características principales de las madres y los hogares, contribuyendo así a la mejora del diseño e implementación de políticas públicas focalizadas.

El trabajo se organiza de la siguiente manera: en primer lugar, se presenta la revisión de la literatura de bajo peso al nacer relacionada con factores socioeconómicos, como también se mencionan los trabajos que utilizan enfoques predictivos para predecir el bajo peso al nacer

(sección 2). En la sección 3 se describe la fuente de datos, seguidamente en la sección 4 se presenta la metodología adoptada y luego se analizan exploratoriamente los datos de entrenamiento (sección 5). En la sección 6 se muestran los resultados de los modelos de predicción implementados. Finalmente, en la sección 7 se presentan las conclusiones del trabajo.

2. Revisión de la literatura

2.1 Factores socioeconómicos y el bajo peso al nacer

Los determinantes y factores asociados al bajo peso al nacer pueden clasificarse en cuatro grupos: factores biológicos, factores demográficos, factores socioeconómicos y factores nutricionales (Khan et al., 2020), por lo que un gran número de variables socioeconómicas se asocian al desarrollo físico de los niños. En este sentido, se estima que el 60% de la variación del peso al nacer puede ser atribuida al entorno en el que el feto crece (Belitsky et al., 1992 como se cita en Rendón y Apaza, 2009).

La literatura que investiga el bajo peso al nacer y su relación con las condiciones socioeconómicas es amplia. Entre ellos, los factores maternos se encuentran como los determinantes más importantes (Zaveri et al., 2020). En este sentido, se suele citar a la edad de la madre (Silva, 2012; Restrepo-Méndez et al., 2015; Mahumud et al., 2017; Ratowiecki et al., 2018; Hidalgo-Lopezosa et al., 2019), su nivel educativo (Som et al., 2004; Juárez y Revuelta Eugercios, 2013; Grytten et al., 2014; Ratowiecki et al., 2018; Hidalgo-Lopezosa et al., 2019; Mahmoodi et al., 2013; Mahumud et al., 2017; Zaveri et al., 2020; Khan et al., 2020), al estado civil (Phung et al., 2003; Hidalgo-Lopezosa et al., 2019) y los cuidados prenatales (Anjum et al., 2011; Mahumud et al., 2017; Falcão et al., 2020) como los más influyentes.

Una edad materna temprana implica un mayor riesgo de tener recién nacidos con BPN. Algunas explicaciones de este hecho lo relacionan con factores biológicos, como la competencia por nutrientes entre una persona embarazada que aún está en etapa de crecimiento y el feto. Por otro lado, la preocupación reciente por los resultados adversos de los nacimientos también se ha desplazado hacia las madres de mayor edad, a consecuencia de los nacimientos de madres en edad adulta que han aumentado en los últimos años (Restrepo-Méndez et al., 2015). Una edad materna avanzada también es asociada con una disminución en el potencial de crecimiento fetal, lo que posiblemente refleja el envejecimiento biológico o efectos acumulativos de enfermedades previas o no detectadas (Aras, 2013; Restrepo-Méndez et al., 2015).

Por su parte, el nivel educativo refleja la preparación de las personas para una paternidad o maternidad responsable (Hendricks, 1967). La educación es uno de los factores que influye en la toma de decisiones de las personas, a la par que las madres más educadas pueden tener mejores trabajos y por ende una mejor condición de vida (Mahmoodi et al., 2013). La mayor escolaridad influye en el conocimiento de la mujer sobre la necesidad de proporcionarse cuidados prenatales y alimentación adecuada. Por lo que se puede intuir que, a mayor escolaridad, mejor ingreso económico y menor probabilidad de BPN (Rendón y Apaza, 2009).

A su vez, las madres solteras presentan una mayor probabilidad de tener hijos con bajo peso al nacer en comparación con las madres casadas (Padilla y Reichman, 2001). Las primeras pueden presentar mayor probabilidad de enfrentar condiciones socioeconómicas desfavorables y el estrés puede afectar el embarazo de manera directa a través de la liberación de sustancias químicas naturales como el cortisol (Masho et al., 2010). Con respecto a los cuidados prenatales, la asociación entre los mismos y el BPN ha sido ampliamente documentada en la literatura científica. Si bien esta relación tuvo dificultades iniciales de interpretación desde la perspectiva médica, ya que las visitas prenatales estándar parecían tener una influencia limitada en la reducción del BPN (Alexander y Korenbrot, 1995), investigaciones posteriores han demostrado que la atención prenatal efectivamente es beneficiosa (Supadmi et al., 2020).

La atención prenatal constituye una de las herramientas más valiosas para los profesionales de la salud en la detección oportuna de factores de riesgo modificables que pueden impactar en el desarrollo fetal y el peso final al nacer. Una atención prenatal completa, iniciada tempranamente y con un número adecuado de visitas, resulta fundamental para reducir la incidencia del BPN y prevenir complicaciones asociadas (Pinzón-Rondón et al., 2015). Evidencia adicional sugiere que los recién nacidos de madres con seguro médico presentan mejores indicadores de salud al nacimiento en comparación con aquellos de madres sin seguro (Kumar y Gonzalez, 2018).

En otro orden, se afirma que la etnia es relevante en los casos de bajo peso al nacer (Johnelle Sparks, 2009). Sin embargo, no se ha logrado establecer explicaciones socioeconómicas y conductuales para las diferencias étnicas en el peso al nacer. Esta aparente falta de evidencia ha llevado a algunos a sugerir que los pesos más bajos al nacer en ciertos grupos étnicos son de alguna manera normales (Kelly et al., 2009).

El bajo nivel socioeconómico materno se asocia con un mayor riesgo de BPN (Martinson y Reichman, 2016; Mahumud et al., 2017; Khan et al., 2020; Zaveri et al., 2020; Mishra et al.,

2021; Spencer y Logan, 2002; Finch, 2003; Faulks et al., 2023), creando un ciclo intergeneracional de disparidades en salud y riqueza (Anderson, 2022). Esto se debe a que las mujeres con desventaja económica tienen mayor probabilidad de una ingesta alimentaria y viviendas inadecuadas, y menor acceso a atención médica y medicamentos (Som et al., 2004; Elshibly y Schmalisch, 2008). Un mejor estatus económico se relaciona con la disponibilidad de alimentos y una adecuada ingesta durante el embarazo, reduciendo la incidencia del BPN (Supadmi et al., 2020; Wulandari et al., 2023). La desventaja socioeconómica se vincula a bajo nivel educativo (Lin y Liu, 2009; Burdette et al., 2012), raza o etnia (Rauh et al., 2001; Hoggatt et al., 2012) y ser madre soltera (Burdette et al., 2012; Aizer y Currie, 2014), lo que ha sido foco de estudio en cómo afecta la salud del recién nacido (Aizer y Currie, 2014; Spencer y Logan, 2002; Zaveri et al., 2020; Phung et al., 2003; Altenhöner et al., 2016).

2.2 Machine learning y el bajo peso al nacer

A la hora de predecir el BPN, los antecedentes muestran el uso de herramientas estadísticas avanzadas de machine learning, así como de métodos más tradicionales como la regresión logística. Esta última es además una de las técnicas de predicción más utilizadas en el campo de la medicina (Shipe et al., 2019; Nusinovici et al, 2020). Entre los antecedentes, Islam Pollob et al. (2022) hace uso de la misma, junto con otra más avanzada, en un ejercicio de predicción del BPN con datos de Bangladesh. A su vez, Senthilkumar y Paulraj (2015) hacen lo propio con datos de la India. Por lo general, al utilizar la técnica tradicional, la comparan con técnicas más avanzada de machine learning. En este sentido, los resultados de desempeño predictivo de los modelos tradicionales en relación a los modelos de machine learning no son del todo concluyentes.

Algunos de los trabajos que hacen uso de estas técnicas de predicción más avanzadas utilizan datos de Afganistán (Zahirzada y Lavangnananda, 2021), Bangladesh (Borson et al., 2020; Islam Pollob et al., 2022; Mansur et al., 2024; Reza y Salma, 2024), Etiopía (Bekele, 2022), India (Hussain y Borah, 2020; Priya et al., 2024), Indonesia (Faruk et al., 2018; Kurniawati et al., 2022) e Irán (Ahmadi et al., 2017; Arayeshgari et al., 2023; Ranjbar et al., 2023). Entre las técnicas de aprendizaje automático empleadas en la literatura relacionada con la temática de este trabajo se destacan random forest, árbol de decisión y extreme gradient boosting (XGBoost, por sus siglas en inglés). Además, trabajos que buscan estudiar la selección de variables y su relevancia utilizan el método lasso (Kurniawati et al., 2022). El porcentaje de accuracy (precisión) de estas técnicas fue superior al 60% en una serie de estudios revisados (Mane et al., 2024). En general, los resultados de los antecedentes mostraron que el uso de la técnica random forest fue superior en la predicción del bajo peso

al nacer en comparación a la regresión logística (Ahmadi et al., 2017; Faruk et al., 2018) y contra otras técnicas empleadas (Hussain y Borah, 2020; Zahirzada y Lavangnananda, 2021; Bekele, 2022). No obstante, en otros estudios la regresión logística presentó mejores resultados para predecir el bajo peso al nacer, en relación a diferentes estimadores de machine learning (Islam Pollob et al., 2022; Arayeshgari et al., 2023).

3. Datos

Este trabajo utiliza datos de la Encuesta Nacional de Niñas, Niños y Adolescentes (MICS, por sus siglas en inglés) llevada a cabo para Argentina en 2019-2020 por el Fondo de las Naciones Unidas para la Infancia (UNICEF, por sus siglas en inglés). La encuesta consistió en entrevistas con los miembros de los hogares sobre una variedad de temas, centrándose principalmente en aquellos asuntos que afectan directamente la vida de los niños y las mujeres. Esta encuesta tiene representatividad a nacional y regional ya que proporciona estimaciones en áreas urbanas y para las regiones AMBA, Pampeana, Cuyo, NEA, NOA y Patagonia (UNICEF, 2021).

La base de datos consta de cinco cuestionarios con información general sobre el hogar, los respectivos miembros del mismo, uno específico para las mujeres de 15 a 49 años de edad, uno para niños menores de 5 años y otro para niños entre 5 y 17 años. En total la encuesta MICS completó el cuestionario a 12202 mujeres de entre 15-49 años de edad, 6157 niños hasta 5 años de edad y 6536 niños en el rango de 5 a 17 años. Para este trabajo se utilizaron las bases de hogares, los miembros y el cuestionario a mujeres.

Entre las ventajas que ofrece la utilización de esta base de datos se destaca la amplitud de variables relevadas y el alcance a nivel nacional; por el contrario, entre sus limitaciones se encuentra la falta de periodicidad, ya que la primera ronda de esta encuesta se llevó a cabo en 2011-2012 y la siguiente en 2019-2020. A su vez, la diferencia entre ambos períodos es que no permite hacer un seguimiento completo o analizar la evolución de las variables, debido a que el cuestionario se amplió en el último relevamiento. En este sentido, el peso al nacer fue incluido únicamente en la ronda 2019-2020.

Un total de 2134 mujeres afirmaron haber tenido algún nacimiento en los últimos dos años de realizada la encuesta.[1] De ellas, 2110 respondieron que el niño fue pesado al nacer. La variable dependiente binaria de interés en este trabajo fue construida a partir de la variable que mide el peso al nacer, clasificando los casos de bajo peso al nacer a aquellos menores

[1] La forma de recolección de la variable dependiente tiene limitaciones que serán mencionadas más adelante.

o iguales a 2500 gramos. Las variables socioeconómicas seleccionadas para el análisis pueden verse en el Anexo.

4. Metodología

Frente a una variable dependiente de tipo binaria, se está en presencia de un problema de predicción de respuesta cualitativa, conocido como clasificación. En este caso, la variable dependiente binaria es el bajo peso al nacer, donde la misma toma el valor 1 en caso de bajo peso al nacer (instancia positiva) y 0 en caso contrario (instancia negativa). Para este trabajo se consideró a los casos <= 2500 como bajo peso.

De acuerdo a Pesantez-Narvaez et al. (2019), teniendo datos de N individuos y P covariables, existe la variable de respuesta binaria Y_i con $i = 1,\dots,n$ que toma valores 0 y 1. A su vez, el conjunto de covariables se denota como X_{ip} con $p = 1,\dots,P$ y la función de probabilidad condicional de $Y_i = t(t = 0{,}1)$ dado $X_i(X_{i1},\dots,X_{iP})$ se expresa como $\pi_t(X_i)$. Esto último puede ser expresado como $Prob(Y_i = t) = \pi_t(X_i)$ tal que $E(Y_i) = Prob(Y_i = 1) = \pi_1(X_i)$.

4.1 Regresión logística

En primer lugar, se utilizará el modelo de regresión logística, definido como uno tradicional, por su predominancia a la hora de predecir diferentes outcomes cualitativos, fuera del campo de aprendizaje automático. Este modelo es uno de los más utilizados por la literatura que analiza los determinantes del bajo peso al nacer. El principal objetivo del mismo es la clasificación (Cokluk, 2010) y se caracteriza por el hecho que la variable de respuesta es una variable binaria en vez de una continua.

La regresión logística utiliza la función logit como un enlace canónico, es decir, el logaritmo del ratio de las funciones de probabilidad $\pi_t(x_i)$ es una función lineal de X, esto es:

$$log\frac{\pi_1(X_i)}{\pi_0(X_i)} = log\frac{Prob(Y_i = 1)}{Prob(Y_i = 0)} = \beta_0 + \sum_{p=1}^{p} X_{ip}\beta_p \quad (1)$$

con $\beta_0, \beta_1, \dots, \beta_p$ como parámetros del modelo, $Prob(Y_i = 1)$ como la probabilidad de observar el evento en la variable dependiente binaria y $Prob(Y_i = 0)$ como la probabilidad de no observarlo.

Reescribiendo $\pi_1(X_i) = \pi$ y $\pi_0(X_i) = 1 - \pi$, y aplicando exponenciales a ambos términos, se puede expresar la ecuación (1) cuando $Prob(Y_i = 1)$ como:

$$\frac{\pi}{1-\pi} = exp(\beta_0 + \sum_{p=1}^{p} X_{ip}\beta_p) \quad (2)$$

A partir de (2) se puede expresar:

$$\pi = \frac{exp(\beta_0 + \sum_{p=1}^{p} X_{ip}\beta_p)}{1 + exp(\beta_0 + \sum_{p=1}^{p} X_{ip}\beta_p)}$$

En este modelo, el incremento de una unidad en x_i implica el cambio en una unidad en β_i de la probabilidad del suceso del evento. Equivalentemente ello puede interpretarse como el cambio en una unidad de x_i multiplica la probabilidad del suceso del evento por exp(β_i). Los coeficientes β_i se estiman en este modelo a través de la función de máxima verosimilitud, al tener mejores propiedades estadísticas que la estimación mediante mínimos cuadrados ordinarios (James et al., 2013; Pesantez-Narvaez et al., 2019).

$$\ell(\theta) = \sum_{i=1}^{N}\{y_i(\beta_0 + \sum_{p=1}^{p} X_{ip}\beta_p - \log(1+\exp(\beta_0 + \sum_{p=1}^{p} X_{ip}\beta_p))\} \quad (3)$$

4.1.1 Método de selección hacia adelante

Las principales ventajas en el uso del modelo tradicional logit en estudios de determinantes del BPN puede ser mejorado desde el enfoque predictivo, mediante una selección de variables guiada por los datos. El método de selección hacia adelante elige las variables que le agregan mayor ganancia adicional al modelo en la predicción (James et al., 2013). El proceso de selección inicia desde una especificación sin predictores y agrega las variables, una a la vez, hasta llegar al modelo con el set completo de variables. El criterio de selección se realiza en base al AUC-ROC sobre el grupo de validación, es decir, se ajusta el modelo con cada una de los predictores en el grupo de entrenamiento, pero se selecciona aquel que genere el mayor AUC-ROC adicional a la hora de predecir nuevos datos (Staudt, 2022). A su vez, el procedimiento usualmente utilizado ajusta y selecciona sobre el grupo de entrenamiento y la selección del modelo final se realiza evaluando esa especificación óptima en el grupo de validación (James et al., 2013). Antes de ejecutar el algoritmo de selección se aplica la técnica de one-hot-encoding, la cual transforma cada categoría de una variable string en una variable binaria con un 1 indicando si la observación tiene la categoría en cuestión (Staudt, 2022). De esta manera, las variables categóricas fueron transformadas en variables de tipo dummy.

4.2 Regresión lasso

En el método tradicional, cuando la cantidad de observaciones (n) no es mucho más grande que la cantidad de variables independientes (p), se podría generar un escenario de mucha variabilidad en el ajuste, resultando en un sobreajuste del estimador y por lo tanto en malas predicciones realizadas sobre las observaciones de testeo. A su vez, algunas variables utilizadas en el modelo de regresión podrían no estar asociadas con la respuesta de la variable dependiente, por lo que incluir tales predictores irrelevantes podría conducir a una complejidad innecesaria en el modelo resultante (Szretter Noste, 2019). Al eliminar estas variables (ajustando a cero las estimaciones del coeficiente correspondiente), utilizando de base la estimación logística, se puede obtener un modelo de mayor consistencia, sin perder de vista la interpretabilidad de los resultados.

