UNIVERSIDAD NACIONAL DE SAN MARTIN

RCSI
https://revistas.unsm.edu.pe/index.php/rcsi
e-ISSN: 2709-992X

Artículo Original
Original Article
Jul-Dic, 2026

Predicción e interpretación de factores asociados al bajo peso al nacer mediante aprendizaje automático con datos del MINSA, Perú

Prediction and interpretation of factors associated with low birth weight using machine learning with MINSA data, Peru

Andrade-Girón, Daniel[1]
Neri-Ayala, Abrahan[1]
Aguilar-Luna-Victoria, Miguel[1]
Oscuvicla-Tapia, Elsa[1*]
Peña, Américo[1]
Cuevas-Huari, Edgardo[1]

[1]Universidad Nacional José Faustino Sánchez Carrión, Huacho, Perú

Recibido: 11 Dec. 2025 | **Aceptado:** 10 Jun. 2026 | **Publicado:** 20 Jul. 2026

Autor de correspondencia*: eoscuvilca@unjfsc.edu.pe

Cómo citar este artículo: Andrade-Girón, D., Oscuvicla-Tapia, E., Neri-Ayala, A., Peña, A., Aguilar-Luna-Victoria, M. & Cuevas-Huari, E. (2026). Prediction and interpretation of factors associated with low birth weight using machine learning with MINSA data, Peru. *Revista Científica de Sistemas e Informática, 6*(2), e1388. https://doi.org/10.51252/rcsi.v6i2.1388

RESUMEN

El bajo peso al nacer constituye un problema relevante de salud pública en el Perú. El objetivo fue desarrollar, evaluar e interpretar modelos de aprendizaje automático para predecirlo mediante registros nacionales del Ministerio de Salud. Se realizó un estudio observacional y retrospectivo con 4 873 146 nacimientos registrados entre 2015 y 2025. Se utilizaron 16 predictores maternos, obstétricos, neonatales, territoriales y asistenciales. CatBoost, LightGBM y XGBoost fueron comparados mediante Average Precision y diez particiones estratificadas de 2023. El modelo seleccionado fue entrenado con datos de 2015-2022, calibrado con 2023 y evaluado temporalmente en 2024 y 2025. CatBoost obtuvo el mejor desempeño, con Average Precision de 0,680 y ROC-AUC de 0,910 en 2023; en 2024 y 2025 alcanzó ROC-AUC de 0,913 y 0,915, respectivamente. SHAP identificó la duración gestacional como el predictor más influyente. Su eliminación redujo sustancialmente el rendimiento. Los resultados muestran que los registros del MINSA permiten desarrollar modelos interpretables y temporalmente estables para apoyar la vigilancia perinatal.

Palabras clave: aprendizaje automático; bajo peso al nacer; CatBoost; explicabilidad; registros de salud; SHAP

ABSTRACT

Low birth weight is a significant public health issue in Peru. The objective was to develop, evaluate, and interpret machine learning models to predict it using national records from the Ministry of Health. A retrospective observational study was conducted involving 4,873,146 births registered between 2015 and 2025. Sixteen predictors—covering maternal, obstetric, neonatal, territorial, and healthcare-related factors—were used. CatBoost, LightGBM, and XGBoost were compared using Average Precision and ten stratified partitions from 2023. The selected model was trained on data from 2015–2022, calibrated using 2023 data, and temporally evaluated on data from 2024 and 2025. CatBoost achieved the best performance, with an Average Precision of 0.680 and an ROC-AUC of 0.910 in 2023; in 2024 and 2025, it achieved ROC-AUC values of 0.913 and 0.915, respectively. SHAP analysis identified gestational duration as the most influential predictor; its removal substantially reduced performance. The results demonstrate that MINSA records enable the development of interpretable and temporally stable models to support perinatal surveillance.

Keywords: machine learning; low birth weight; CatBoost; explainability; health records; SHAP

1. INTRODUCCIÓN

El bajo peso al nacer continúa siendo un problema relevante de salud pública por su relación con la supervivencia, el crecimiento y el desarrollo posterior del recién nacido. La Organización Mundial de la Salud define esta condición como un peso inferior a 2 500 g al momento del nacimiento, independientemente de la edad gestacional. A escala mundial, se estimó que, en 2020 alrededor de 19,8 millones de recién nacidos, equivalentes al 14,7 % de los nacimientos, presentaron bajo peso, mientras que el progreso hacia su reducción permaneció por debajo de las metas internacionales previstas. Además de incrementar la vulnerabilidad durante el periodo neonatal, esta condición se ha relacionado con mayor morbilidad, alteraciones del crecimiento y desarrollo y un mayor riesgo de enfermedades crónicas durante etapas posteriores de la vida (WHO & UNICEF, 2023).

En el Perú, los registros nacionales han evidenciado que el bajo peso al nacer presenta importantes diferencias temporales, geográficas y sociales. Carrillo-Larco et al. (2021), a partir de 2 927 761 nacimientos registrados entre 2012 y 2019, observaron que su prevalencia nacional disminuyó de 6,9 % a 6,2 % durante dicho periodo; sin embargo, las diferencias subnacionales fueron considerablemente mayores. En 2019, por ejemplo, dentro del departamento de Áncash se registraron prevalencias de 1,0 % en Huarmey y 11,4 % en Antonio Raymondi. Asimismo, los indicadores de peso al nacer presentaron peores resultados en zonas altoandinas y entre madres con menor nivel educativo. Estudios posteriores con 3 841 531 nacimientos entre 2012 y 2021 confirmaron la persistencia de desigualdades geográficas y socioeconómicas en los fenotipos neonatales vulnerables del país (Cajachagua-Torres et al., 2024).

La disponibilidad de registros administrativos de nacimientos representa una oportunidad para ampliar este tipo de análisis. El Certificado de Nacido Vivo del Ministerio de Salud del Perú reúne información individual relacionada con las características del nacimiento, antecedentes maternos, duración del embarazo, tipo y condición del parto, lugar de residencia y establecimiento donde se produjo la atención. Su integración con fuentes complementarias, como el Registro Nacional de Instituciones Prestadoras de Servicios de Salud y los códigos oficiales de ubicación geográfica, permite incorporar información territorial e institucional a gran escala. Estudios previos han demostrado la utilidad del registro nacional peruano para describir patrones de peso al nacer y desigualdades subnacionales; sin embargo, su utilización se ha concentrado principalmente en análisis epidemiológicos descriptivos, lo que abre la posibilidad de avanzar hacia enfoques predictivos capaces de aprovechar simultáneamente múltiples características individuales y contextuales.