El método lasso (*Least Absolute Shrinkage and Selection Operator*) es un método de regularización que impone una restricción o penalización sobre los coeficientes de regresión, para reducir la varianza del estimador (a costa de un aumento en su sesgo). Ante contextos de mucha variabilidad, la reducción de la varianza puede compensar el aumento del sesgo, mejorando así los resultados de la predicción ante nuevas observaciones. A su vez, lasso utiliza una regularización que contrae los coeficientes menos relevantes hacia exactamente cero, lo cual genera indirectamente un ranking de atributos más importantes de la predicción, a partir del valor absoluto de los coeficientes que permanecen en el resultado final de la estimación (Tibshirani, 1996). Para ello, el método agrega al estimador de máxima verosimilitud de logit un hiper parámetro de regularización que contrae los coeficientes hacia cero, manteniendo un modelo con estructura lineal (Lee et al., 2022). La ecuación lasso de regresión logística penalizada a maximizar será, a partir de la Ecuación 3:

$$max_{\theta}\left\{\sum_{i=1}^{N}\left[y_i(\beta_0+\sum_{p=1}^{p}x_{ip}\beta_p)-log(1+exp(\beta_0+\sum_{p=1}^{p}x_{ip}\beta_p))\right]-\lambda\sum_{j=1}^{p}|\beta_j|\right\} \quad (4)$$

El término λ debe ser determinado previamente, ya que no se obtiene a través de la optimización. Cuando $\lambda = 0$, el resultado es el mismo que sin penalización, es decir, resultará aproximadamente en el modelo logit. En el otro extremo, cuando $\lambda \to \infty$ los coeficientes estimados serán forzados a cero. Por lo que un valor de λ lo suficientemente grande producirá selección de variables.

En este caso, para obtener el parámetro de regularización (λ) se realizó un escalamiento de los datos utilizados, para normalizar las variables predictoras y otorgar un peso balanceado de las clases (siendo que las clases se encuentran desbalanceadas porque los casos de bajo peso al nacer son menores). Seguidamente se implementó la búsqueda de hiper parámetros para encontrar el mejor valor de λ, utilizando una validación cruzada de 5 pliegues y empleando el AUC-ROC como métrica de evaluación. De esta manera se buscó seleccionar

el parámetro de regularización λ que maximice la capacidad predictiva del modelo, medida por el AUC-ROC.

4.3 Random forest

Random forest (Breiman, 2001) es una técnica de machine learning basada en la construcción y combinación de múltiples árboles de decisión, utilizada para la resolución de problemas de regresión o de clasificación de datos. Un árbol de decisión destinado a clasificación genera una respuesta cualitativa que trata de predecir a qué categoría pertenece cada observación, asignándole a aquella con la cual tiene más propiedades en común en función de las observaciones de entrenamiento (Sardaña, 2022).[2]

La idea detrás de este método es la de generar un conjunto de árboles independientes, los cuales de manera conjunta presentan mejor performance que de manera individual. De esta forma, las decisiones de clasificación son decididas por el "voto" de la mayoría de los árboles construidos individualmente (Chakraborty y Joseph, 2017). Al emplear este mecanismo son propensos a presentar mayor aleatoriedad, ya que cada árbol es entrenado de manera independiente y al final se agregan los resultados de cada uno de ellos (Sardaña, 2022). A su vez, a la agregación aplicada el algoritmo hace uso de la técnica de bootstrap: se toman muestras repetidas del mismo set de entrenamiento.

El modelo puede ser expresado de la siguiente manera (Hanafy y Ming, 2021):

$$g(x) = f0(x) + f1(x) + f2(x) + \ldots + fn(x)$$

donde g es el modelo de random forest final, es decir, la suma de todos los árboles de decisión, mientras que cada $f(x)$ corresponde a cada uno de los árboles de decisión individuales.

En los casos de clasificación, la técnica de bagging agrega los resultados eligiendo la clasificación de la mayoría de los múltiples árboles entrenados en random forest. Formalmente:

[2] Los árboles de decisión presentan la desventaja que son propensos al sobreajuste del conjunto de datos de entrenamiento y debido a ello no clasifican bien los datos *no vistos.* Este defecto implica que, si se dividen los datos de entrenamiento en dos partes al azar y se ajusta un árbol de decisión a ambas mitades, los resultados que se obtienen son muy dispares, resultando en una varianza muy alta (James et al., 2013). Random forest emplea un método de agregación o ensamble de múltiples árboles para reducir esta varianza, llamado *bagging* (Breiman, 2001).

Sea $\hat{C}_b(x)$ la predicción de clase del b-ésimo árbol de random forest. Entonces $\hat{C}_{rf}^{B}(x) = voto\ mayoritario\ \{\hat{C}_b(x)\}_1^B$.[3]

Lo que diferencia a random forest de los árboles de decisión es que la técnica *bagging* en el primero, independiza la relación entre los árboles. Esta no-correlación se debe a que los múltiples árboles de decisión, generados en random forest, seleccionan al azar un número limitado de variables para cada nodo del árbol creado y un subconjunto de observaciones. Por lo que, mientras los árboles de decisión observan todas las covariables o predictores en cada nodo, random forest sólo observa unas pocas elegidas al azar. En general, el número seleccionado de variables predictoras observadas en cada nodo es la raíz cuadrada del número máximo de variables que contiene el conjunto de datos original. Así, el algoritmo al no poder utilizar todas las variables elimina la correlación entre los árboles de decisión, y al tener en cuenta toda la variabilidad potencial de los datos, tanto en las variables observadas, como en las observaciones, se puede reducir el riesgo de sobreajuste, sesgo y varianza general, lo que da lugar a predicciones más precisas que la estimación de un sólo árbol de clasificación (Sardaña, 2022).

Entre las principales ventajas de esta técnica se destacan su capacidad para manejar la no linealidad y la selección de variables. Se trata de un algoritmo capaz de elegir de manera independiente qué variables ajustan mejor, en base a la importancia o ganancia en información que realizan (Varian, 2014; Chaluh, 2023). Por el contrario, uno de sus defectos radica en que actúa como una caja negra, en el sentido que no ofrece explicaciones simples de las relaciones entre los datos (Varian, 2014).

Existe una variedad de parámetros que se deben elegir previo a la ejecución del algoritmo. En este caso, se calibraron los siguientes, buscando maximizar el AUC-ROC:

- *n_estimators*: Número de árboles en el bosque
- *max_depth*: Profundidad máxima de cada árbol
- *min_samples_split*: Mínimo de muestras para dividir un nodo
- *min_samples_leaf*: Mínimo de muestras en cada hoja
- *max_features*: Número de features a considerar en cada división.

4.4 Métrica de evaluación: AUC-ROC

Para evaluar el desempeño predictivo de los modelos de machine learning en un enfoque de clasificación se emplea el área bajo la curva ROC (AUC, por su siglas en inglés). AUC se

[3] El desarrollo puede verse en Hastie et al. (2009).

emplea habitualmente para valorar la bondad de un modelo de clasificación binario, teniendo en cuenta diferentes umbrales de decisión para considerar la pertenencia de una observación a un grupo u otro (Martínez et al., 2021).

Esta métrica es utilizada en la literatura de machine learning y es reconocida por su robustez al momento de evaluar el desempeño predictivo en problemas con clases desbalanceadas[4]. La misma toma valores en el intervalo [0, 1], donde mayor AUC implica mejor desempeño predictivo, ya que este mide la capacidad del modelo de separar correctamente las clases del output (un valor igual a 1 representa una separación perfecta). A su vez, un valor igual a 0.5 indica que el modelo predice acorde a lo que predeciría un modelo totalmente aleatorio, mientras que un valor menor a 0.5 sugiere un desempeño peor que el azar (Soules, 2020).

La curva ROC muestra el desempeño de un modelo de clasificación para todos los umbrales o puntos de cortes, donde grafica la Tasa de Verdaderos Positivos (TPR) o sensibilidad y la Tasa de Falsos Positivos (FPR) o 1 − especificidad. La TPR (o recall) se define como

$$TPR = \frac{TP}{TP + FN}$$

A su vez, FPR viene dado por

$$FPR = \frac{FP}{FP + TN}$$

Por su parte, a la hora de comparar el desempeño de cada estimador se utiliza el test de DeLong et al. (1988) que determina si las diferencias de predicción existentes entre modelos son estadísticamente significativas.

4.5 Enfoque de validación

En los ejercicios de predicción con machine learning es indispensable separar el conjunto de datos en sets de entrenamiento, validación y testeo, para evaluar la capacidad de los modelos de predecir de manera acertada en datos desconocidos. El subconjunto de datos de entrenamiento es utilizado para entrenar los distintos candidatos modelos (es decir, de dónde se aprenden de manera directa los patrones predictivos). Mientras que los datos de validación

[4] La curva ROC grafica TPR vs. FPR en diferentes umbrales de clasificación y como los modelos estiman probabilidades, para clasificar un resultado dentro de una u otra clase se debe establecer un umbral de decisión, comúnmente se elige 0.5. Sin embargo, este umbral puede no generar resultados satisfactorios cuando las clases de la variable dependiente se encuentran desbalanceadas (Staudt, 2022). El área bajo la curva ROC (AUC-ROC) mide toda el área por debajo de la curva, por lo que provee una medida agregada del desempeño entre todos los posibles umbrales de clasificación.

son utilizados para medir el desempeño predictivo de los modelos, y su relevancia radica en que se elige el modelo que se considera mejor a los fines de predecir en datos cuyo valor de la variable a predecir se desconoce. Por último, una vez elegido el mejor modelo a partir de su entrenamiento y validación, se utiliza el subconjunto de testeo para tener una estimación final de su desempeño en datos desconocidos (Soules, 2020).

En síntesis, se hace uso del set de entrenamiento para estimar el modelo, el de validación para elegir al mejor modelo y al de testeo para evaluar qué tan buena es la performance del modelo elegido para predecir nuevos datos (Varian, 2014). En el caso de este trabajo, se divide la base de datos en tres particiones aleatorias de entrenamiento, validación y testeo. La distribución de observaciones se puede ver en la siguiente tabla.

Tabla 1: Subconjuntos del dataset

Enfoque de validación

Conjunto	Observaciones	Porcentaje del dataset
Entrenamiento	1350	64,0%
Validación	338	16,0%
Test	422	20,0%

Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF, 2021).

5. Análisis exploratorio

Utilizando el conjunto de datos de entrenamiento, la etapa exploratoria hace foco en las diferentes características socioeconómicas de las madres para los casos de bajo y no bajo peso al nacer. Debe dejarse en claro que las estimaciones realizadas presentan una connotación exploratoria, es decir que las asociaciones que se encuentran no deben interpretarse como relaciones causales. En esta sección se analizarán las variables que la literatura encuentra como relevantes.

En primer lugar, se muestra la distribución de casos de bajo y no bajo peso al nacer de la base de datos (Figura 1). Los mismos fueron clasificados según el criterio utilizado frecuentemente en la literatura: considerando como casos de bajo peso al nacer aquellos

menores a 2.5 kg. La proporción de estos casos es de 8.8% en la base de datos, 0.9 puntos porcentuales por encima del porcentaje de los nacidos vivos que presentaron bajo peso al nacer según el último dato disponible de 2022, el cual alcanzó 7.9% (Ministerio de Salud, 2024). Teniendo ello en cuenta, la variable objetivo está desbalanceada, ya que los casos de BPN son considerablemente menores que los casos de peso normal. Ello se tendrá en cuenta a la hora de evaluar el desempeño predictivo de los modelos.

Figura 1: Proporción de casos de con bajo peso y sin bajo peso al nacer

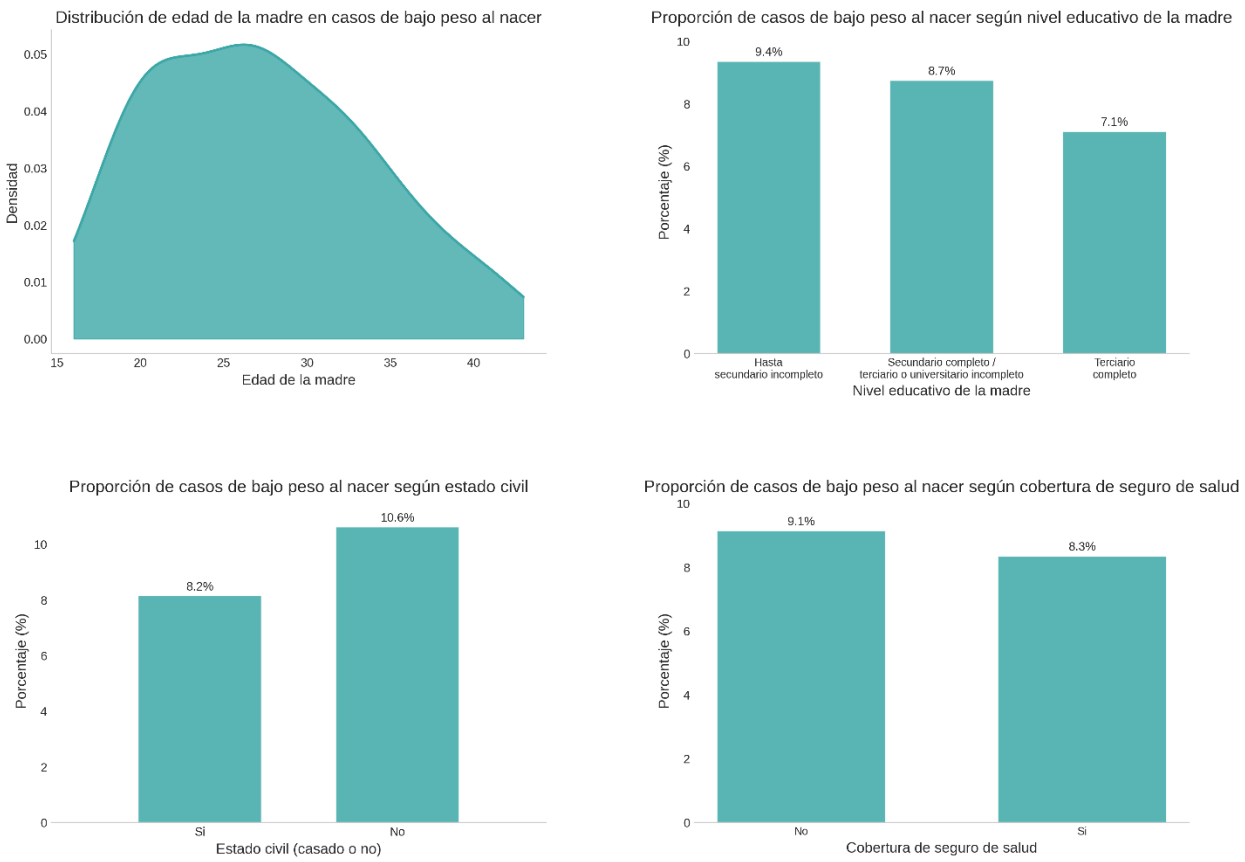
Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF, 2021).

La Figura 2 muestra los casos de bajo peso al nacer según una selección de variables socioeconómicas[5]. En el primer panel, se puede observar que las madres con casos de bajo peso al nacer alcanzan un pico cerca de los 30 años de edad, para disminuir progresivamente a medida que aumenta la edad de la madre. En el siguiente cuadrante se observa la proporción de casos según el nivel educativo alcanzado por la madre, donde se observan mayor proporción de casos en los menores niveles de instrucción.

Para los casos de bajo peso al nacer según el estado civil declarado por las madres, la variable fue reconstruida y reclasificada en dos categorías: casadas (casado civil o en pareja) y solteras (divorciado/separado/nunca tuvo o tiene pareja) para una interpretación más directa de los resultados. Se puede observar que en los casos de madres solteras se encontró el mayor porcentaje de casos de bajo peso al nacer, tal como indica la literatura. Con respecto a los cuidados prenatales, se puede observar que la mayor proporción de casos de BPN se da en madres que no cuentan con cobertura paga de salud.