La incorporación del aprendizaje automático en el campo de la salud se ha extendido hacia diferentes aplicaciones relacionadas con la prevención, la promoción sanitaria y el análisis de factores que influyen en los resultados de salud, lo que evidencia el creciente interés por estas técnicas dentro de la investigación biomédica (Dominguez Miranda & Rodriguez Aguilar, 2024). En este contexto, el aprendizaje automático ofrece herramientas para reconocer relaciones complejas y no lineales entre múltiples características maternas, obstétricas, neonatales y asistenciales. En Brasil, Victor et al. (2025) compararon Random Forest, XGBoost, CatBoost y LightGBM para predecir bajo peso al nacer a partir de una cohorte de gestantes, incorporando además valores de Shapley para interpretar la contribución de las variables. En China, Wu et al. (2025) evaluaron regresión logística y cuatro algoritmos de aprendizaje automático en 10 227 nacimientos, encontrando que la edad gestacional, la edad materna y diferentes características clínicas contribuyeron de manera importante a las predicciones; asimismo, utilizaron SHAP para explicar el comportamiento del modelo. Estos antecedentes muestran que los métodos de aprendizaje automático pueden complementar los enfoques estadísticos convencionales al capturar interacciones entre predictores y generar estimaciones individualizadas del riesgo.

No obstante, el rendimiento predictivo por sí solo resulta insuficiente cuando los modelos se aplican a información sanitaria. Las probabilidades estimadas deben mantener correspondencia con la frecuencia observada del evento, dado que un algoritmo puede presentar una adecuada capacidad discriminativa y, al

mismo tiempo, producir probabilidades deficientemente calibradas (Van Calster et al., 2019). Del mismo modo, los modelos basados en ensambles de árboles pueden alcanzar un elevado desempeño a costa de una menor transparencia sobre el proceso que conduce a cada predicción. Los valores SHAP permiten abordar parcialmente esta limitación al cuantificar la contribución de cada característica sobre las predicciones individuales y globales del modelo (Lundberg & Lee, 2017). En estudios desarrollados con registros longitudinales, resulta además relevante comprobar que el rendimiento se mantenga en periodos posteriores a aquellos utilizados durante el desarrollo, en lugar de depender únicamente de particiones aleatorias de una misma cohorte.

A pesar de estos avances, en el Perú los estudios nacionales sobre bajo peso al nacer han estado dirigidos principalmente a describir tendencias, diferencias territoriales y fenotipos neonatales vulnerables, mientras que permanece menos explorada la utilización de registros nacionales de nacimientos para desarrollar modelos predictivos interpretables y evaluar su comportamiento en periodos temporales posteriores. En respuesta a esta brecha, el objetivo del presente estudio fue desarrollar, comparar e interpretar modelos de aprendizaje automático para predecir el bajo peso al nacer utilizando registros del Ministerio de Salud del Perú. Para ello, se integraron registros nacionales correspondientes al periodo 2015-2025, se compararon CatBoost, LightGBM y XGBoost, se evaluó estadísticamente su desempeño y se seleccionó un modelo definitivo que posteriormente fue calibrado y validado temporalmente. Asimismo, se aplicaron valores SHAP para identificar las características con mayor contribución a las predicciones. El enfoque propuesto busca aportar evidencia reproducible para el análisis del bajo peso al nacer a gran escala y explorar el potencial de los registros administrativos como fuente para el desarrollo de herramientas predictivas en salud materno-neonatal.

2. MATERIALES Y MÉTODOS

2.1 Diseño y alcance del estudio

La investigación fue de tipo aplicada, con enfoque cuantitativo, diseño observacional, no experimental y retrospectivo, orientada al desarrollo, comparación e interpretación de modelos de aprendizaje automático para la predicción del bajo peso al nacer. La unidad de análisis correspondió a cada registro individual de nacimiento contenido en el Certificado de Nacido Vivo del Ministerio de Salud del Perú (CNV-MINSA).

El estudio comprendió registros correspondientes al periodo 2015-2025 y adoptó una estrategia de validación temporal. Los datos de 2015 a 2022 se destinaron al desarrollo histórico de los modelos; los registros de 2023 se utilizaron para la comparación mediante particiones estratificadas y, posteriormente, para la calibración y definición del punto operativo del modelo seleccionado; 2024 constituyó el periodo de evaluación temporal independiente; y los registros disponibles de 2025 se reservaron para examinar la estabilidad posterior del desempeño.

Esta separación cronológica permitió evaluar la capacidad de generalización del modelo frente a observaciones posteriores a las utilizadas durante su desarrollo, reduciendo el riesgo de obtener estimaciones excesivamente optimistas derivadas de particiones exclusivamente aleatorias.

2.2 Procedimiento metodológico

El procedimiento comprendió la auditoría estructural y semántica de las fuentes, la corrección de inconsistencias del archivo principal, la normalización de variables y claves de integración, la construcción de la variable objetivo y la vinculación con información territorial y de establecimientos de salud.

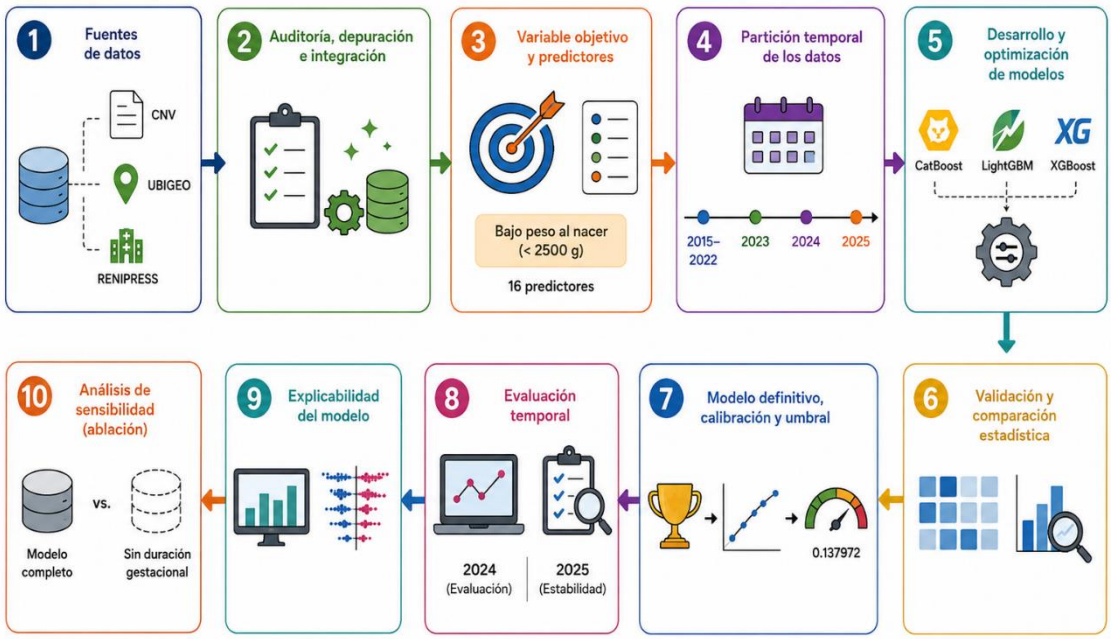

Figura 1. Flujograma metodológico del desarrollo y evaluación del modelo predictivo de bajo peso al nacer

Posteriormente, se realizó la partición cronológica de los registros, la preparación de los predictores, la optimización de hiperparámetros y la comparación de tres algoritmos de gradient boosting: CatBoost, LightGBM y XGBoost. Los tres modelos fueron evaluados utilizando el mismo conjunto de predictores, las mismas particiones de validación y una métrica principal común.

La selección del algoritmo se complementó mediante procedimientos estadísticos pareados. Finalmente, el modelo seleccionado se reentrenó con el histórico completo de 2015-2022, sus probabilidades fueron calibradas con los registros de 2023 y se estableció un umbral operativo basado en una sensibilidad mínima predefinida. El modelo permaneció sin modificaciones durante su evaluación en 2024 y el análisis de estabilidad en 2025.

El procedimiento se complementó con análisis de explicabilidad mediante valores SHAP, evaluación de calibración, estimación de intervalos de confianza mediante bootstrap y un análisis de sensibilidad por ablación de la duración gestacional.

Fuentes de datos y entorno de análisis

La fuente principal estuvo constituida por los registros del Certificado de Nacido Vivo del Ministerio de Salud del Perú, contenidos en el archivo con fecha de corte al 30 de noviembre de 2025. La base inicial presentó un tamaño aproximado de 793,79 MB y estuvo conformada por 4 874 510 registros.

Como fuentes complementarias se utilizaron el diccionario de datos del CNV, el catálogo de Ubicaciones Geográficas del Instituto Nacional de Estadística e Informática y el Registro Nacional de Instituciones Prestadoras de Servicios de Salud (RENIPRESS). El catálogo de UBIGEO permitió incorporar información de departamento, provincia y distrito, mientras que RENIPRESS aportó información relacionada con la institución, categoría y ubicación geográfica del establecimiento de salud.

El procesamiento se realizó en Google Colaboratory mediante Python. Para la manipulación e integración de los datos se utilizaron pandas y NumPy; para el preprocesamiento, particiones, métricas y calibración, scikit-learn; para la optimización de hiperparámetros, Optuna; para el aprendizaje supervisado, CatBoost, LightGBM y XGBoost; para el análisis estadístico, SciPy y Statsmodels; y para la representación gráfica, Matplotlib y Seaborn.

Los algoritmos evaluados corresponden a implementaciones de gradient boosting ampliamente utilizadas en problemas con datos tabulares. XGBoost incorpora regularización y optimizaciones para la construcción

eficiente de árboles (Chen & Guestrin, 2016); LightGBM introduce estrategias orientadas a mejorar la eficiencia computacional del entrenamiento (Ke et al., 2017); que CatBoost incorpora procedimientos específicos para el tratamiento de variables categóricas y la reducción del sesgo durante su codificación (Prokhorenkova et al., 2019).

Auditoría, depuración e integración de los datos

Inicialmente se realizó una auditoría estructural para identificar la codificación, el delimitador, las dimensiones, las columnas disponibles, los valores especiales y la consistencia general de los archivos. Los archivos CSV fueron procesados utilizando codificación UTF-8-SIG y punto y coma como delimitador.

Durante esta etapa se detectaron separadores adicionales dentro del campo correspondiente a la descripción de la ocupación materna, por lo que estos registros fueron reparados reconstruyendo la estructura original de las columnas antes de continuar con el procesamiento.

Posteriormente, se efectuó una auditoría semántica de las variables numéricas, categóricas, temporales y de identificación. Se examinaron rangos, categorías, valores especiales, cobertura de integración con UBIGEO y RENIPRESS y posibles registros duplicados.

La depuración definitiva se ejecutó por fragmentos de 200 000 registros para reducir el consumo de memoria. Se consideraron consistentes los pesos al nacer comprendidos entre 300 y 6 000 g, las edades maternas entre 10 y 59 años y las duraciones del embarazo entre 20 y 45 semanas.

Los registros cuyo peso no cumplió los criterios establecidos fueron excluidos, debido a que esta variable se utilizó directamente para definir el resultado de interés. En cambio, los valores de edad materna y duración gestacional situados fuera de sus respectivos rangos fueron transformados en valores ausentes y tratados posteriormente durante el modelado.

De los 4 874 510 registros iniciales, se excluyeron 1 364 por inconsistencias en el peso al nacer. De estos, 1 315 presentaron valores no positivos, dos registraron pesos positivos inferiores a 300 g y 47 mostraron valores superiores a 6 000 g. Como resultado, la base analítica quedó conformada por 4 873 146 nacimientos.

Las variables categóricas fueron normalizadas mediante eliminación de espacios adicionales, conversión a mayúsculas y unificación de diferentes representaciones de ausencia. Los códigos UBIGEO fueron estandarizados a seis dígitos. Para la integración con RENIPRESS se normalizaron los códigos IPRESS antes de efectuar la vinculación con la base principal.

Los conteos relacionados con antecedentes maternos fueron transformados a formato numérico. En las variables correspondientes al número de embarazos y número de hijos vivos, la categoría “5 o más” fue representada mediante el valor 5. En otras variables de antecedentes maternos, la categoría “ninguno” fue representada por cero y “11 o más” mediante el valor 11.

La ocupación materna fue agrupada determinísticamente en categorías ocupacionales mediante reglas textuales previamente definidas.

Se conservaron 974 registros potencialmente duplicados debido a que la fuente no disponía de un identificador individual que permitiera establecer con certeza que correspondían al mismo nacimiento. Asimismo, la variable relacionada con el número de hijos fallecidos de la madre fue excluida debido a que aproximadamente el 96,97 % de sus registros correspondía a valores ignorados.

Definición de la variable objetivo y predictores

El bajo peso al nacer se definió como un peso inferior a 2 500 g, criterio utilizado internacionalmente para identificar este resultado neonatal (Blencowe et al., 2019).

La variable objetivo se codificó con valor 1 para los nacimientos cuyo peso se encontraba entre 300 y menos de 2 500 g y con valor 0 para aquellos con peso entre 2 500 y 6 000 g.

La variable correspondiente al peso del nacido se utilizó exclusivamente para construir el resultado y posteriormente fue eliminada del conjunto de predictores. Asimismo, la talla del nacido fue excluida debido a su proximidad directa con el resultado neonatal y al riesgo de introducir información excesivamente relacionada con la variable que se pretendía predecir.

El experimento comparativo utilizó un conjunto común de 16 predictores: duración del embarazo, categoría de la IPRESS, tipo de parto, distrito de la IPRESS, provincia de residencia, provincia de la IPRESS, departamento de la IPRESS, nivel de instrucción materna, número de hijos vivos, financiador del parto, condición del parto, número de embarazos maternos, institución de la IPRESS, profesional o personal que atendió el parto, distrito de residencia y sexo del nacido.

La duración del embarazo, el número de hijos vivos y el número de embarazos maternos fueron tratados como variables numéricas; los 13 predictores restantes fueron procesados como variables categóricas. El mismo conjunto fue utilizado en los tres algoritmos con el propósito de mantener condiciones comparables durante la evaluación experimental.