Figura 2: Proporción de casos de bajo peso según variables socioeconómicas de la madre

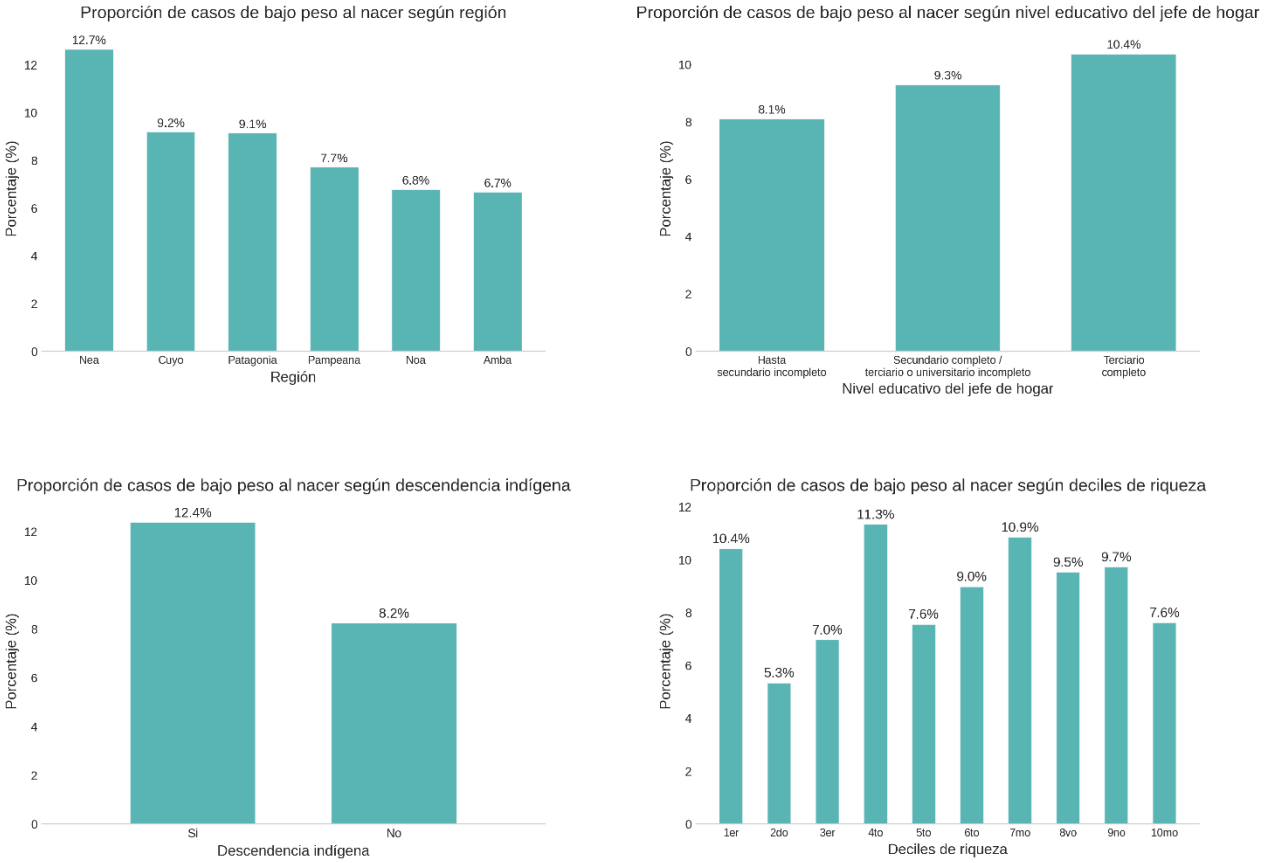
[5] El resto de las figuras puede visualizarse en el Anexo.



Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF, 2021).

En la figura 3 se observan cuatro variables más. El análisis respecto a los casos de BPN según la región del país arroja que fueron mayores los casos en las zonas menos desarrolladas. En el Noreste (NEA) se registró la mayor proporción de casos (12,7%), seguido por Cuyo (9,2%). En este sentido, la región NEA había registrado en el primer semestre de 2020 (el período más cercano al momento de realizada la encuesta) una tasa de pobreza en personas del 42.8%, siendo la región peor posicionada en ese momento en el indicador mencionado. Observando respecto al nivel educativo del jefe de hogar, la mayor proporción de casos de BPN se produjo en donde el jefe de hogar alcanzó terciario completo. A su vez, la menor proporción de casos se observó en donde el jefe de hogar alcanzó hasta secundario incompleto, lo cual no coincide con la literatura. Es decir que en esta ocasión los menores casos de BPN se dieron en el menor nivel educativo de los jefes de hogar.

Figura 3: Proporción de casos de bajo peso según variables socioeconómicas del hogar y el jefe del hogar



Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF, 2021).

Con respecto a la etnia, la literatura afirma la existencia de disparidades en el peso al nacer entre distintas descendencias. En este sentido, en la base de datos se observa que quienes afirman tener descendencia indígena presentaron una proporción de casos de BPN mayor que quienes afirmaron no tener esta descendencia. Siguiendo con las variables socioeconómicas, el análisis de casos de BPN según los deciles de riqueza de los hogares donde habitan las madres parece no mostrar evidencia clara de un predominio exclusivo de casos de BPN sobre los deciles más bajos. Los menores porcentajes de casos se encontraron en el segundo y tercer decil, mientras que la mayor proporción se ubicó en el cuarto. Estos resultados contrastan con los examinados por la literatura, que indica que los menores deciles de riqueza tenderían a mostrar una mayor proporción de casos de BPN.

Por último, el análisis de las características que se vinculan a la educación, los cuidados prenatales, el estado civil, las regiones del país, la cobertura de salud, y la descendencia indígena presentadas anteriormente se alinean con los resultados encontrados por la literatura de bajo peso al nacer, exceptuando el caso del nivel educativo del jefe del hogar. Estos resultados exploratorios sugieren que dichas variables podrían influir negativamente en la probabilidad de bajo peso al nacer.

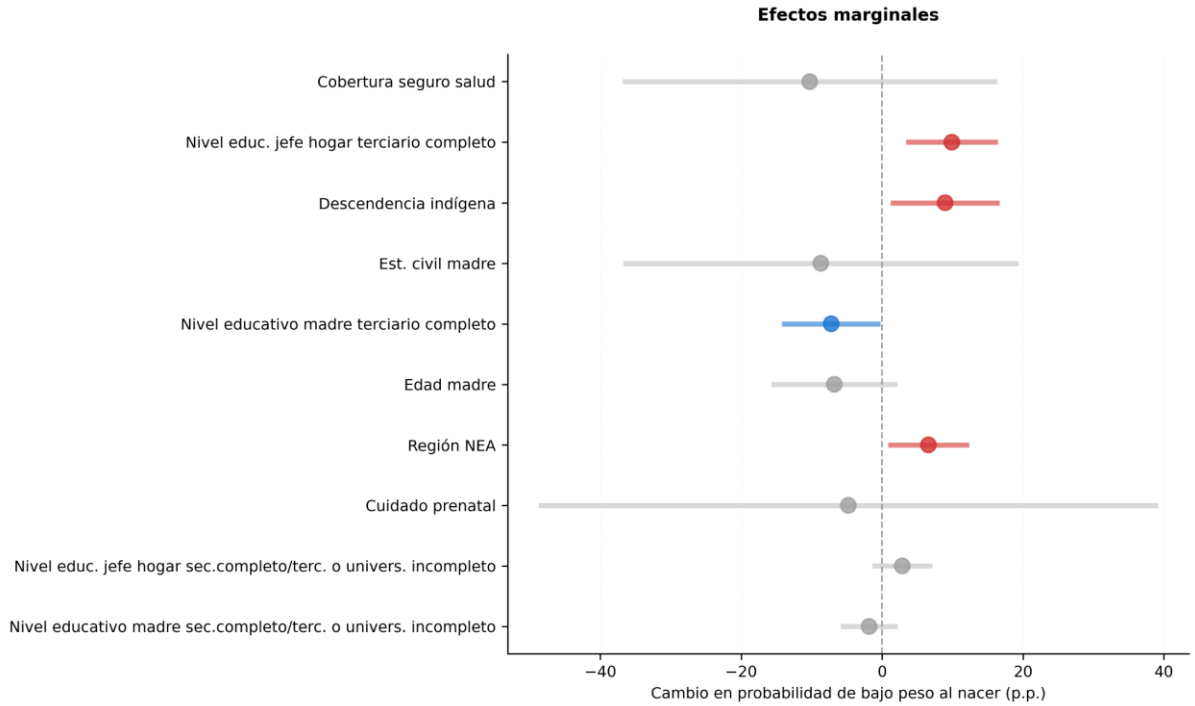
6. Resultados

En esta sección se presentan los resultados de los modelos utilizados para predecir el bajo peso al nacer. En primer lugar, se toma como punto de partida el modelo de regresión logística, incluyendo una ampliación del conjunto de variables predictoras utilizando una técnica de variables de selección hacia adelante. Posteriormente, se compara este modelo con los enfoques de machine learning empleados: lasso y random forest.

6.1 Regresión logística

En primer lugar, en base a las variables socioeconómicas analizadas en la literatura, se corrió un modelo de regresión logística.[6] En el siguiente gráfico se presentan los coeficientes del modelo expresados en probabilidades con las variables estudiadas en la sección de revisión de la literatura.

Figura 4: Coeficientes del modelo con variables recomendadas por la literatura



[6] La salida de regresión completa se puede observar en el Anexo. Para interpretar los coeficientes de regresión logística se debe tener en cuenta que los mismos representan el cambio en el logaritmo de la odds ratio (OR) de que un evento ocurra (la categoría 1 de la variable dependiente) en respuesta a un cambio unitario en la variable independiente del coeficiente, manteniendo constantes las demás variables independientes del modelo (Molina, 2024). Los odds ratio se definen como el cociente de la probabilidad de presentar una característica y la probabilidad de no presentarla, o lo que es lo mismo el cociente del número de casos que presentan la característica entre el número de casos que no la presentan (Peláez, 2016). Para la presentación de los resultados en este apartado, se realiza un cálculo adicional para obtener las probabilidades de las variables reportadas.

Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF, 2021).

Las variables estadísticamente significativas del modelo ($p < 0.05$) resultaron descendencia indígena, nivel educativo de la madre terciario completo, región NEA y nivel educativo del jefe de hogar terciario completo.[7] Analizando los resultados en base a estas variables, la descendencia indígena presenta 8.9 p.p. más de probabilidad de tener un niño con BPN comparado con quienes no presentan descendencia de este tipo[8]. A su vez, las madres con un nivel educativo terciario completo presentan una probabilidad de caso de BPN de sus hijos en 7.2 p.p. menor al nivel educativo hasta secundario incompleto. Contrariamente, a mayor nivel educativo del jefe de hogar la probabilidad de bajo peso aumenta (9.8 p.p.).

En cuanto a las regiones, nacer en la región NEA implica 6.5 p.p. más de probabilidad de tener un hijo con BPN, mientras que los deciles de riqueza no mostraron influencia. El decil más cercano a ser significativo fue el segundo, pero aun así no alcanzó el umbral de significancia[9].

El desempeño del modelo con las variables socioeconómicas investigadas en la literatura alcanzó un poder de clasificación moderado, con un área bajo la curva ROC de 0.67.[10] El AUC-ROC obtenido indica que en promedio el modelo puede diferenciar correctamente entre un caso de bajo peso al nacer y uno sin esta condición el 67% de las veces.[4] Si bien este valor es mayor que 0.5, lo que implicaría un rendimiento equivalente al azar, el mismo sugiere que el poder predictivo de las variables socioeconómicas incluidas es limitado, por lo que podrían existir otros factores importantes no incluidos en el modelo, que influyen en el bajo peso al nacer.

Para optimizar la selección de variables, se utilizó el método de selección hacia adelante. Las variables edad, región NEA, descendencia indígena, estado civil y región Pampeana fueron las que más aportaron a la mejora en el valor del AUC-ROC final[11]. Con estas variables seleccionadas, el modelo mejoró su performance a 0.77 (+10 puntos porcentuales que el modelo sin selección). Es decir que, bajo este modelo con las variables seleccionadas, existe

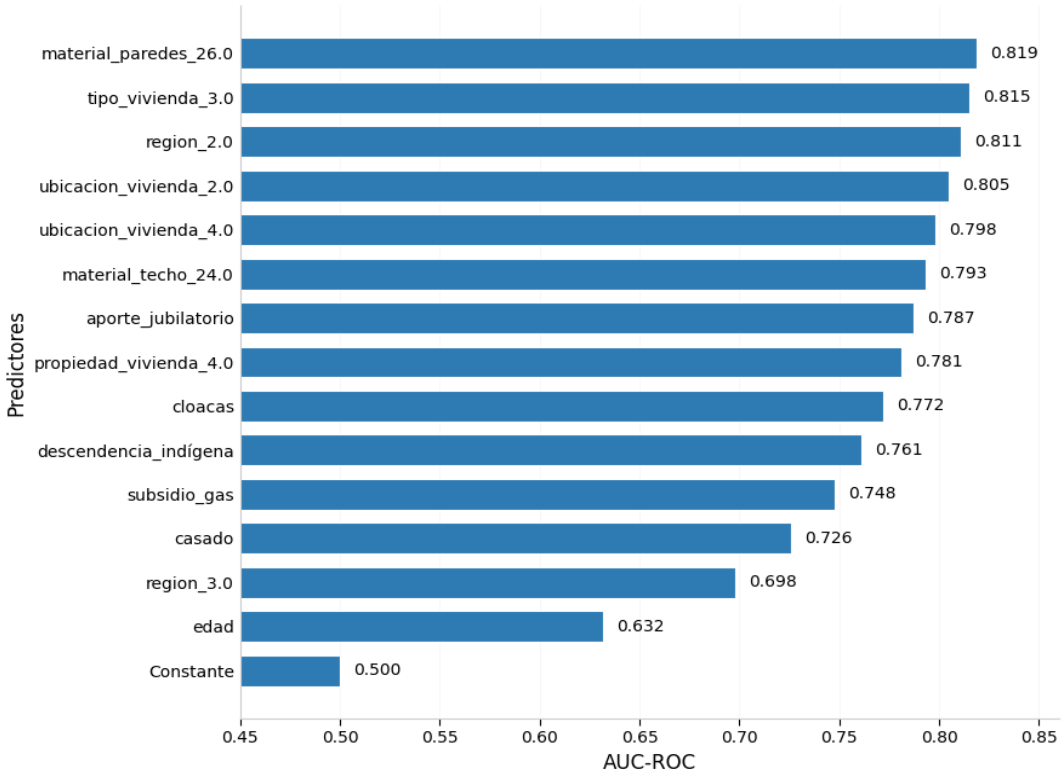
[7] Los intervalos de confianza de estas variables no incluyen el 0, confirmando la significancia estadística de las mismas.
[8] Estos resultados deben tomarse con cautela, ya que los mismos presentan una connotación exploratoria, no así hallazgos de índole causal.
[9] Los coeficientes de dichas variables se encuentran en el Anexo en la salida de regresión completa. Se omitieron en la figura.
[10] El gráfico puede visualizarse en el Anexo.
[11] La tabla con las variables seleccionadas y su aporte al AUC-ROC puede visualizarse en el Anexo.

un 77% de probabilidad de clasificar correctamente entre un caso positivo (bajo peso al nacer) y uno negativo (caso contrario)[12].

6.1.1 Ampliación del conjunto de variables

Considerando que el modelo inicial presentaba un poder predictivo limitado, se decidió extender el rango de predictores disponibles. En este sentido, la base de datos de UNICEF (2021) contiene información adicional que podría ayudar a mejorar la estimación y resultados del modelo para predecir el bajo peso al nacer. Se incluyen variables relacionadas con el hogar y los miembros del mismo como ser edad del jefe del hogar, materiales de los pisos, paredes y techos, si la madre percibe descuento jubilatorio, categoría ocupacional de la madre, tipo de vivienda, si cobra seguro de desempleo, entre otras. De esta manera se trata de captar información socioeconómica no presente en las variables seleccionadas en la literatura, que permitiría mejorar la predicción del bajo peso al nacer medida por la performance de la curva AUC-ROC.