Partición temporal y preparación para el modelado

Después de la depuración, los registros fueron separados cronológicamente en cuatro conjuntos. El periodo 2015-2022 reunió 3 726 200 nacimientos, de los cuales 233 235 correspondieron a bajo peso al nacer, equivalente a una prevalencia de 6,2593 %. El año 2023 comprendió 410 422 registros, con una prevalencia de 7,0013 %; 2024 incluyó 390 047 registros, con una prevalencia de 6,9020 %; y 2025 estuvo conformado por 346 477 registros disponibles hasta la fecha de corte, con una prevalencia de 6,7699 %.

Para reducir el costo computacional durante la comparación inicial de los algoritmos, se obtuvo del histórico 2015-2022 una muestra de 600 000 observaciones. El muestreo fue estratificado simultáneamente por año y clase de la variable objetivo, utilizando una semilla aleatoria de 42, con el propósito de conservar la estructura temporal y la proporción de casos de bajo peso.

Las variables numéricas faltantes fueron imputadas mediante la mediana calculada exclusivamente con los registros destinados al entrenamiento en cada etapa. Las categorías ausentes fueron representadas mediante una categoría explícita de valor desconocido. Los parámetros obtenidos con los datos de entrenamiento fueron posteriormente aplicados a los conjuntos de validación sin reajuste.

Debido al desbalance existente entre las clases, se incorporó ponderación durante el entrenamiento. El peso asignado a los nacimientos con bajo peso se calculó a partir de la relación entre el número de observaciones de la clase mayoritaria y el número de observaciones de la clase minoritaria. Este procedimiento permitió otorgar mayor importancia a los errores cometidos sobre los casos de bajo peso sin modificar artificialmente la distribución de los datos mediante sobremuestreo.

Desarrollo y optimización de los modelos

Se compararon CatBoost, LightGBM y XGBoost, tres algoritmos basados en gradient boosting con árboles de decisión. En términos generales, estos métodos construyen secuencialmente múltiples árboles, de manera que cada nueva iteración busca corregir parte de los errores producidos por los árboles anteriores.

CatBoost incorpora procedimientos específicos para el tratamiento de variables categóricas y ordered boosting (Prokhorenkova et al., 2019); LightGBM introduce estrategias destinadas a aumentar la eficiencia computacional del aprendizaje (Ke et al., 2017); y XGBoost emplea regularización y diferentes optimizaciones dirigidas al entrenamiento eficiente de árboles (Chen & Guestrin, 2016).

La optimización de hiperparámetros se realizó mediante Optuna, utilizando el estimador Tree-structured Parzen Estimator y 15 ensayos por algoritmo. Para esta etapa, los registros históricos de la muestra fueron

separados en entrenamiento correspondiente a 2015-2021 y validación correspondiente a 2022. En cada ensayo se utilizaron como máximo 150 000 observaciones para entrenamiento y 50 000 para validación, obtenidas mediante muestreo estratificado.

Para CatBoost se exploraron entre 300 y 650 iteraciones, profundidades entre 5 y 7, tasas de aprendizaje entre 0,035 y 0,09 y regularización L2 entre 3 y 10. Se mantuvieron constantes otros parámetros relacionados con el tratamiento de variables categóricas, el tipo de boosting, el procedimiento de bootstrap y la proporción de submuestreo.

Para LightGBM se evaluaron tasas de aprendizaje entre 0,02 y 0,08, entre 31 y 127 hojas, profundidades entre 5 y 9, un mínimo de observaciones por hoja entre 30 y 100 y regularización L2 entre 1 y 10. El número máximo de estimadores se fijó en 1 500 y la cantidad efectiva se determinó mediante parada temprana.

Para XGBoost se evaluaron tasas de aprendizaje entre 0,02 y 0,08, profundidades entre 5 y 8, pesos mínimos de nodo hijo entre 2 y 10, regularización L2 entre 1 y 10 y valores de gamma entre 0 y 1. Se estableció un máximo de 1 800 estimadores, junto con proporciones de submuestreo de observaciones y variables de 0,90.

La optimización buscó maximizar Average Precision sobre la clase correspondiente al bajo peso al nacer. Después de los 15 ensayos se realizó una confirmación de los mejores hiperparámetros utilizando los registros disponibles de 2015-2021 y 2022 dentro de la muestra histórica. La cantidad definitiva de árboles se determinó mediante parada temprana utilizando 2022 como conjunto de validación y posteriormente permaneció fija durante la comparación entre algoritmos.

Validación cruzada y comparación estadística

La comparación de los algoritmos se realizó mediante diez particiones estratificadas de los registros correspondientes a 2023. Las asignaciones de cada observación a los folds fueron generadas una sola vez, utilizando barajado y semilla aleatoria 42. Las mismas particiones fueron reutilizadas por CatBoost, LightGBM y XGBoost con el propósito de garantizar una comparación pareada.

En cada iteración, el conjunto de entrenamiento estuvo constituido por los registros históricos de 2015-2022 y nueve décimas partes de los datos de 2023, mientras que la décima parte restante de 2023 funcionó como conjunto de validación. De esta manera, cada observación de 2023 obtuvo una única predicción fuera de fold.

La métrica principal fue Average Precision, debido al desbalance entre las clases y al interés por evaluar específicamente la capacidad de identificación de nacimientos con bajo peso. Las métricas basadas en precisión y sensibilidad resultan especialmente informativas cuando la frecuencia de la clase positiva es reducida (Saito & Rehmsmeier, 2015).

Como métricas complementarias se calcularon ROC-AUC, pérdida logarítmica, Brier score, precisión, sensibilidad, especificidad, F1-score, exactitud balanceada, coeficiente de correlación de Matthews y valor predictivo negativo.

Para cada algoritmo se resumieron los resultados obtenidos en los diez folds mediante media, desviación estándar, mediana, rango intercuartílico, valores mínimo y máximo e intervalo de confianza del 95 %.

La distribución de Average Precision se examinó mediante la prueba de Shapiro-Wilk. La comparación global entre los tres algoritmos se efectuó mediante la prueba no paramétrica de Friedman y el tamaño del efecto se estimó mediante el coeficiente W de Kendall.

Cuando se requirieron comparaciones pareadas posteriores, se aplicó la prueba de rangos con signo de Wilcoxon. Los valores de significancia fueron ajustados mediante el procedimiento de Holm para controlar el efecto de las comparaciones múltiples, utilizando un nivel de significancia de 0,05. Estos procedimientos

son utilizados en estudios comparativos de algoritmos evaluados bajo particiones comunes (Demšar, 2006; García et al., 2010).

Debido a que los diez resultados procedieron de particiones de una misma cohorte temporal y no de diez conjuntos de datos completamente independientes, las pruebas estadísticas fueron interpretadas como evidencia comparativa complementaria dentro del diseño experimental y no como sustituto de la evaluación temporal independiente.