Figura 5: Ganancia en AUC-ROC de las variables seleccionadas por el método de selección hacia adelante

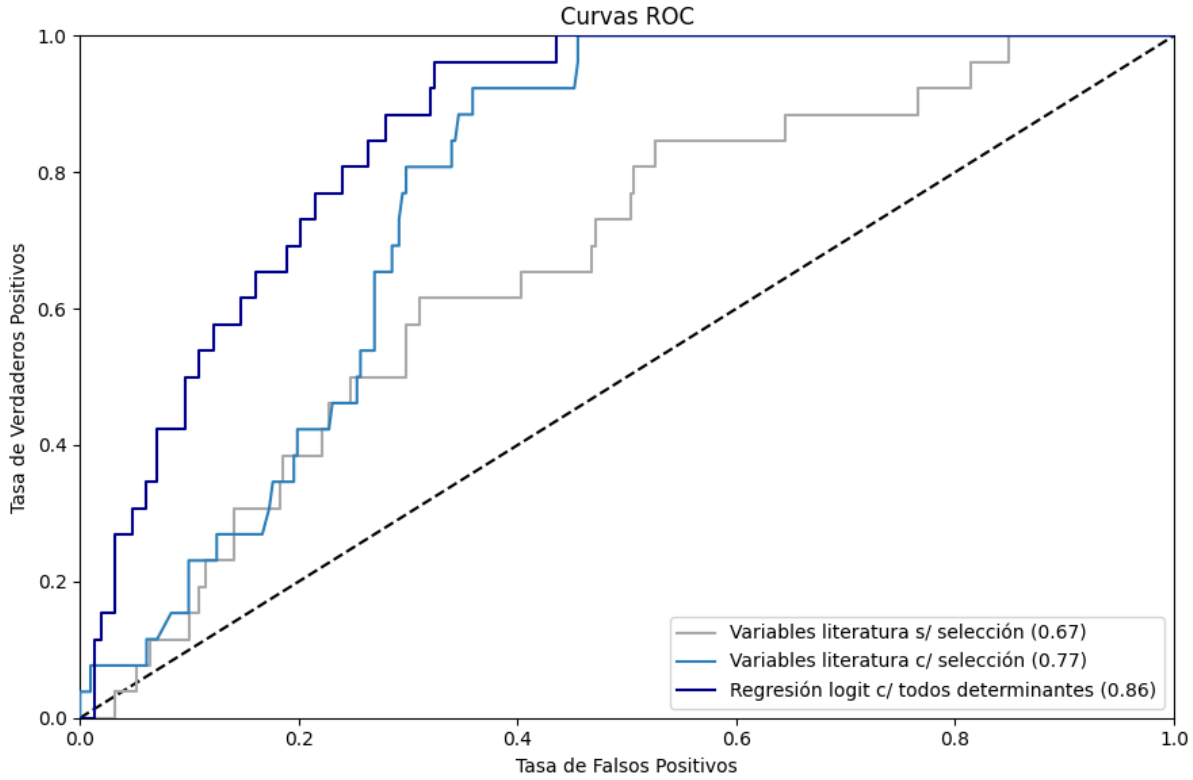


Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF, 2021).

[12] El gráfico comparativo de las AUC-ROC puede visualizarse en el Anexo.

Nuevamente en la selección de variables las más relevantes fueron edad y región NEA. Sin embargo, aparecen otras como: si el hogar percibe subsidio de gas o los tipos de propiedad de la vivienda. En la totalidad, con la técnica de selección de un mayor set de covariables que la especificación inicial, el modelo de regresión logística arrojó un valor AUC-ROC de 0.86. Se puede destacar en el gráfico comparativo (Figura 6) la brecha que se observa en el desempeño de predicción, a partir de valores intermedios de Tasa de Falsos Positivos (TFP). Desde una TFP de 0.3 aproximadamente la capacidad predictiva del modelo bajo selección mejora notablemente, ya que para una misma TFP la Tasa de Verdaderos Positivos (TVP) aumenta considerablemente. De hecho, el modelo bajo selección llega a una TVP de 1 a costa de una TFP cercana al 0.5 únicamente, mientras que en el modelo sin selección una TVP de 1 se alcanza a costa de una TFP de más del 0.8.

Figura 6: Curvas AUC-ROC de cada modelo logit



Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF, 2021).

6.2 Comparación de modelos: Regresión logística, lasso y random forest

Una vez establecida la regresión logística como modelo base, se procede a comparar su desempeño con los modelos de machine learning: lasso y random forest[13]. Esta comparación

[13] La optimización de parámetros de cada modelo puede visualizarse en el Anexo.

se realiza desde dos perspectivas: la selección de variables más importantes de cada modelo y el desempeño predictivo global. En la Tabla 2 pueden visualizarse las 10 primeras variables seleccionadas para cada modelo y sus respectivos coeficientes (para logit [14] y lasso) e importancia de variables (en el caso de random forest).

Tabla 2: Comparación de las 10 primeras variables seleccionadas por los modelos

Regresión logística	**Coeficiente**	**Lasso**	**Coeficiente**	**Random forest**	**Importancia variable**
Edad de la madre	-0.03	2° decil riqueza	-0.14	Edad jefe de hogar	0.09
Región NEA	0.05	Región NEA	0.11	Edad de la madre	0.06
Estado civil casado	-0.02	Nivel educativo jefe hogar (secundario completo/terc. o univ. incom.)	0.09	Cant. habitaciones	0.04
Subsidio gas	0.02	Descendencia indígena	0.09	Descendencia indígena	0.03
Descendencia indígena	0.04	Material piso (parqué)	-0.08	Categoría ocupacional	0.02
Cloacas	0.006	Edad de la madre	-0.07	Desagüe	0.02
Propiedad vivienda (cedida por trabajo)	0.009	Electricidad por red	-0.05	Subsidio gas	0.02
Aporte jubilatorio	0.02	Material techo (tabla)	-0.04	7° decil de riqueza	0.02
Material techo (chapa)	-0.08	Material techo (palma)	0.03	Nivel educativo jefe hogar (secundario completo/ terc. o Univ. incomp.)	0.019
Ubicación vivienda (otros)	0.06	Material techo (otro)	0.03	Región NEA	0.019
AUC-ROC logit	0.86	**AUC-ROC Lasso**	0.60	**AUC-ROC Random forest**	0.69

[14] En el caso de la regresión logística los coeficientes se expresaron en términos de probabilidades (efectos marginales) para una interpretación más intuitiva de los resultados. Cabe aclarar, que la interpretación tiene una connotación puramente correlacional y no se busca establecer vínculos causales.

Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF, 2021).

En términos de edad, en el modelo logit un año adicional reduce la probabilidad de bajo peso en 3.4 puntos porcentuales (p.p.), mientras que pertenecer a la región NEA implica una probabilidad de bajo peso 5.5 p.p. mayor que pertenecer a la región de referencia (AMBA). Por su parte, el estado civil de la madre (casada) se corresponde con una probabilidad de bajo peso de 2.0 p.p. menor que en el caso de las madres con estado civil soltera. Con respecto al modelo lasso, la variable región NEA también implicó una probabilidad de bajo peso al nacer mayor. Nuevamente, la descendencia indígena es otra de las variables principales en este modelo, presentando mayor probabilidad de bajo peso al nacer al tener esta descendencia. A su vez, mayor edad materna se asocia con menor riesgo de bajo peso al nacer.

En el modelo random forest, la importancia de variables indicó que la edad del jefe de hogar fue la variable más destacada, ya que su contribución al modelo fue de 9.9%. En segundo lugar, se encontró la edad de la madre, con 6.1%. La descendencia indígena vuelve a aparecer también en este modelo, en este caso con 3.3%. Tanto lasso como random forest remarcaron el subsidio de gas entre las principales variables relevantes. Nuevamente aparece la región NEA como determinante, en el caso de random forest, en una posición mucho menor que en los casos anteriores (1.9%).

La comparación revela que ciertas variables aparecen consistentemente en los tres modelos como predictores importantes: edad (de la madre o del jefe de hogar), región NEA, descendencia indígena y subsidio de gas. Esto sugiere que las mismas tienen un efecto robusto en la predicción del bajo peso al nacer, independientemente del método de estimación utilizado.

El modelo lasso seleccionó 18 variables, priorizando aquellas con mayor capacidad predictiva mientras penaliza la complejidad del modelo. El modelo random forest, por su parte, seleccionó 85 de las 95 variables disponibles, lo que refleja su capacidad para manejar un mayor número de predictores sin sufrir sobreajuste, a través de la técnica del bagging.

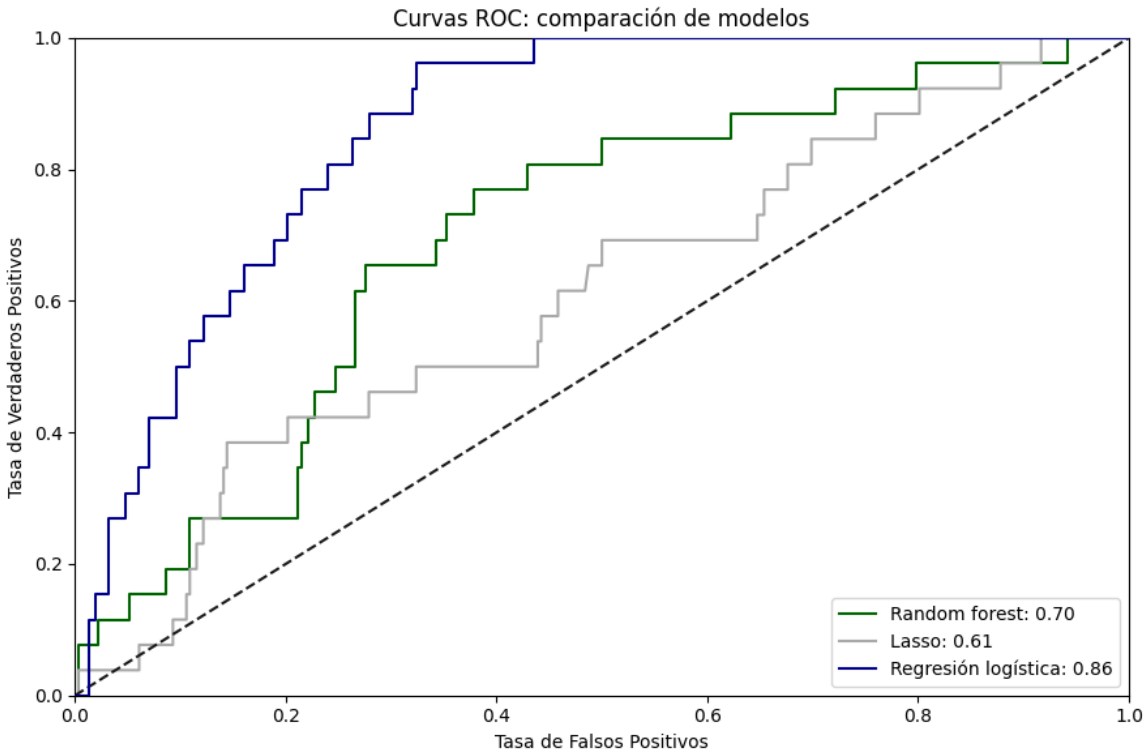
6.2.2 Desempeño predictivo

El análisis comparativo del desempeño predictivo muestra diferencias significativas entre los modelos. La regresión logística obtuvo un área bajo la curva de 0.86, indicando una capacidad predictiva superior a los otros modelos analizados. Este valor sugiere que el modelo de

regresión logística es eficaz para distinguir entre los casos de bajo peso al nacer y los que no lo son.

El modelo random forest obtuvo un AUC-ROC de 0.70, indicando una capacidad predictiva buena pero inferior a la regresión logística. Por su parte, el modelo lasso obtuvo un AUC-ROC de 0.61, apenas por encima del azar (0.5), lo que indica una capacidad predictiva limitada para este conjunto de datos específico.

Figura 7: AUC-ROC de los mejores modelos de cada estimador

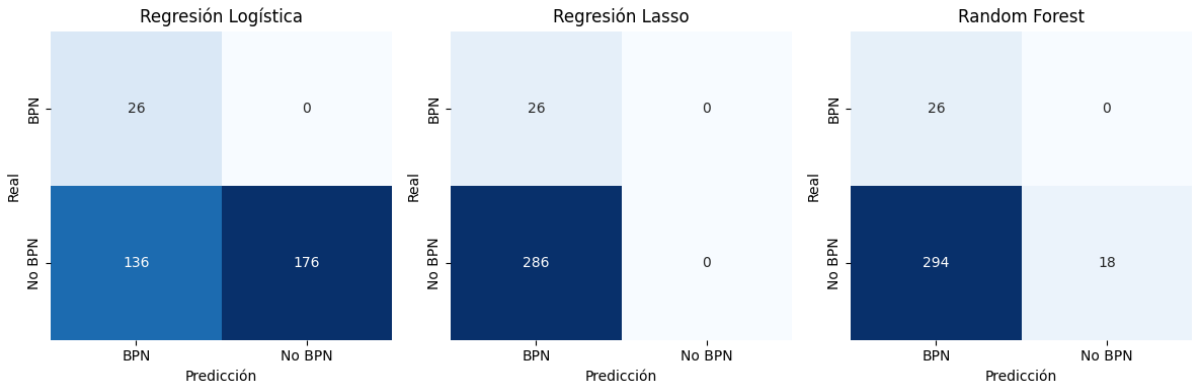


Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF, 2021).

6.2.3 Análisis de matrices de confusión

Para profundizar el análisis de la capacidad predictiva de los modelos se forzó a que cada uno de ellos alcance una Tasa de Verdaderos Positivos (TVP) del 100%, es decir, que acierten de manera correcta todos los casos de bajo peso al nacer.

Figura 8: Matrices de confusión



Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF, 2021).

Al analizar la performance de los modelos, la regresión logística arroja un mejor desempeño general entre los tres modelos evaluados. Este modelo logra mantener un equilibrio razonable de Verdaderos Negativos (176/312, es decir, 56%), con una exactitud del 60% y una media ponderada de 0.69. La tasa de falsos positivos fue de 136 casos, considerablemente menor que en los otros modelos.

Tabla 3: Umbral de decisión que maximiza la Tasa de Verdaderos Positivos (TVP) y métricas de evaluación

Modelo	Umbral de decisión que maximiza TVP	Precisión	F1-score	Exactitud	Macro Average	Weighted Average
Regresión logística	0.48	0.16	0.28	0.60	0.50	0.69
Lasso	0.40	0.08	0.15	0.15	0.15	0.15
Random forest	0.05	0.08	0.15	0.13	0.13	0.11

Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF) (2021).

Por otro lado, tanto la regresión lasso como random forest tuvieron un rendimiento significativamente inferior, con tasas de falsos positivos de 286 y 294 casos respectivamente, lo que implica que la gran mayoría de casos de peso normal serían clasificados erróneamente como de bajo peso.

Del análisis comparativo se selecciona al modelo de regresión logística como el de mejor desempeño para la tarea objetivo de este trabajo. La evaluación final del modelo sobre el

conjunto de datos de prueba arrojó un AUC-ROC de 0.84, representando una caída de sólo 2 puntos porcentuales respecto al desempeño en validación, lo que demuestra robustez en la predicción ante datos desconocidos.

Si bien supera a los objetivos propuestos en este trabajo, una de las razones hipotéticas de porqué el modelo de regresión logística podría haberse desempeñado mejor en este contexto radica en que los métodos más avanzados de machine learning se destacan al existir relaciones complejas y no lineales. Pero la temática de este trabajo podría seguir una estructura donde las relaciones entre los factores socioeconómicos y el bajo peso se comportan de manera lineal y aditiva.

El modelo final identifica correctamente 33 de 43 casos positivos observados, reflejando una tasa de verdaderos positivos del 76% y una tasa de falsos positivos del 28%. El AUC-ROC de 0.84 indica un rendimiento sólido del modelo para distinguir entre las clases, confirmando que la regresión logística constituye la mejor opción para la predicción del bajo peso al nacer utilizando variables socioeconómicas en este contexto específico.

6.2.4 Resultado final

En la Figura N° 9 se trazan las proporciones de casos de bajo peso sobre cada categoría de las variables región, edad de la madre, el nivel educativo de las mismas y el estado civil (casada o no) sobre la base de datos de prueba con los valores observados de la variable dependiente. Mientras que en la Figura 10 se realiza el mismo cálculo pero para los valores predichos del BPN que arroja el modelo logit de mejor desempeño.[15] Como se puede observar en ambas figuras, las distribuciones de cada predictor no presentan cambios abruptos entre lo observado y los valores predichos, mostrando consistencia en el modelo ante la predicción de nuevas observaciones. No obstante, la proporción de casos positivos predichos es mayor que el valor real, debido a una sobreestimación del modelo ante la presencia de casos de BPN. Hecho que de todas formas es esperable, dada la búsqueda de mejorar la detección de casos de bajo peso.