Modelo definitivo, calibración y selección del umbral

Después de la comparación experimental y estadística se seleccionó CatBoost como algoritmo definitivo. El modelo fue reentrenado utilizando los 3 726 200 registros completos correspondientes al periodo 2015-2022. Esta etapa utilizó nuevamente el conjunto histórico completo y no la muestra de 600 000 observaciones empleada durante el desarrollo inicial.

Los hiperparámetros quedaron establecidos antes de la evaluación sobre los periodos temporales posteriores. El modelo definitivo utilizó 630 iteraciones, profundidad de 7, tasa de aprendizaje de 0,0422639, regularización L2 de 6,59964, fuerza aleatoria de 1, 64 puntos de discretización, complejidad máxima de interacción categórica de 1, bootstrap Bernoulli y una proporción de submuestreo de 0,80.

Las variables numéricas ausentes fueron imputadas mediante medianas obtenidas exclusivamente a partir de los registros de 2015-2022. Los valores resultantes fueron 39 semanas para duración del embarazo, dos para el número de hijos vivos y dos para el número de embarazos maternos.

El peso asignado a la clase positiva fue recalculado utilizando la totalidad del conjunto histórico y alcanzó un valor de 14,9762.

Las probabilidades producidas por CatBoost fueron posteriormente calibradas utilizando los 410 422 registros correspondientes a 2023 mediante escalamiento de Platt. Este procedimiento ajustó una regresión logística sobre las probabilidades transformadas del modelo con el propósito de mejorar la correspondencia entre las probabilidades estimadas y la frecuencia observada del evento.

La calibración resulta especialmente importante cuando las probabilidades producidas por un modelo se pretenden interpretar como medidas de riesgo y no únicamente como valores utilizados para ordenar observaciones (Niculescu-Mizil & Caruana, 2005; Van Calster et al., 2019).

El umbral de clasificación tampoco se estableció convencionalmente en 0,50. Debido al interés en reducir los falsos negativos, se definió una sensibilidad mínima del 70 %. Entre todos los puntos de corte que cumplieron este criterio utilizando los registros de 2023, se seleccionó aquel que presentó la mayor especificidad. El procedimiento produjo un umbral definitivo de 0,137972.

Tanto el calibrador como el umbral operativo fueron determinados exclusivamente con información de 2023 y posteriormente se aplicaron sin reajuste a los registros correspondientes a 2024 y 2025.

Evaluación temporal y estabilidad del modelo

El desempeño definitivo se evaluó inicialmente sobre los registros correspondientes a 2024. Estos datos no participaron en la optimización de hiperparámetros, el entrenamiento definitivo, la calibración ni la selección del umbral.

Sobre este conjunto se calcularon Average Precision, ROC-AUC, pérdida logarítmica y Brier score como métricas independientes del punto de corte. Utilizando el umbral previamente establecido se calcularon además precisión, sensibilidad, especificidad, F1-score, exactitud balanceada, coeficiente de correlación de Matthews, valor predictivo negativo y matriz de confusión.

Para cuantificar la incertidumbre asociada a las métricas se estimaron intervalos de confianza del 95 % mediante 300 remuestreos bootstrap estratificados. En cada repetición se conservaron las proporciones de las clases y se recalcularon las métricas correspondientes.

Posteriormente, el mismo modelo, el mismo procedimiento de calibración y el mismo umbral fueron aplicados sin ningún reajuste a los registros correspondientes a 2025 disponibles hasta el 30 de noviembre. Esta segunda evaluación permitió examinar la estabilidad temporal de la discriminación, calibración y desempeño de clasificación ante un periodo posterior al conjunto principal de prueba.

Explicabilidad del modelo

La interpretación del modelo definitivo se realizó mediante medidas globales de importancia y valores SHAP. Este procedimiento se fundamenta en los valores de Shapley y permite distribuir entre las características la diferencia entre una predicción individual y el valor esperado del modelo (Lundberg & Lee, 2017).

Para evitar que la interpretación se realizara exclusivamente sobre observaciones empleadas durante el entrenamiento, los valores SHAP fueron calculados sobre una muestra estratificada de 5 000 nacimientos pertenecientes al conjunto temporal independiente de 2024.

La importancia global de cada predictor se obtuvo a partir del promedio del valor absoluto de sus contribuciones SHAP. De esta manera, las variables con mayores valores promedio fueron consideradas aquellas con mayor participación en las predicciones generadas por el modelo.

Los valores SHAP positivos fueron interpretados como contribuciones que desplazaron las predicciones hacia una mayor probabilidad de bajo peso al nacer, mientras que los valores negativos representaron contribuciones en dirección hacia una menor probabilidad del evento.

Para los predictores numéricos se calculó adicionalmente la correlación de Spearman entre sus valores observados y las correspondientes contribuciones SHAP. En las variables categóricas, las contribuciones fueron resumidas por categoría con el propósito de identificar patrones asociados con incrementos o reducciones de la predicción.

Los resultados de explicabilidad se interpretaron únicamente en términos predictivos. Los valores SHAP permiten examinar cómo participan las variables en las predicciones del modelo, pero no demuestran relaciones causales entre los predictores y el bajo peso al nacer.

Análisis de sensibilidad

Debido a la elevada importancia predictiva observada para la duración gestacional, se realizó un análisis descriptivo complementario de la prevalencia de bajo peso al nacer según semana de embarazo utilizando los registros correspondientes a 2023.

Para cada semana gestacional se calculó el porcentaje de nacimientos con bajo peso respecto del total de nacimientos registrados en esa misma semana. Este análisis permitió examinar directamente el comportamiento observado del resultado a través de las distintas duraciones del embarazo, independientemente de las explicaciones producidas por el modelo.

Finalmente, se efectuó un análisis de sensibilidad mediante ablación de la duración gestacional. Para ello se entrenó un segundo modelo CatBoost eliminando únicamente esta variable y manteniendo los demás predictores y los mismos hiperparámetros utilizados por el modelo principal. No se realizó una nueva optimización.

El desempeño del modelo reducido fue comparado con el del modelo completo mediante Average Precision y ROC-AUC. La reducción observada en estas métricas permitió estimar en qué medida el rendimiento predictivo dependía de la información proporcionada por la duración gestacional. Este análisis se utilizó

únicamente como procedimiento de sensibilidad y no para presentar el modelo reducido como una alternativa al clasificador definitivo.

3. RESULTADOS Y DISCUSIÓN

3.1 Conformación de la base analítica

La base inicial estuvo conformada por 4 874 510 registros de nacimientos. Durante la depuración se excluyeron 1 364 observaciones por inconsistencias en el peso al nacer: 1 315 presentaron valores no positivos, dos registraron pesos inferiores a 300 g y 47 superaron los 6 000 g. En consecuencia, la base analítica final quedó constituida por 4 873 146 nacimientos correspondientes al periodo 2015-2025.

Como se presenta en la Tabla 1, el conjunto histórico de 2015-2022 concentró 3 726 200 registros, de los cuales 233 235 correspondieron a bajo peso al nacer, con una prevalencia de 6,26 %. En 2023 la prevalencia aumentó a 7,00 %, mientras que en 2024 y 2025 alcanzó 6,90 % y 6,77 %, respectivamente. Estas proporciones mostraron una frecuencia relativamente estable del evento en los periodos utilizados para la validación y evaluación temporal.

Tabla 1. Características generales de la base analítica

Partición	Periodo	Registros	Bajo peso	Peso ≥ 2500 g	Prevalencia de bajo peso
Desarrollo histórico	2015-2022	3 726 200	233 235	3 492 965	6,26 %
Validación operativa	2023	410 422	28 735	381 687	7,00 %
Evaluación temporal	2024	390 047	26 921	363 126	6,90 %
Estabilidad temporal	2025	346 477	23 456	323 021	6,77 %
Total	**2015-2025**	**4 873 146**	**312 347**	**4 560 799**	**6,41 %**

3.2 Evaluación de los modelos de aprendizaje automático

La optimización inicial mostró resultados próximos entre los tres algoritmos. En la confirmación realizada con 2022, CatBoost alcanzó un Average Precision (AP) de 0,655, seguido de LightGBM con 0,650 y XGBoost con 0,647. Posteriormente, los modelos fueron comparados utilizando las mismas diez particiones de los registros de 2023.

CatBoost obtuvo el mayor AP promedio, con 0,680 ± 0,005, seguido de XGBoost con 0,678 ± 0,006 y LightGBM con 0,678 ± 0,005. Asimismo, CatBoost alcanzó el mayor ROC-AUC promedio, con 0,910. Aunque las diferencias absolutas de AP fueron pequeñas, CatBoost presentó el mejor resultado en nueve de los diez folds frente a cada uno de los otros algoritmos y obtuvo el menor rango promedio, como se muestra en la Tabla 2.

Tabla 2. Desempeño de los algoritmos en la validación de diez folds sobre 2023

Algoritmo	AP promedio	DE	IC 95 % del AP	ROC-AUC promedio	Rango promedio
CatBoost	0,6800	0,0050	0,6764-0,6836	0,9100	1,20
LightGBM	0,6783	0,0053	0,6745-0,6821	0,9081	2,40
XGBoost	0,6784	0,0055	0,6744-0,6823	0,9061	2,40

La prueba de Friedman identificó diferencias globales estadísticamente significativas entre los algoritmos (χ^2 = 9,60; p = 0,0082), con un W de Kendall de 0,48, correspondiente a un efecto moderado. En las comparaciones pareadas, CatBoost presentó un AP significativamente superior al de LightGBM y XGBoost después de la corrección de Holm (p ajustado = 0,0117 en ambos casos). En contraste, no se encontraron diferencias significativas entre LightGBM y XGBoost (p ajustado = 0,5566). En función de estos resultados, CatBoost fue seleccionado como algoritmo definitivo.

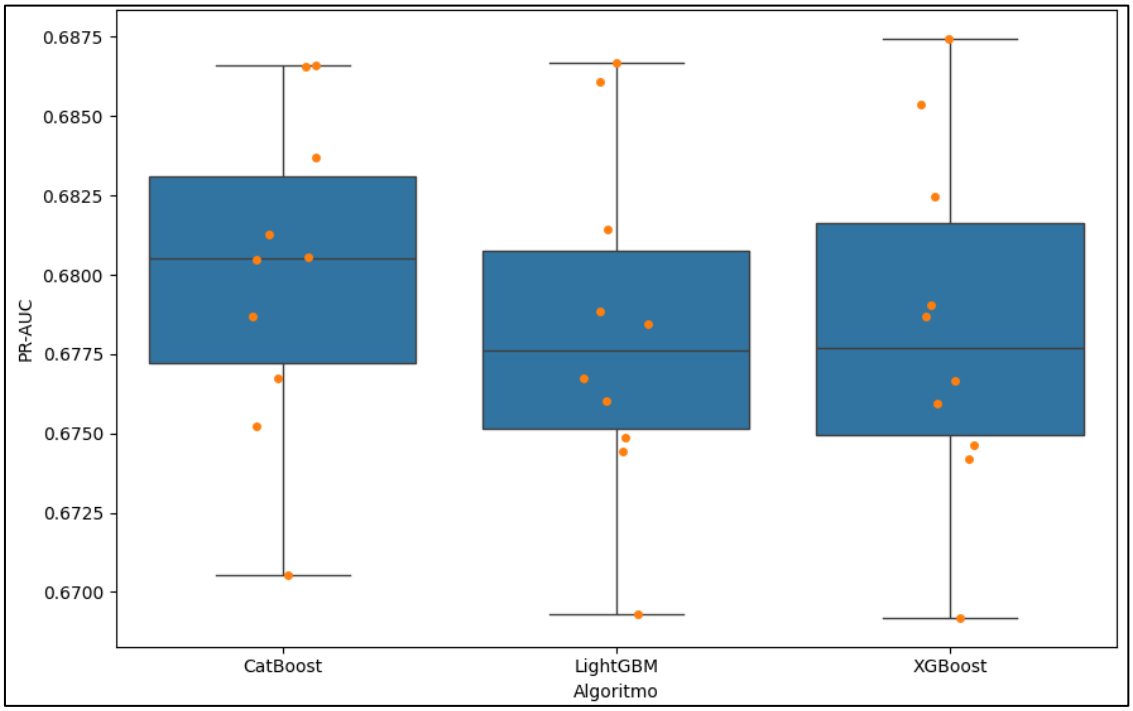

Figura 2. Distribución del Average Precision obtenido por CatBoost, LightGBM y XGBoost en los diez folds de validación

3.3 Modelo definitivo y selección del punto operativo

El modelo CatBoost seleccionado fue reentrenado con la totalidad de los 3 726 200 nacimientos correspondientes al periodo 2015-2022. La configuración definitiva utilizó 630 iteraciones, profundidad de 7 y una tasa de aprendizaje de 0,0423. Posteriormente, sus probabilidades fueron calibradas mediante escalamiento de Platt utilizando los 410 422 registros de 2023.

La selección del punto de corte se realizó priorizando una sensibilidad mínima de 70 %. Bajo este criterio, el umbral definitivo fue de 0,137972, con una sensibilidad de 70,00 %, especificidad de 94,66 %, precisión de 49,69 % y F1-score de 0,581 en 2023. Aunque un umbral de 0,321889 permitió alcanzar un F1-score superior, de 0,632, su sensibilidad se redujo a 58,52 %, por lo que no fue seleccionado como punto operativo.

La calibración modificó la escala de las probabilidades sin alterar el orden discriminativo producido por el modelo. La correspondencia entre las probabilidades estimadas y las frecuencias observadas mostró un comportamiento próximo a la referencia después de la calibración, manteniéndose este patrón durante la evaluación temporal posterior.