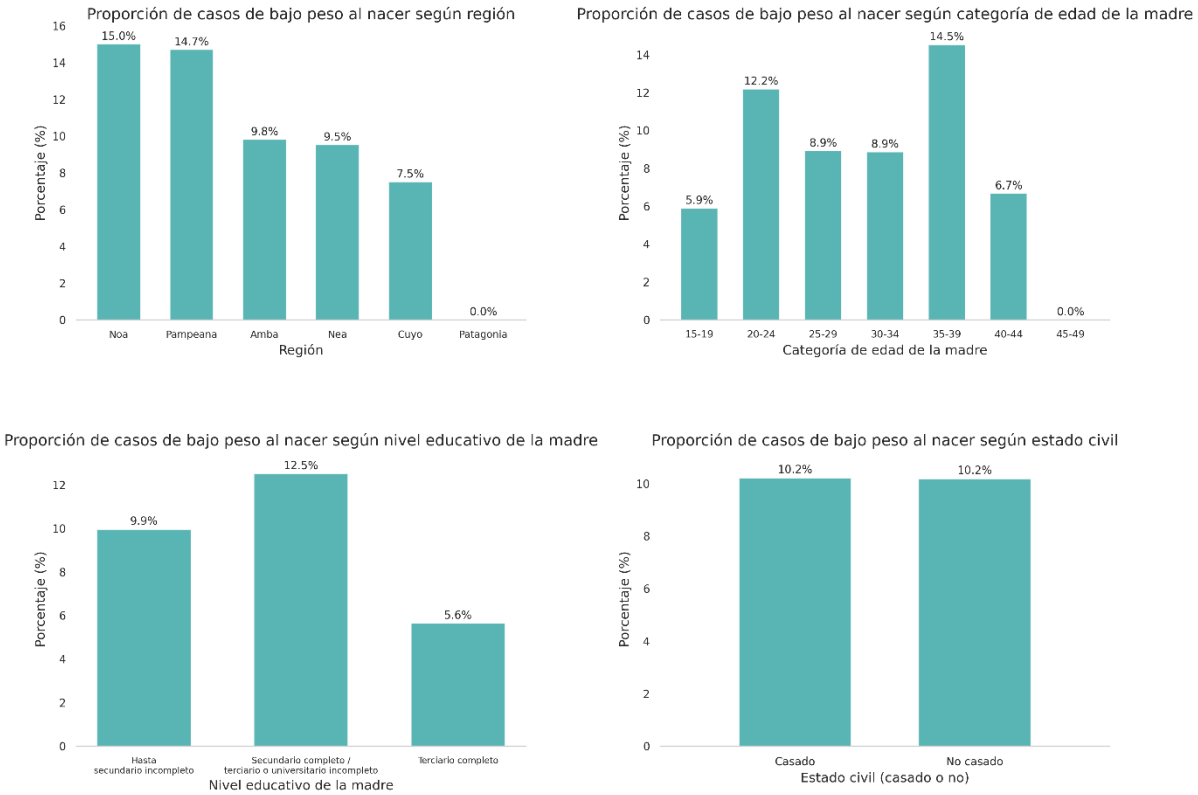
El análisis realizado en apartados anteriores muestra que el modelo logit presenta un mayor desempeño global en relación a los otros modelos, dado por los valores computados en el AUC-ROC. No obstante, al traducir estos resultados a escenarios concretos para la toma de

[15] A diferencia de lo realizado en el apartado de comparación de modelos, para clasificar las probabilidades estimadas por logit se tuvo en cuenta el umbral de decisión que maximiza el f1-score. Dicha métrica balancea la precisión y la sensibilidad (recall), siendo de utilidad cuando las clases están desbalanceadas (Núñez Nepomuceno, 2025), como es el caso de este trabajo. Sin recurrir a una estrategia fuerte de maximizar los verdaderos positivos.

decisiones debemos pensar en umbrales de decisión para clasificar las probabilidades predichas en casos positivos o negativos, lo cual depende de las prioridades del tomador de decisiones. El trabajo no profundiza este análisis, sino que toma una métrica convencional (como es el f1-score) para poner en práctica el modelo. Sin embargo, queda para propuestas futuras dedicarle una mayor profundización a establecer umbrales de decisión óptimos basados en el contexto de análisis estudiado. Siguiendo los resultados del modelo predicho, en las regiones NOA, Pampeana, AMBA y NEA se detecta la mayor proporción de casos positivos. En términos de edad, los menores casos se identifican en la edad más avanzada del rango, sin embargo, entre los 20 y 24 años se verifica un porcentaje relevante de casos de bajo peso, en conjunto con el rango de entre 35 y 39 años. Por su parte, el nivel educativo parece indicar un mayor porcentaje de casos de bajo peso al nacer en madres con nivel educativo hasta secundario incompleto y terciario/universitario sin finalizar.

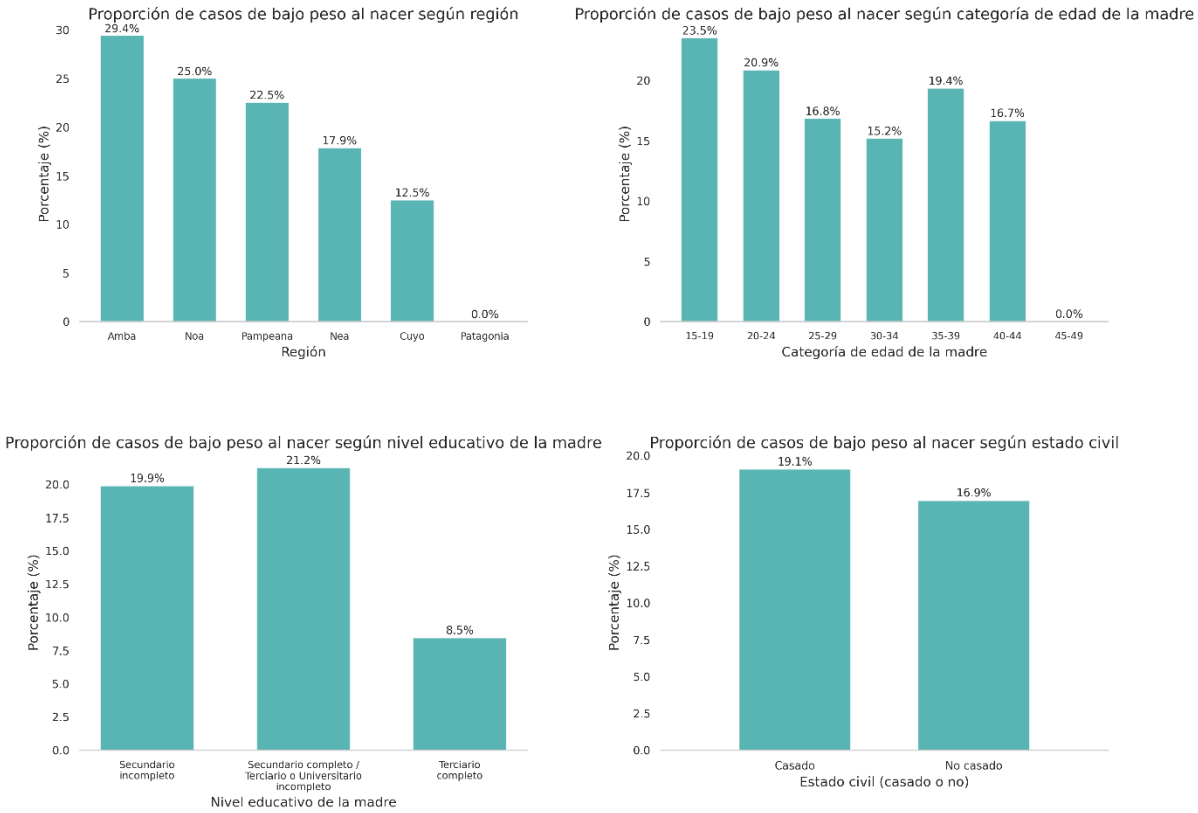
En términos de implicancias de política pública, los resultados hallados permitirían, bajo una serie de encuestas que capten la misma información de las covariables de nuestro modelo, identificar potenciales casos de bajo peso al nacer. Ello daría la posibilidad de brindar el acompañamiento adecuado a estos hogares para mitigar potenciales instancias de BPN. En este mismo sentido, la Fundación Abrazar trabaja con casos de violencia infantil en el hogar, realizando este tipo de intervención a partir de algoritmos que le permiten a los gobiernos locales identificar hogares que necesiten un espacio de orientación para mejorar el vínculo de crianza con sus hijos (Fundación Abrazar, s.f.). Por otro lado, la caracterización de la población analizada genera evidencia para poder redirigir los esfuerzos y recursos de la implementación de políticas a los grupos poblacionales más propensos a tener un recién nacido con bajo peso. No sólo a partir de las características vistas en la Figura 9, sino teniendo otros rasgos como la descendencia indígena, el aporte jubilatorio o el hecho de recibir subsidio de gas.

Figura 9: Proporción de casos de bajo y no bajo peso al nacer según región, edad materna, nivel educativo de la madre y estado civil (casada). Con variables observadas.



Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF, 2021).

Figura 10: Proporción de casos de bajo y no bajo peso al nacer según región, edad materna, nivel educativo de la madre y estado civil (casada) (c/ umbral maximiza el f1-score)



Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF, 2021).

Como remarca Wulandari et al. (2023) la parte más crítica de la implementación de políticas para la prevención del BPN es la educación de mujeres embarazadas y jóvenes que se preparan para el embarazo. A su vez, la OMS (2017) remarca que para atacar a la problemática del bajo peso al nacer los países deben aplicar políticas basadas en evidencia como el impulso al nivel de instrucción de las mujeres (como se vio anteriormente, hay evidencia de mayores casos de bajo peso al nacer en mujeres con menor nivel educativo). Además, la OMS hace hincapié en el uso de los sistemas de información para un mejor monitoreo de la salud de los recién nacidos, a través de un sistema universal simplificado de recopilación de datos perinatales que permita un seguimiento exhaustivo y su posterior intervención.

7. Conclusiones

El presente trabajo se propuso predecir el bajo peso al nacer, comparando diferentes modelos predictivos y analizando los principales determinantes socioeconómicos detrás de las predicciones realizadas. De esta manera, se buscó el modelo de mayor consistencia y poder predictivo, como también, se indagó acerca de la influencia que las características de la madre y del hogar tienen sobre dicha predicción. Se utilizaron tres estimadores bajo un

análisis comparativo: la regresión logística, el estimador lasso y el modelo de random forest, utilizados comúnmente en la literatura de predicción del bajo peso al nacer.

Con respecto a la comparación de modelos, el estimador de regresión logística presenta un poder predictivo significativamente mayor a lasso y random forest, siendo este último el segundo más competitivo, cuyos valores AUC-ROC son significativamente mayores a lasso. El resultado hallado implica que las técnicas más avanzadas no superaron en término de performance a logit para predecir el bajo peso al nacer. Este mismo hallazgo, puede encontrarse en otros trabajos con la misma índole (Borson et al., 2020; Islam Pollob et al., 2022; Khan et al., 2022; Arayeshgari et al., 2023; Patterson et al., 2023; Mathew y Thinakaran, 2025).

El mejor modelo de regresión logística tiene un 84% de probabilidades de que, si se eligen dos observaciones desconocidas al azar, una positiva (bajo peso al nacer) y otra negativa (de no bajo peso al nacer), la probabilidad predicha para la primera sea mayor que para la de no bajo peso al nacer, lo que equivale a 14 puntos porcentuales más que el desempeño estimado en random forest. Resultado que muestra la consistencia del mejor modelo, como también, un poder predictivo competitivo por parte de logit. No obstante, aún existe un margen no despreciable de error del estimador, lo cual sugiere que podrían existir inobservables, y variables de baja correlación con los factores socioeconómicos, que son relevantes para predecir el BPN.

Entre las variables más influyentes se encuentra la edad de la madre. Utilizando técnicas de selección dicho predictor se encuentra entre los más influyentes de la predicción. En el caso de logit la variable edad de la madre es la más importante, mientras que en la regresión lasso se ubica en el sexto lugar. En random forest, se posiciona en segundo lugar. Una edad materna temprana como una muy avanzada implican un potencial riesgo de bajo peso al nacer. En el primer caso, por un posible escenario de competencia por nutrientes, y en el segundo, por la disminución en el potencial de crecimiento fetal (Aras, 2013; Restrepo-Méndez et al., 2015).

Además de la edad, se encuentran otras características de la madre relevantes que se alinean con los hallazgos de la literatura de bajo peso al nacer. El estado civil, la región de pertenencia, el nivel educativo, la descendencia indígena y variables como los materiales de la vivienda, o la recepción de subsidios, son algunos de los factores que se relacionan con la probabilidad de bajo peso al nacer de los niños. Por lo que este trabajo destaca la necesidad de considerar como factor central el estatus socioeconómico para el estudio del bajo peso al nacer.

Los resultados hallados implican que el abordaje de la problemática del BPN puede verse beneficiada con la aplicación de técnicas estadísticas de predicción, utilizando las mismas como herramientas de identificación de potenciales casos para facilitar el trabajo de los profesionales y/o tomadores de decisiones de política pública, sin reemplazar su trabajo base. A su vez, este trabajo aporta evidencia a favor de que la regresión logística supera a los clasificadores de machine learning, aunque la exploración de otras técnicas más avanzadas podría traer mayores beneficios en la búsqueda de detectar casos de bajo peso al nacer.
En cuanto a la caracterización socioeconómica de los casos de bajo peso al nacer, los resultados encontrados aportan evidencia para poder diseñar políticas públicas enfocadas a los grupos poblacionales más propensos a tener un recién nacido con bajo peso. Ello daría la posibilidad de brindar el acompañamiento adecuado a estos hogares para mitigar potenciales instancias de riesgo de bajo peso que, como documenta la literatura, conlleva un efecto negativo en los resultados futuros del recién nacido.

Por supuesto, el estudio presenta algunas limitaciones. En primer lugar, la utilización de una base de datos tendiente al análisis de características estructurales de los hogares y el faltante de algunas variables relevantes en términos de salud podrían no captar toda la influencia hacia el fenómeno bajo estudio. Como se mencionó anteriormente, los determinantes y factores asociados al bajo peso al nacer incluye factores biológicos, factores demográficos, factores socioeconómicos y factores nutricionales (Khan et al., 2020). Este trabajo se limita a los factores socioeconómicos.

Por otra parte, respecto al relevamiento de la variable dependiente, al ser una pregunta retrospectiva puede presentar sesgos dado por el recuerdo y el contexto donde el niño fue concebido. Estas limitaciones pueden generar errores de clasificación y/o subestimación a la hora de realizar las predicciones. Un hecho a remarcar es que la base de datos utilizada cuenta con el dato de bajo peso al nacer sólo de un momento en el tiempo, ya que fue incluida únicamente en el último relevamiento realizado.

Finalmente, como recomendación de trabajos futuros es necesario explorar otras estrategias para predecir el bajo peso al nacer, como también, pensar en alternativas superadoras de fijación del umbral de decisión, adaptado al contexto de estudio y las decisiones específicas de política. Entre las estrategias de predicción a futuro, se podría poner el foco en explorar interacciones de las variables en los modelos lineales y la exploración de otras técnicas de predicción más avanzadas para mejorar la performance de los modelos estudiados en el presente estudio.

Referencias bibliográficas

Ahmadi, P., Alavimajd, H., Khodakarim, S., Tapak, L., Kariman, N., Amini, P., y 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. https://doi.org/10.22037/jps.v8i3.15412

Aizer, A., y Currie, J. (2014). The intergenerational transmission of inequality: maternal disadvantage and health at birth. Science, 344(6186), 856-861. https://doi.org/10.1126/science.1251872

Akbulut-Yuksel, M., Cilasun, S., y Turan, B. (2020). Children of crisis: The effects of economic shocks on newborns. IZA Discussion Paper No. 12898. https://dx.doi.org/10.2139/ssrn.3525225

Alderman, H., y Behrman, J. R. (2006). Reducing the incidence of low birth weight in low-income countries has substantial economic benefits. The World Bank Research Observer, 21(1), 25-48. https://doi.org/10.1093/wbro/lkj001

Alexander, G. R., y Korenbrot, C. C. (1995). The role of prenatal care in preventing low birth weight. The future of children, 103-120. https://doi.org/10.2307/1602510

Almond, D., Chay, K. Y., y Lee, D. S. (2005). The costs of low birth weight. The Quarterly Journal of Economics, 120(3), 1031-1083. https://doi.org/10.1093/qje/120.3.1031

Altenhöner, T., Köhler, M., y Philippi, M. (2016). The relevance of maternal socioeconomic characteristics for low birth weight–a case-control study. Geburtshilfe und Frauenheilkunde, 76(03), 248-254. https://doi.org/10.1055/s-0042-100204

Anderson, K. M. (2022). Maternal Socioeconomic Status and Infant Low Birth Weight: Interactions Across Generations. University Honors Theses. Paper 1204. https://doi.org/10.15760/honors.1203

Anjum, F., Javed, T., Afzal, M. F., y Sheikh, G. A. (2011). Maternal risk factors associated with low birth weight: A case control study. Annals of King Edward Medical University, 17(3), 223-223. https://doi.org/10.21649/akemu.v17i3.338

Aras, R. Y. (2013). Is maternal age risk factor for low birth weight?. Archives of medicine and health sciences, 1(1), 33-37. https://doi.org/10.4103/2321-4848.113558

Arayeshgari, M., Najafi-Ghobadi, S., Tarhsaz, H., Parami, S., y Tapak, L. (2023). Machine learning-based classifiers for the prediction of low birth weight. Healthcare Informatics Research, 29(1), 54-63. https://doi.org/10.4258/hir.2023.29.1.54

Behrman, J. R., y Rosenzweig, M. R. (2004). Returns to birthweight. Review of Economics and statistics, 86(2), 586-601. https://doi.org/10.1162/003465304323031139

Bekele, W. T. (2022). Machine learning algorithms for predicting low birth weight in Ethiopia. BMC medical informatics and decision making, 22(1), 232. https://doi.org/10.1186/s12911-022-01981-9

Berniell, L., y De La Mata, D. (2022). RED 2022 - Capítulo 3: La formación del capital humano y la movilidad intergeneracional. https://scioteca.caf.com/handle/123456789/1977

Borson, N. S., Kabir, M. R., Zamal, Z., y Rahman, R. M. (2020). Correlation analysis of demographic factors on low birth weight and prediction modeling using machine learning techniques. In 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4) (pp. 169-173). IEEE. https://doi.org/10.1109/WorldS450073.2020.9210338

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32. https://link.springer.com/article/10.1023/a:1010933404324

Burdette, A. M., Weeks, J., Hill, T. D., y Eberstein, I. W. (2012). Maternal religious attendance and low birth weight. Social Science & Medicine, 74(12), 1961-1967. https://doi.org/10.1016/j.socscimed.2012.02.021

Chakraborty, C., y Joseph, A. (2017). Machine learning at central banks. Bank of England Working Paper No. 674. http://dx.doi.org/10.2139/ssrn.3031796

Chaluh, M. (2023). Estimando la Importancia Relativa de las Expectativas de Inflación: Un Estudio de Machine Learning en el Contexto de Argentina, 2010-2023. 15° Premio de Investigación Económica “Dr. Raúl Prebisch” 2023. Banco Central de la República Argentina (BCRA). https://www.bcra.gob.ar/institucional/DescargaPDF/DownloadPDF.aspx?Id=1112

Cokluk, O. (2010). Logistic Regression: Concept and Application. Educational Sciences: Theory and Practice, 10(3), 1397-1407. https://eric.ed.gov/?id=EJ919857

Conti, G., Hanson, M., Inskip, H., Crozier, S., Cooper, C., y Godfrey, K. M. (2020). Beyond birthweight: The origins of human capital. IZA Discussion Paper No. 13296. https://dx.doi.org/10.2139/ssrn.3614244

Cruces, G., Glüzmann, P., y Calva, L. F. L. (2012). Economic crises, maternal and infant mortality, low birth weight and enrollment rates: evidence from Argentina’s downturns. World Development, 40(2), 303-314. https://doi.org/10.1016/j.worlddev.2011.07.014

Cuestas, E., Gómez-Flores, M. E., Charras, M. D., Peyrano, A. J., Montenegro, C., Sosa-Boye, et al. (2021). Socioeconomic inequalities in low birth weight risk before and during the COVID-19 pandemic in Argentina: a cross-sectional study. The Lancet Regional Health–Americas, 2. https://doi.org/10.1016/j.lana.2021.100049

DeLong, E. R., DeLong, D. M., y Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 837-845. https://www.jstor.org/stable/2531595

Elshibly, E. M., y Schmalisch, G. (2008). The effect of maternal anthropometric characteristics and social factors on gestational age and birth weight in Sudanese newborn infants. BMC Public Health 8, 1-7. https://doi.org/10.1186/1471-2458-8-244

Falcão, I. R., Ribeiro-Silva, R. D. C., de Almeida, M. F., Fiaccone, R. L., dos S Rocha, A., Ortelan et al. (2020). Factors associated with low birth weight at term: a population-based linkage study of the 100 million Brazilian cohort. BMC Pregnancy Childbirth 20, 536 (2020). https://doi.org/10.1186/s12884-020-03226-x

Faruk, A., Cahyono, E. S., Eliyati, N., y 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. https://doi.org/10.17509/ijost.v3i1.10799

Faulks, F., Shafiei, T., McLachlan, H., Forster, D., Mogren, I., Copnell, B., y Edvardsson, K. (2023). Perinatal outcomes of socially disadvantaged women in Australia: a population-based retrospective cohort study. BJOG: An International Journal of Obstetrics & Gynaecology, 130(11), 1380-1393. https://doi.org/10.1111/1471-0528.17501

Finch, B. K. (2003). Socioeconomic gradients and low birth-weight: Empirical and policy considerations. Health services research, 38(6p2), 1819-1842. https://doi.org/10.1111/j.1475-6773.2003.00204.x

Fondo de las Naciones Unidas para la Infancia. (2021). Encuesta Nacional de Niñas, Niños y Adolescentes (MICS) 2019-2020. https://www.unicef.org/argentina/informes/mics-2019-2020

Fondo de las Naciones Unidas para la Infancia. (2023). Low birthweight. A good start in life begins in the womb. UNICEF Data: Monitoring the situation of children and women. https://data.unicef.org/topic/nutrition/low-birthweight/

Fundación Abrazar. (s.f). Nuestros Talleres. https://fundacionabrazar.org/talleres/

Gillion, L. (2017). Birth Weight as Destiny? How Parental Investment Reinforces the Birth Weight Educational Gap. Journal of Social, Behavioral, and Health Sciences, 11(1), 6. https://doi.org/10.5590/JSBHS.2017.11.1.6

Grytten, J., Skau, I., y Sørensen, R. J. (2014). Educated mothers, healthy infants. The impact of a school reform on the birth weight of Norwegian infants 1967–2005. Social Science & Medicine, 105, 84-92. https://doi.org/10.1016/j.socscimed.2014.01.008

Hanafy, M., y Ming, R. (2021). Machine learning approaches for auto insurance big data. Risks, 9(2), 42. https://doi.org/10.3390/risks9020042

Hastie, T., Tibshirani, R., Friedman, J. H., y Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: Springer. https://doi.org/10.1007/978-0-387-84858-7

Hendricks, C. H. (1967). Delivery patterns and reproductive efficiency among groups of differing socioeconomic status and ethnic origins. American Journal of Obstetrics and Gynecology, 97(5), 608-624. https://doi.org/10.1016/0002-9378(67)90449-8

Hidalgo-Lopezosa, P., Jiménez-Ruz, A., Carmona-Torres, J. M., Hidalgo-Maestre, M., Rodríguez-Borrego, M. A., y López-Soto, P. J. (2019). Sociodemographic factors associated with preterm birth and low birth weight: A cross-sectional study. Women and Birth, 32(6), e538-e543. https://doi.org/10.1016/j.wombi.2019.03.014

Hoggatt, K. J., Flores, M., Solorio, R., Wilhelm, M., y Ritz, B. (2012). The "Latina epidemiologic paradox" revisited: the role of birthplace and acculturation in predicting infant low birth weight for Latinas in Los Angeles, CA. Journal of immigrant and minority health, 14(5), 875-884. https://doi.org/10.1007/s10903-011-9556-4

Hussain, Z., y Borah, M. D. (2020). Birth weight prediction of new born baby with application of machine learning techniques on features of mother. Journal of Statistics and Management Systems, 23(6), 1079-1091. https://doi.org/10.1080/09720510.2020.1814499

Islam Pollob, S. A., Abedin, M. M., Islam, M. T., Islam, M. M., y Maniruzzaman, M. (2022). Predicting risks of low birth weight in Bangladesh with machine learning. PloS one, 17(5), e0267190. https://doi.org/10.1371/journal.pone.0267190

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: Springer. https://doi.org/10.1007/978-1-4614-7138-7

Johnelle Sparks, P. (2009). One size does not fit all: an examination of low birthweight disparities among a diverse set of racial/ethnic groups. Maternal and child health journal, 13, 769-779. https://doi.org/10.1007/s10995-009-0476-z

Juárez, S., y Revuelta Eugercios, B. (2013). Diferencias socioeconómicas en el bajo peso al nacer: revisitando enfoques epidemiológicos. Revista Española de Investigaciones Sociológicas (REIS), 144(1), 73-96. https://doi.org/10.5477/cis/reis.144.73

Kelly, Y., Panico, L., Bartley, M., Marmot, M., Nazroo, J., y Sacker, A. (2009). Why does birthweight vary among ethnic groups in the UK? Findings from the Millennium Cohort Study. Journal of public health, 31(1), 131-137. https://doi.org/10.1093/pubmed/fdn057

Khan, N., Mozumdar, A., y Kaur, S. (2020). Determinants of low birth weight in India: An investigation from the National Family Health Survey. American Journal of Human Biology, 32(3), e23355. https://doi.org/10.1002/ajhb.23355

Khan, W., Zaki, N., Masud, M. M., Ahmad, A., Ali, L., Ali, N., & Ahmed, L. A. (2022). Infant birth weight estimation and low birth weight classification in United Arab Emirates using machine learning algorithms. Scientific reports, 12(1), 12110. https://doi.org/10.1038/s41598-022-14393-6

Kumar, S., y Gonzalez, F. (2018). Effects of health insurance on birth weight in Mexico. Health Economics, 27(8), 1149-1159. https://doi.org/10.1002/hec.3662

Kurniawati, Y., Notodiputro, K. A., y Sartono, B. (2022). Selection of variables in logistic linear mixed model with L1-penalty (Case study: Low birth weight in Indonesia). In AIP Conference

Proceedings (Vol. 2662, No. 1, p. 020034). AIP Publishing LLC. https://doi.org/10.1063/5.0110962

Lee, J. H., Shi, Z., & Gao, Z. (2022). On LASSO for predictive regression. Journal of Econometrics, 229(2), 322-349. https://doi.org/10.1016/j.jeconom.2021.02.002

Lin, M. J., y Liu, J. T. (2009). Do lower birth weight babies have lower grades? Twin fixed effect and instrumental variable method evidence from Taiwan. Social science & medicine, 68(10), 1780-1787. https://doi.org/10.1016/j.socscimed.2009.02.031

Mahmoodi, Z., Karimlou, M., Sajjadi, H., Dejman, M., Vameghi, M., y Dolatian, M. (2013). Working conditions, socioeconomic factors and low birth weight: path analysis. Iranian Red Crescent medical journal, 15(9), 836–842. https://doi.org/10.5812/ircmj.11449

Mahumud, R. A., Sultana, M., y Sarker, A. R. (2017). Distribution and Determinants of Low Birth Weight in Developing Countries. Journal of preventive medicine and public health = Yebang Uihakhoe chi, 50(1), 18–28. https://doi.org/10.3961/jpmph.16.087

Mane, D. T., Mante, J., Bakare, A. A., Gandhi, Y., Khetani, V., et al. (2024). A Systematic Review on the Applications of Machine Learning for Fetal Birth Weight Prediction. Int J Med Net, 2(2), 01-12. https://doi.org/10.21203/rs.3.rs-3440424/v1

Mansur, M., Alam, M. M., y Rayhan, M. I. (2024). Unraveling birth weight determinants: Integrating machine learning, spatial analysis, and district-level mapping. Heliyon, 10(5). https://doi.org/10.1016/j.heliyon.2024.e27341

Martínez, M. R., Ibáñez, P. C., y Campillo, J. P. (2021). La predicción del fracaso empresarial de las cooperativas españolas. Aplicación del Algoritmo Extreme Gradient Boosting. CIRIEC-España, revista de economía pública, social y cooperativa, (101), 255-288. https://doi.org/10.7203/CIRIEC-E.101.15572

Martinson, M. L., y Reichman, N. E. (2016). Socioeconomic inequalities in low birth weight in the United States, the United Kingdom, Canada, and Australia. American Journal of Public Health, 106(4), 748-754. https://doi.org/10.2105/AJPH.2015.303007

Masho, S. W., Chapman, D., y Ashby, M. (2010). The impact of paternity and marital status on low birth weight and preterm births. Marriage & Family Review, 46(4), 243-256. https://doi.org/10.1080/01494929.2010.499319

Mathew, D., y Thinakaran, K. (2025). Early prediction of Low Birth Weight (LBW) using logistic regression algorithm with regularization technique over Gradient-boosting algorithm for improved accuracy. In Hybrid and Advanced Technologies (pp. 248-253). CRC Press. ISBN 978-1-032-90254-8.

https://www.researchgate.net/publication/388512416_Early_prediction_of_Low_Birth_Weight_LBW_using_logistic_regression_algorithm_with_regularization_technique_over_Gradient-boosting_algorithm_for_improved_accuracy

Ministerio de Salud. (2024). Secretaría de Acceso a la Salud. Dirección de Estadísticas e Información de la Salud (DEIS). Estadísticas Vitales. Anuario Estadístico de la República Argentina 2022. https://www.argentina.gob.ar/sites/default/files/serie_5_nro_66_anuario_vitales_2022_3.pdf

Mishra, P. S., Sinha, D., Kumar, P., Srivastava, S., y Bawankule, R. (2021). Newborn low birth weight: do socio-economic inequality still persist in India?. BMC Pediatrics, 21, 1-12. https://doi.org/10.1186/s12887-021-02988-3

Molina, M. (2024). La paradoja del aire: Interpretar un modelo de regresión logística. Revista Electrónica AnestesiaR, 16(3). https://doi.org/10.30445/rear.v16i3.1246

Núñez Nepomuceno, D. (2025). Evaluación de técnicas de preprocesamiento para problemas de clasificación con datos desequilibrados. Universidad de Granada. https://digibug.ugr.es/handle/10481/102576

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., ... y Cheng, C. Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. Journal of clinical epidemiology, 122, 56-69. https://doi.org/10.1016/j.jclinepi.2020.03.002

Organización Mundial de la Salud (2017). Metas mundiales de nutrición 2025: documento normativo sobre bajo peso al nacer. https://www.who.int/es/publications/i/item/WHO-NMH-NHD-14.5

Padilla, Y. C., y Reichman, N. E. (2001). Low birthweight: Do unwed fathers help?. Children and Youth Services Review, 23(4-5), 427-452. https://doi.org/10.1016/S0190-7409(01)00136-0

Paneth, N. S. (1995). The problem of low birth weight. The future of children, 19-34. https://doi.org/10.2307/1602505

Patterson, J.K., Thorsten, V.R., Eggleston, B. et al. (2023). Building a predictive model of low birth weight in low- and middle-income countries: a prospective cohort study. BMC Pregnancy Childbirth 23, 600. https://doi.org/10.1186/s12884-023-05866-1

Peláez, I. M. (2016). Modelos de regresión: lineal simple y regresión logística. Revista Seden, 14, 195-214. https://www.revistaseden.org/files/14-cap%2014.pdf

Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. Risks, 7(2), 70. https://doi.org/10.3390/risks7020070

Petrou, S., Sach, T., y Davidson, L. (2001). The long-term costs of preterm birth and low birth weight: Results of a systematic review. Child: care, health and development, 27(2), 97-115. https://doi.org/10.1046/j.1365-2214.2001.00203.x

Phung, H., Bauman, A., Nguyen, T. V., Young, L., Tran, M., y Hillman, K. (2003). Risk factors for low birth weight in a socio-economically disadvantaged population: parity, marital status,

ethnicity and cigarette smoking. European Journal of Epidemiology, 18, 235-243. https://doi.org/10.1023/A:1023384213536

Pinzón-Rondón, Á. M., Gutiérrez-Pinzon, V., Madriñan-Navia, H., Amin, J., Aguilera-Otalvaro, P., y Hoyos-Martínez, A. (2015). Low birth weight and prenatal care in Colombia: a cross-sectional study. BMC pregnancy and childbirth, 15, 1-7. https://doi.org/10.1186/s12884-015-0541-0

Priya, T., Sarkar, B. K., y Sahana, S. K. (2024). Regression based machine learning models for forecasting preterm birth cases. Procedia Computer Science, 235, 830-839. https://doi.org/10.1016/j.procs.2024.04.079

Ranjbar, A., Montazeri, F., Farashah, M. V., Mehrnoush, V., Darsareh, F., y Roozbeh, N. (2023). Machine learning-based approach for predicting low birth weight. BMC Pregnancy and Childbirth, 23(1), 803. https://doi.org/10.1186/s12884-023-06128-w

Ratowiecki, J., Poletta, F. A., Giménez, L. G., Güi, J. A., Pawluk, M. S., y López Camelo, J. S. (2018). Prevalencia del bajo peso al nacer en un escenario de depresión económica en Argentina. Archivos Argentinos de Pediatría, 116(5), 322-327. http://dx.doi.org/10.5546/aap.2018.322

Rauh, V. A., Andrews, H. F., y Garfinkel, R. S. (2001). The contribution of maternal age to racial disparities in birthweight: a multilevel perspective. American journal of public health, 91(11), 1815-1824. https://doi.org/10.2105/AJPH.91.11.1815

Rendón, M. T., y Apaza, D. H. (2009). Influencia de la escolaridad materna en el peso del recién nacido en hospitales del Ministerio de Salud del Perú. Revista Médica Basadrina, 3(1), 5-8. https://doi.org/10.33326/26176068.2009.1.692

Restrepo-Méndez, M. C., Lawlor, D. A., Horta, B. L., Matijasevich, A., Santos, I. S., Menezes, A. M., et al. (2015). The association of maternal age with birthweight and gestational age: a cross-cohort comparison. Paediatric and perinatal epidemiology, 29(1), 31-40. https://doi.org/10.1111/ppe.12162

Reza, T. B., y Salma, N. (2024). Prediction and Feature selection of Low Birth Weight using Machine Learning Algorithms. https://doi.org/10.21203/rs.3.rs-3972884/v1

Rosenwaike, I. (1971). The influence of socioeconomic status on incidence of low birth weight. HSMHA Health Reports, 86(7), 641. https://pmc.ncbi.nlm.nih.gov/articles/PMC1937090/

Sardaña, M. S. (2022). Modelo predictivo de venta cruzada en productos de Vida y Salud: Random Forest vs XGBoost. https://hdl.handle.net/10016/36523