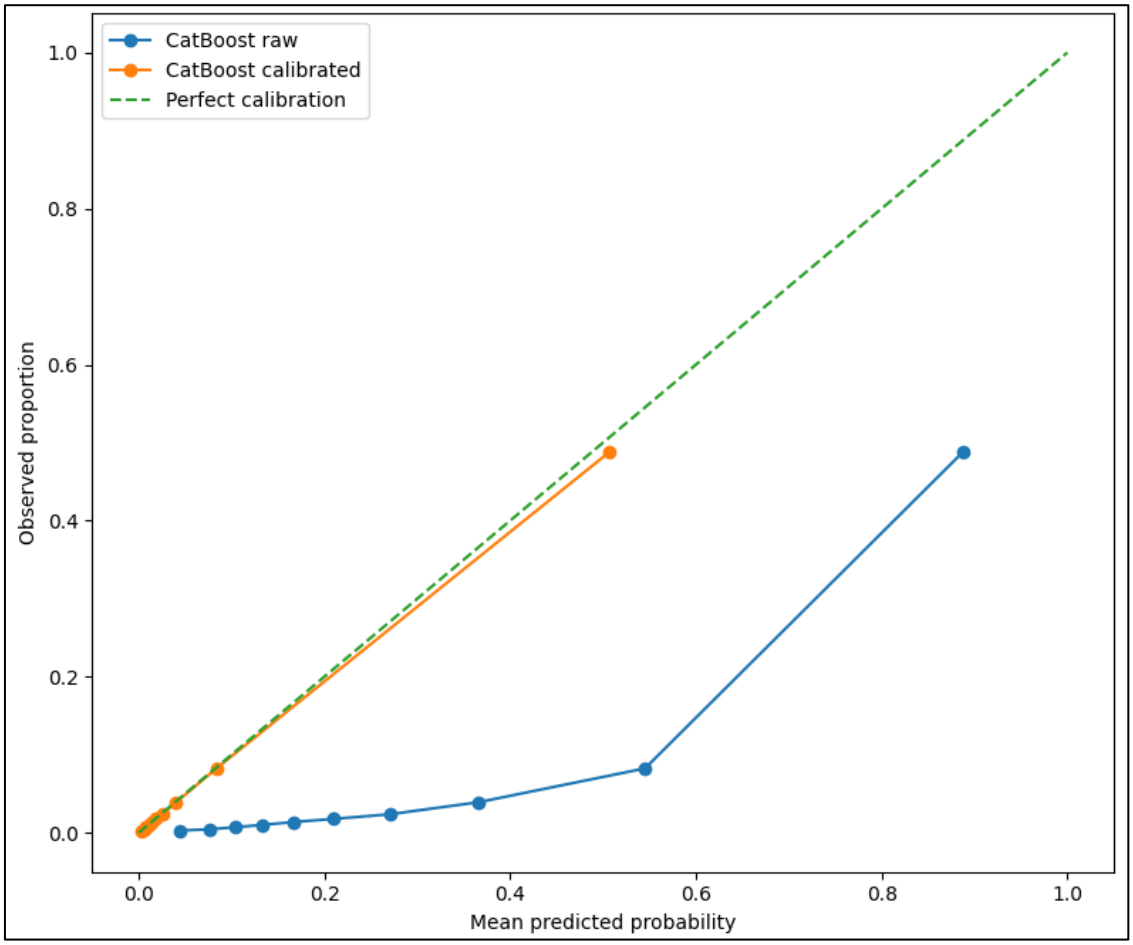

Figura 3. Calibración de las probabilidades del modelo CatBoost en la evaluación temporal de 2024

3.4 Evaluación temporal y estabilidad del modelo

El modelo definitivo, el calibrador y el umbral establecidos previamente fueron aplicados sin reajuste a los registros de 2024. En este conjunto independiente, CatBoost alcanzó un AP de 0,681, un ROC-AUC de 0,913 y una exactitud balanceada de 0,827. La sensibilidad fue de 71,12 % y la especificidad de 94,31 %, mientras que la precisión y el F1-score alcanzaron 48,09 % y 0,574, respectivamente.

De los 26 921 casos de bajo peso registrados en 2024, el modelo identificó correctamente 19 145 y dejó sin detectar 7 776. Entre los 363 126 nacimientos con peso igual o superior a 2 500 g, clasificó correctamente 342 462 y generó 20 664 resultados falsos positivos. El valor predictivo negativo alcanzó 97,78 %, lo que indicó una elevada proporción de resultados negativos correctamente identificados.

La aplicación sin modificaciones a los registros de 2025 mostró un comportamiento similar. El AP aumentó ligeramente a 0,685 y el ROC-AUC a 0,915; asimismo, la sensibilidad alcanzó 71,85 % y la especificidad se mantuvo en 94,21 %. El F1-score fue de 0,571 y la exactitud balanceada de 0,830. Los valores de Brier score y pérdida logarítmica también permanecieron estables, como se observa en la Tabla 3.

Tabla 3. Desempeño temporal del modelo CatBoost definitivo

Métrica	Validación 2023	Evaluación 2024	Estabilidad 2025
Average Precision	0,6806	0,6811	0,6847
ROC-AUC	0,9104	0,9130	0,9152
Precisión	0,4969	0,4809	0,4741

Sensibilidad	0,7000	0,7112	0,7185
Especificidad	0,9466	0,9431	0,9421
F1-score	0,5812	0,5738	0,5713
Exactitud balanceada	0,8233	0,8271	0,8303
MCC	0,5534	0,5478	0,5470
Brier score	0,0357	0,0352	0,0344
LogLoss	0,1378	0,1356	0,1326

En conjunto, las variaciones observadas entre 2023, 2024 y 2025 fueron reducidas. La discriminación y sensibilidad aumentaron ligeramente en los periodos posteriores, mientras que la precisión y el F1-score mostraron disminuciones pequeñas. Estos resultados evidenciaron estabilidad temporal del comportamiento predictivo bajo el punto operativo definido previamente.