Senthilkumar, D., y Paulraj, S. (2015). Prediction of low birth weight infants and its risk factors using data mining techniques. In Proceedings of the 2015 international conference on industrial engineering and operations management (pp. 3-5). https://ieomsociety.org/ieom_2015/papers/134.pdf

Shipe, M. E., Deppen, S. A., Farjah, F., y Grogan, E. L. (2019). Developing prediction models for clinical use using logistic regression: an overview. Journal of thoracic disease, 11(Suppl 4), S574. https://doi.org/10.21037/jtd.2019.01.25

Silva, T. R. S. R. D. (2012). Nonbiological maternal risk factor for low birth weight on Latin America: a systematic review of literature with meta-analysis. Einstein (Sao Paulo), 10, 380-385. https://doi.org/10.1590/S1679-45082012000300023

Som, S., Pal, M., Adak, D. K., Gharami, A. K., y Bharati, P. (2004). Effect of socio-economic and biological variables on birth weight. http://library.isical.ac.in:8080/jspui/bitstream/10263/2925/1/Binder1.pdf

Soules, L. M. (2020). Un modelo de aprendizaje automático orientado a predecir mora crediticia sobre la base de datos públicos, abiertos y masivos: desarrollo, evaluación e implicancias prácticas para el mercado crediticio argentino. 12° Premio de Investigación Económica. "Dr. Raúl Prebisch". BCRA. https://bcra.gob.ar/Institucional/DescargaPDF/DownloadPDF.aspx?Id=908

Spencer, N., y Logan, S. (2002). Social influences on birth weight. Journal of Epidemiology & Community Health, 56(5), 326-327. https://doi.org/10.1136/jech.56.5.326

Staudt, A. (2022). Predicción de transiciones laborales de la actividad a la inactividad en el mercado laboral argentino: un enfoque de aprendizaje automático. Tesis de maestría en Economía, Universidad Nacional de La Plata. https://doi.org/10.35537/10915/139421

Supadmi, S., Kusrini, I., Fuada, N., y Laksono, A. D. (2020). The low birth weight in Indonesia: does antenatal care matter?. Children, 14(9). https://www.ijicc.net/images/Vol_14/Iss_9/14939_Kusrini_2020_E1_R.pdf

Szabó, L., y Boros, J. (2023). Socio-economic differences among low-birthweight infants in Hungary. Results of the Cohort '18–Growing Up in Hungary birth cohort study. Plos one, 18(9), e0291117. https://doi.org/10.1371/journal.pone.0291117

Szretter Noste, M. E. (2019). Regularización, Ridge y LASSO. Modelo Lineal. Instituto de Cálculo. FCEyN, UBA. https://cms.dm.uba.ar/academico/materias/1ercuat2019/modelo_lineal/ridgeclase2.pdf

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Torres-Arreola, L. P., Constantino-Casas, P., Flores-Hernández, S., Villa-Barragán, J. P., y Rendón-Macías, E. (2005). Socioeconomic factors and low birth weight in Mexico. BMC Public health, 5, 1-7. https://doi.org/10.1186/1471-2458-5-20

Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of economic perspectives, 28(2), 3-28. https://www.aeaweb.org/articles?id=10.1257/jep.28.2.3

Wulandari, R. D., Laksono, A. D., y Matahari, R. (2023). Policy to Decrease Low Birth Weight in Indonesia: Who Should Be the Target?. Nutrients, 15(2), 465. https://doi.org/10.3390/nu15020465

Zahirzada, A., y Lavangnananda, K. (2021). Implementing predictive model for low birth weight in Afghanistan. 2021 13th International Conference on Knowledge and Smart Technology (KST) (pp. 67-72). IEEE. https://doi.org/10.1109/KST51265.2021.9415792

Zaveri, A., Paul, P., Saha, J., Barman, B., y Chouhan, P. (2020). Maternal determinants of low birth weight among Indian children: Evidence from the National Family Health Survey-4, 2015-16. PLoS One, 15(12), e0244562. https://doi.org/10.1371/journal.pone.0244562

Anexo A: Variables predictoras

A.1 Características de la madre

- Edad (WB4): edad en años
- Casado (MSTATUS): dummy de si la persona está casada o no
- Nivel educativo (welevel)
- Cuidado prenatal (MN2): dummy de si la madre recibió cuidados prenatales durante el embarazo
- Cobertura seguro salud (WB18): dummy de si la madre cuenta con cobertura de seguro de salud
- Descuento jubilatorio (HT9): descuento jubilatorio en la ocupación principal
- Aporte jubilatorio (HT16): aporte jubilatorio en la ocupación principal
- Categoría ocupacional (HT8): categoría ocupacional de la ocupación principal
- Región (stratum): región de pertenencia

A.2 Características del jefe/a del hogar

- Descendencia indígena (ethnicity): dummy de descendencia o pertenencia indígena del jefe/a del hogar
- Sexo del jefe de hogar (HHSEX): representa la dummy del sexo del jefe de hogar, 1 si es varón

A.2 Características del hogar

- Material techo (HC5): material principal del techo del hogar
- Material paredes (HC6):: material principal de las paredes exterior del hogar
- Material piso (HC4): material principal del piso del hogar
- Deciles riqueza (windex10): representa pertenencia del hogar a los deciles de riqueza
- Tipo vivienda (EV2: refiere al tipo de vivienda
- Habitaciones (HC3): número de habitaciones utilizadas para dormir en el hogar
- Basural (EV5B): dummy que representa si la vivienda está localizada en o cerca de un basural permanente que quede a menos de tres cuadras
- Programa estatal (TS5BG): dummy que indica si alguien del hogar recibe algún programa municipal o provincial
- Seguro desempleo (TS5BF): dummy que indica si alguien del hogar recibe seguro de desempleo
- Inundable (EV5A): dummy que representa si la vivienda está localizada o cerca de terrenos o áreas inundables en los últimos 12 meses
- Ubicación vivienda (EV1): indica la ubicación de la vivienda

- AUH (auhhogar): dummy que indica los hogares en el que algún miembro es beneficiario de AUH
- Cloacas (EV4B): dummy que refiere a si en la cuadra de la vivienda existen cloacas
- Comedor (TS7): dummy que hace referencia a si alguien del hogar come o retira vianda regularmente en un comedor comunitario, en forma gratuita
- Alumbrado (EV3C): dummy que refiere a si en la cuadra de la vivienda existe alumbrado público
- Desagüe (EV3B): dummy que hace referencia a si en la cuadra de la vivienda existe calle con desagüe
- Calle pavimentada (EV3A): dummy que hace referencia a si en la cuadra de la vivienda existe calle pavimentada
- Subsidio electricidad (TS8A): dummy que indica si alguien del hogar subsidio, tarifa social o descuento por electricidad
- Subsidio gas (TS8B): dummy que indica si alguien del hogar subsidio, tarifa social o descuento por gas
- Compra alimentos (TS6): variable dummy que hace referencia a si el hogar recibe actualmente tarjetas, tickets, vales o bonos para compra de alimentos
- Red agua (EV4A): dummy que indica si en la cuadra de la vivienda existe red de agua
- Gas red (EV4C): variable dummy que hace referencia a si en la cuadra de la vivienda existe gas de red
- Residuos (EV3D): dummy que refiere a si en la cuadra de la vivienda existe recolección de residuos
- Electricidad red (HC8A): dummy que indica si el hogar tiene electricidad por red

Anexo B: Modelos

Desarrollo de regresión logística

La función log-likelihood para N observaciones y $\beta = \beta_0 + \sum_{p=1}^{p} X_{ip}\beta_p$ es:

$\delta(\theta) = \sum_{i=1}^{N} log\pi_{yi}(x_i;\theta)$ siendo yi una variable binaria que toma valores 0 y 1.

La función log-likelihood puede establecerse de esta forma:

$$\delta(\theta) = \sum_{i=1}^{N} (y_i log\pi_i + (1-y_i)\, log(1-\pi_i)$$

$$\delta(\theta) = \sum_{i=1}^{N} (y_i log\pi_i + log(1 - y_i) - y_i log(1 - \pi_i)$$

$$\delta(\theta) = \sum_{i=1}^{N} \{log(1 - \pi_i) + y_i log(\frac{\pi_i}{1 - \pi_i})\}$$

$$\delta(\theta) = \sum_{i=1}^{N} \{log(\frac{1}{1 + exp(\beta_0 + \sum_{p=1}^{p} x_{ip}\beta_p)}) + y_i(\beta_0 + \sum_{p=1}^{p} x_{ip}\beta_p)\}$$

$$\delta(\theta) = \sum_{i=1}^{N} \{log(\frac{1}{1 + exp(\beta_0 + \sum_{p=1}^{p} x_{ip}\beta_p)}) + y_i(\beta_0 + \sum_{p=1}^{p} x_{ip}\beta_p)\}$$

$$\delta(\theta) = \sum_{i=1}^{N} \{y_i(\beta_0 + \sum_{p=1}^{p} x_{ip}\beta_p) - log(1 + exp(\beta_0 + \sum_{p=1}^{p} x_{ip}\beta_p))\}$$

Regresión lasso - Optimización de parámetros del modelo

Para realizar la estimación se ajustan diferentes regresiones utilizando el dataset de entrenamiento. Luego, cada modelo es evaluado con la métrica AUC-ROC en el grupo de validación. De esta manera, el modelo óptimo es aquel cuyo lambda genera la mejor performance predictiva, esto es, el mayor valor de AUC-ROC para los datos de validación (Staudt, 2022).

Para este trabajo se establece una secuencia de 20 valores lambda que van desde 0.0001 hasta 10000, distribuidos logarítmicamente. El máximo valor alcanzado en términos de AUC-ROC fue de 0.61 en este modelo, con un parámetro de regularización (1/lambda) de 0.0335.

Random forest - Optimización de parámetros del modelo

n_estimators: 50

max_depth: 8

min_samples_split: 5

min_samples_leaf: 1

max_features: sqrt

bootstrap: True

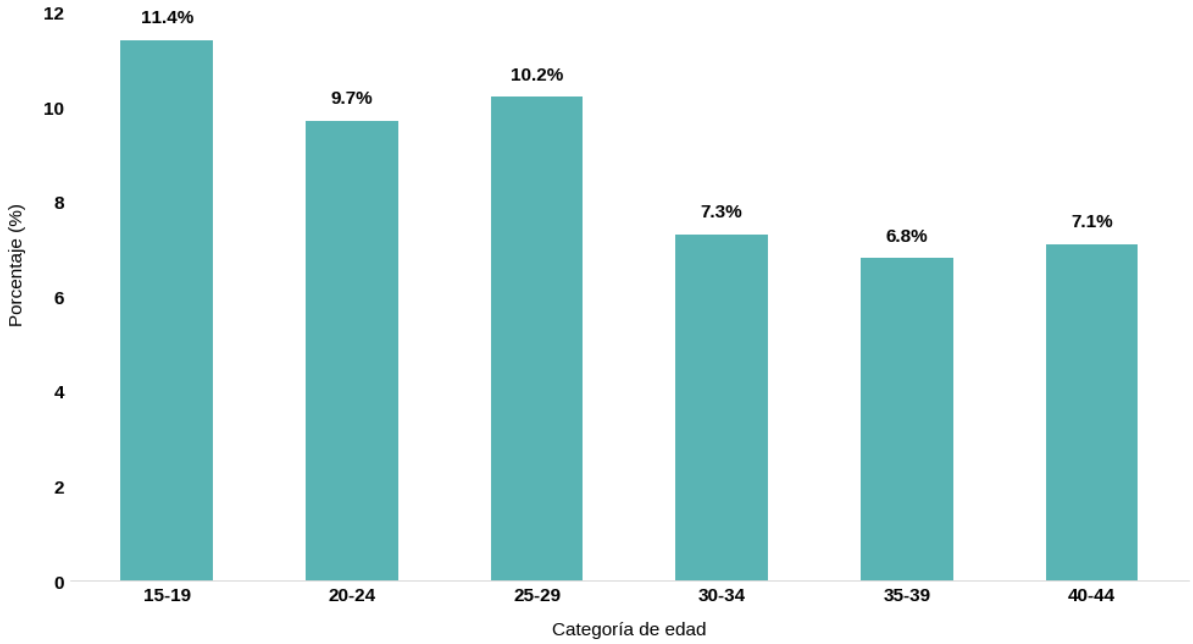
random_state: 123

Mejor AUC-ROC de random forest en conjunto de validación: 0.6970

Anexo C: Gráficos y tablas

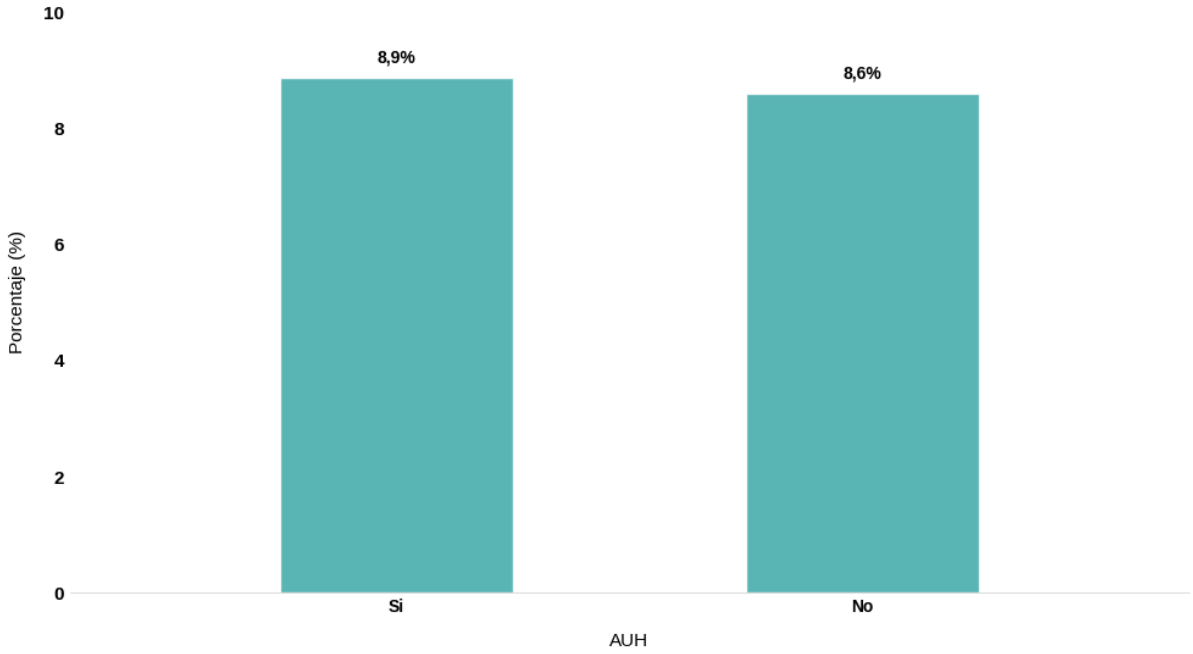
Análisis exploratorio

Figura 11: Proporción de casos de bajo peso al nacer según categoría de edad de la madre



Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF) (2021).

Figura 12: Proporción de casos de bajo peso al nacer según cobro de AUH



Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF) (2021).

Tabla 4: Salida de regresión logística con variables de la literatura

Variable	Log-Odds
cuidado_prenatal	-0.0236 (0.1104)
descendencia_indígena	0.1618* (0.0715)
cobertura_seguro_salud	-0.0944 (0.1247)
edad	-0.1656 (0.1116)
casado	-0.0741 (0.1217)
región Cuyo	0.0526 (0.1369)
región NEA	0.3173* (0.1417)
región NOA	0.0377 (0.1397)

región Pampeana	0.1109 (0.1433)
región Patagonia	0.0424 (0.1284)
nivel_educativo Secundario completo / Terciario o Universitario incompleto	-0.1100 (0.1234)
nivel_educativo Terciario completo	-0.3175* (0.1566)
deciles_riqueza_2	-0.2591 (0.1395)
deciles_riqueza_3	-0.0474 (0.1246)
deciles_riqueza_4	0.0320 (0.1071)
deciles_riqueza_5	-0.0705 (0.1263)
deciles_riqueza_6	0.0597 (0.1226)
deciles_riqueza_7	0.1073 (0.1163)
deciles_riqueza_8	0.1005 (0.1269)
deciles_riqueza_9	0.0613 (0.1335)
deciles_riqueza_10	0.0266 (0.1343)
nivel_educativo_jefe_hogar Secundario completo / Terciario o Universitario incompleto	0.1605 (0.1221)
nivel_educativo_jefe_hogar Terciario completo	0.4044** (0.1360)

AIC:		835.852
Log-Likelihood:		-393.926
Deviance:		787.852
Num.	obs.:	1350

Nota: Signif. códigos: *** p<0.001, ** p<0.01, * p<0.05

Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF) (2021).

Tabla 5: Variables seleccionadas por el método de selección hacia adelante

Variable	AUC-ROC	Ganancia AUC-ROC
Edad	0.632	0.6320
Región NEA	0.698	0.0660
Estado civil casado	0.726	0.0280
Descendencia indígena	0.743	0.0170
Nivel educativo jefe hogar medio	0.749	0.0060
4° decil de riqueza	0.757	0.0080
Región Cuyo	0.762	0.0050
Región Patagonia	0.768	0.0060
Cuidado prenatal	0.774	0.0060
10° decil de riqueza	0.774	0.0001

Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF) (2021).

Tabla 6: Variables seleccionadas por el método de selección hacia adelante utilizando la base completa

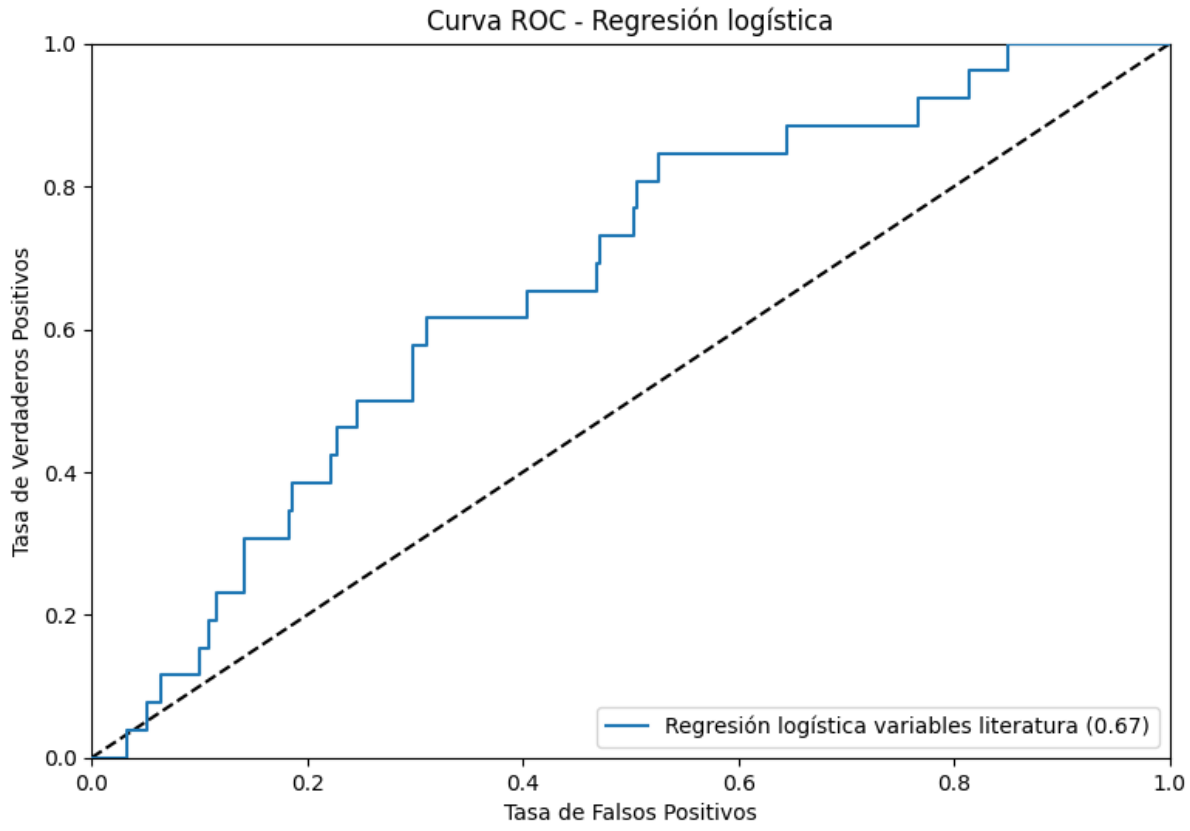
Variable	AUC-ROC	Ganancia AUC-ROC
Edad	0.632	0.6320
Región NEA	0.698	0.0660
Estado civil casado	0.726	0.0280
Subsidio gas	0.748	0.0220

Descendencia indígena	0.761	0.0130
Cloacas	0.772	0.0110
Prop. vivienda cedida	0.781	0.0080
Aporte jubilatorio	0.768	0.0060
Cuidado prenatal	0.787	0.0060
Mat. techo chapa cartón	0.793	0.0060
Ubic. vivienda otros	0.798	0.0050
Ubic. vivienda barrio social	0.805	0.0070
Región Cuyo	0.811	0.0050
Tipo vivienda casilla	0.815	0.0040
Mat. paredes madera	0.819	0.0050
Mat. techo paja/palma	0.823	0.0040
9° decil riqueza	0.826	0.0030
Residuos	0.829	0.0030
Alumbrado	0.835	0.0060
Mat. paredes hormigón	0.841	0.0060
Prop. vivienda alquila	0.845	0.0040
3° decil riqueza	0.847	0.0020
Comedor	0.848	0.0010
Región Patagonia	0.851	0.0020
Nivel ed. jefe hogar NS/NC	0.853	0.0020
Mat. piso vinilo	0.854	0.0010
4° decil riqueza	0.855	0.0010
Mat. paredes tejas	0.856	0.0010

Cuidado prenatal	0.857	0.0010
Prop. vivienda otro	0.858	0.0010
Nivel ed. jefe hogar medio	0.859	0.0020
Tipo vivienda pieza pensión	0.860	0.0002
AUH	0.860	0.0010
Mat. paredes desecho	0.861	0.0004
10° decil riqueza	0.861	0.0002
Mat. techo madera	0.861	0.0004

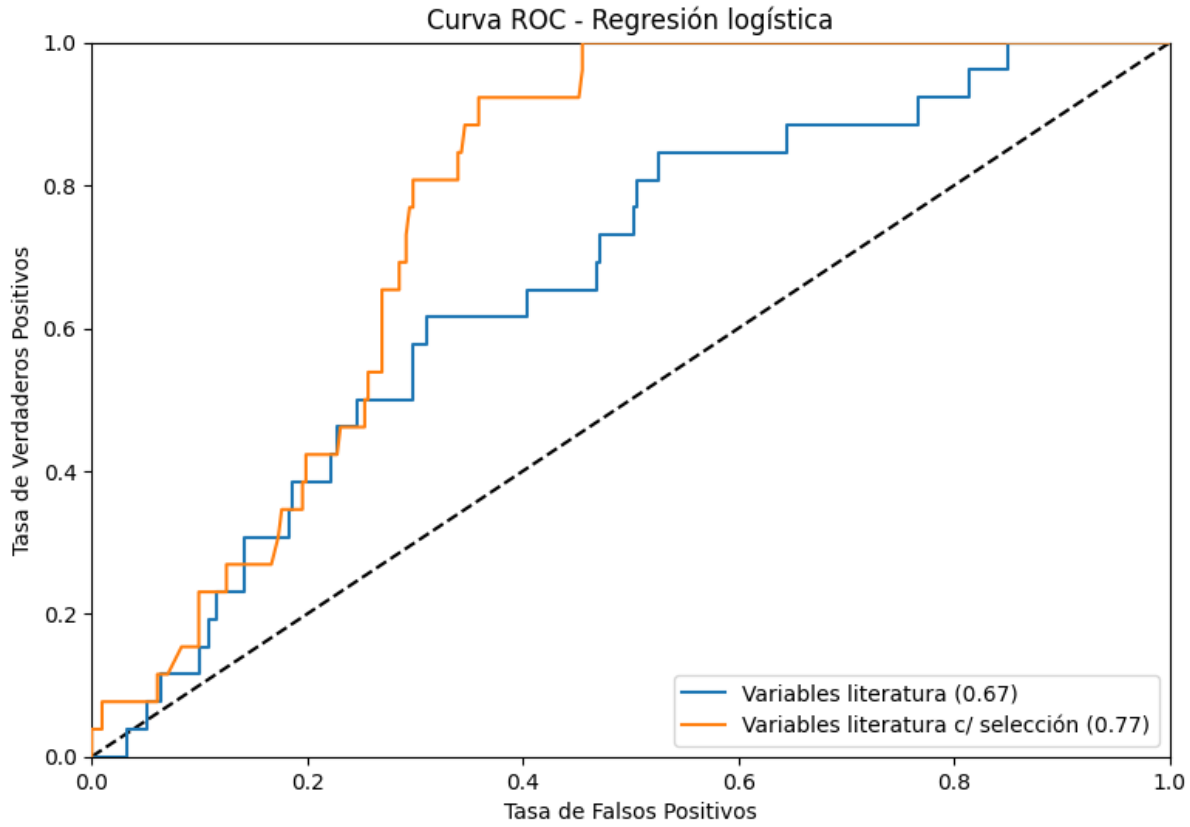
Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF) (2021).

Figura 13: Curva AUC-ROC modelo con variables literatura



Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF) (2021).

Figura 14: Curva AUC-ROC modelo con variables literatura con y sin método de selección



Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF) (2021).

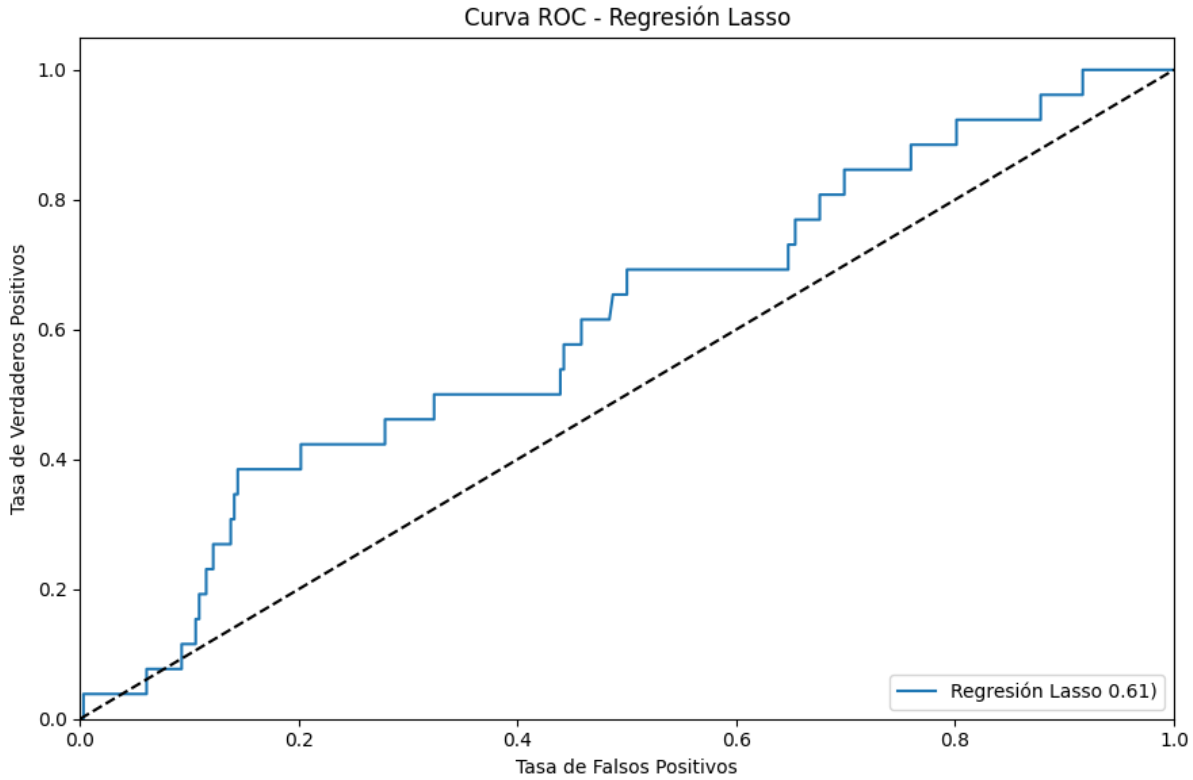
Tabla 7: Variables seleccionadas en el modelo Lasso

Variable	Coeficiente
2° decil de riqueza	-0.1427
Región NEA	0.1103
Nivel educativo jefe hogar alto	0.0937
Descendencia indígena	0.0924
Material piso madera	-0.0848
Edad	-0.0771
Acceso a electricidad por red	-0.0558

Material techo barro	-0.0405
Material techo paja/palma	0.0379
Material techo ns/nr	0.0357
Acceso a red agua	-0.0348
Subsidio gas	0.0345
Tipo de vivienda casilla	-0.0286
Propiedad vivienda prestada	0.0257
Edad jefe de hogar	-0.0218
Ayuda compra alimentos	-0.0090
Material techo baldosa cerámica	0.0031
Material techo tablones de madera	-0.0006

Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF) (2021).

Figura 15: Curva AUC-ROC Modelo Lasso



Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF) (2021).

Tabla 8: Importancia de variables para el modelo Random Forest

Variable	Importancia
Edad jefe de hogar	0.099414
Edad	0.061603
Cantidad de habitaciones	0.040999
Descendencia indígena	0.033358
Categoría ocupacional	0.027257
Desagüe	0.024191
Subsidio gas	0.024038
7° decil de riqueza	0.022315
Nivel educativo jefe hogar medio	0.019648
Región NEA	0.019605
Red de agua	0.019534
Ubicación vivienda otros	0.018942

AUH	0.018115
Cloacas	0.016786
Inundable	0.016212

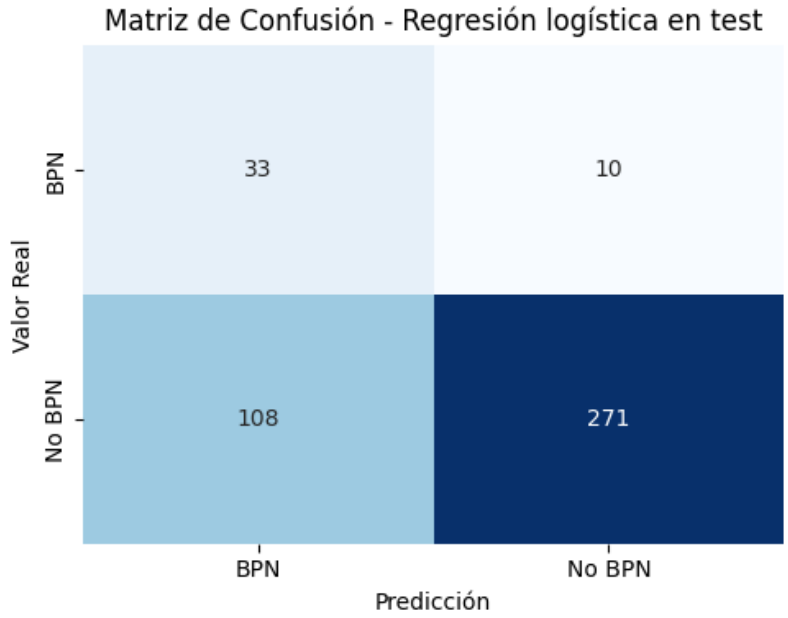
Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF) (2021).

Tabla 9: Métricas de evaluación empleadas

Métrica de evaluación	Definición	Fórmula de cálculo
Precisión	Casos positivos que son realmente positivos. Implica del total de casos que el modelo predice como bajo peso, qué porcentaje realmente lo tiene.	Fórmula: VP / (VP + FP)
Sensibilidad/Recall	Casos positivos reales que fueron correctamente identificados. Del total de casos que realmente tienen bajo peso, qué porcentaje detecta el modelo.	Fórmula: VP / (VP + FN)
Especificidad	Casos negativos reales que fueron correctamente identificados. Implica del total de casos de no bajo peso al nacer, qué porcentaje clasifica correctamente el modelo.	Fórmula: VN / (VN + FP)
F1-score	Métrica que balancea precisión y recall. Es de utilidad cuando las clases están desequilibradas.	Fórmula: 2 * (Precisión * Recall) / (Precisión + Recall)
Exactitud (Accuracy)	Predicciones correctas sobre el total. Porcentaje de casos correctamente clasificados, pero puede no ser correcto cuando las clases están desequilibradas.	Fórmula: (VP + VN) / (VP + VN + FP + FN)
Macro Average	Promedio no ponderado de las métricas para cada clase. Trata todas las clases por igual, independientemente de su frecuencia.	-
Weighted Average	Promedio ponderado de las métricas según la frecuencia de cada clase. Da más importancia a las métricas de las clases más frecuentes.	-
Umbral de Decisión	Valor de probabilidad a partir del cual se clasifica un caso como positivo. Determina el equilibrio entre sensibilidad y especificidad. Umbrales más bajos aumentan la sensibilidad a costa de más falsos positivos.	-

Fuente: Elaboración propia

Figura 16: Matriz de confusión modelo logit sobre conjunto de prueba



Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF) (2021).

Tabla 10: Métricas de evaluación del modelo logit sobre el conjunto de prueba

Métrica de evaluación	Resultado
Recall (sensibilidad)	0.76
Precisión	0.23
Especificidad	0.71
Exactitud	0.72
F1-score	0.35
Macro average	0.59
Weighted average	0.77

Fuente: Elaboración propia a partir del Fondo de las Naciones Unidas para la Infancia (UNICEF) (2021).