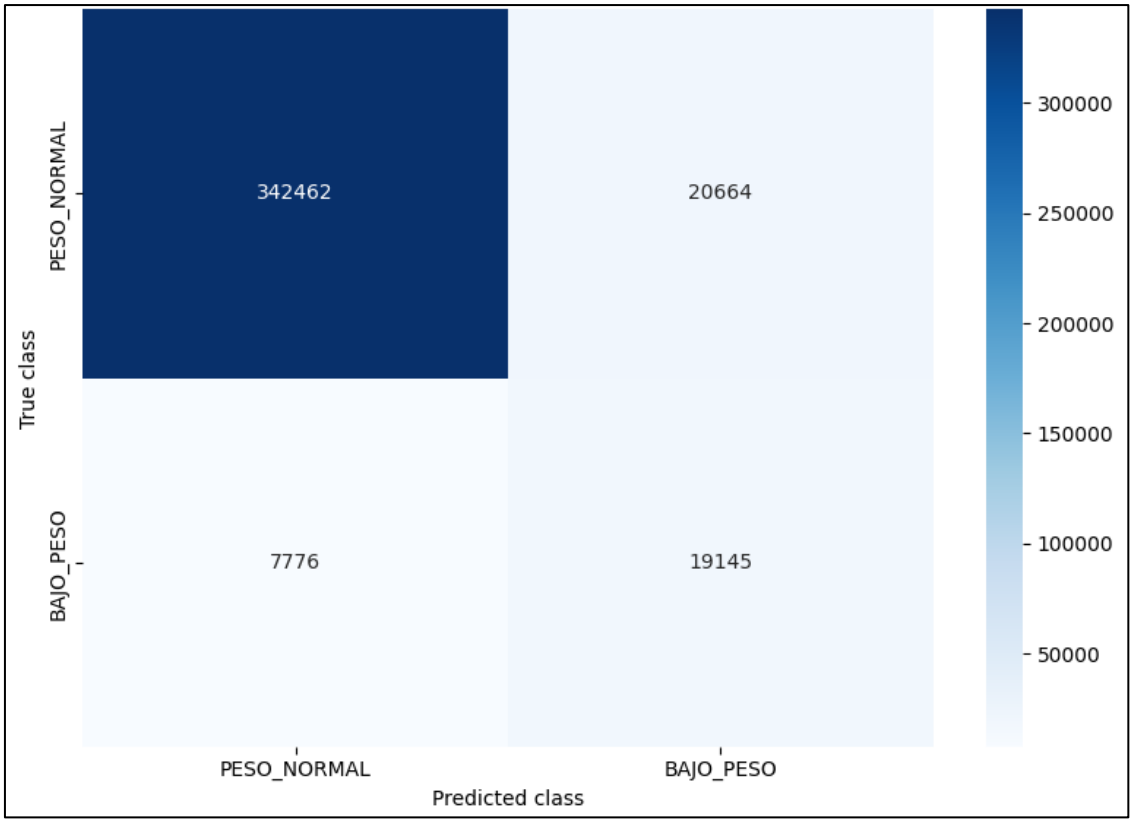

Figura 4. Matriz de confusión del modelo CatBoost en la evaluación temporal de 2024

3.5 Explicabilidad de las predicciones

El análisis mediante valores SHAP mostró una contribución marcadamente predominante de la duración gestacional. Esta variable alcanzó un valor SHAP absoluto promedio de 1,758, considerablemente superior al observado para el tipo de parto (0,170), sexo del nacido (0,168), categoría de la IPRESS (0,138), nivel de instrucción materna (0,120) y número de embarazos maternos (0,119).

La importancia nativa de CatBoost mostró un patrón concordante: la duración gestacional concentró el 66,07 % de la importancia total, seguida a distancia por el tipo de parto, con 6,63 %. Las restantes variables

presentaron contribuciones individuales inferiores al 4 %, lo que evidenció una elevada dependencia del modelo respecto de la información sobre la duración del embarazo.

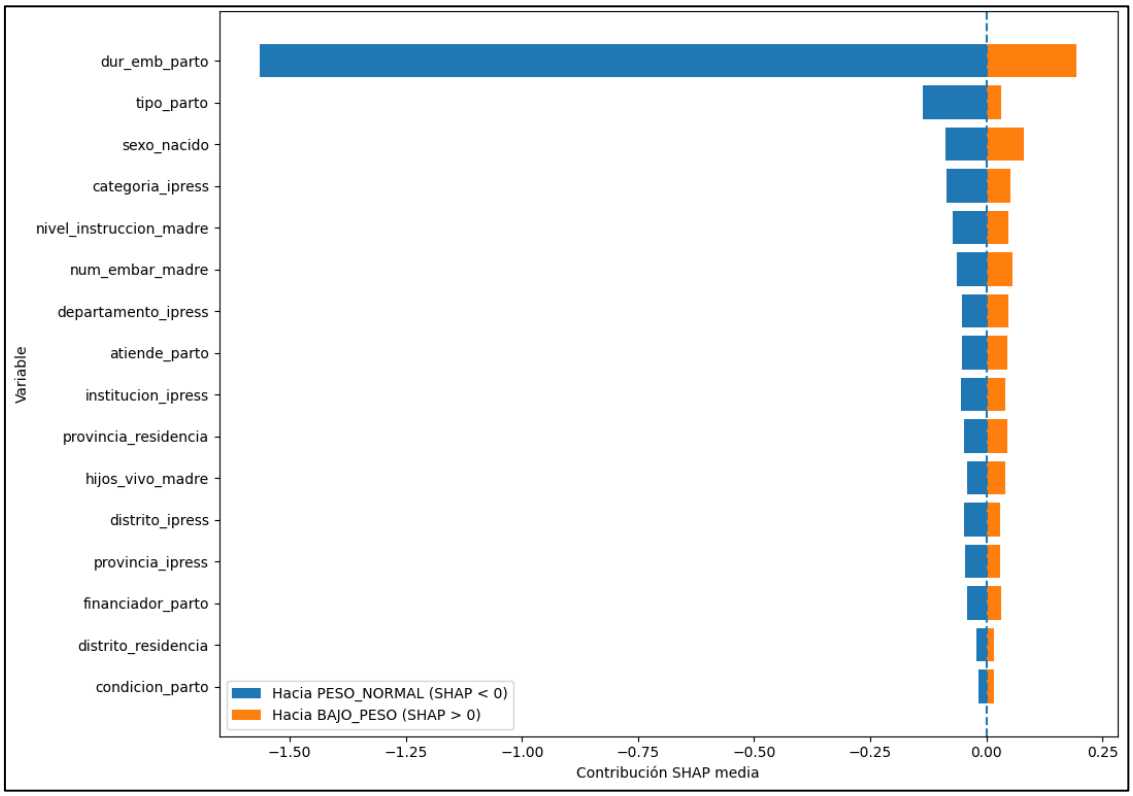

Figura 5. Contribución global de los predictores a las estimaciones de bajo peso al nacer mediante valores SHAP

En las variables numéricas, la duración gestacional presentó una correlación de Spearman de −0,967 entre sus valores y las contribuciones SHAP. También se observaron correlaciones negativas para el número de embarazos maternos (ρ = −0,904) y el número de hijos vivos (ρ = −0,724). En el contexto del modelo, estos resultados indicaron que valores mayores de estas características tendieron a desplazar las predicciones hacia una menor probabilidad de bajo peso, aunque estas relaciones no deben interpretarse como efectos causales.

La relación observada directamente en los registros de 2023 fue consistente con el patrón identificado por el modelo. La prevalencia de bajo peso alcanzó 57,47 % a las 35 semanas, disminuyó a 31,56 % a las 36 semanas y a 11,49 % a las 37 semanas. Posteriormente descendió a 3,44 % a las 38 semanas, 1,60 % a las 39 y 1,03 % a las 40 semanas.

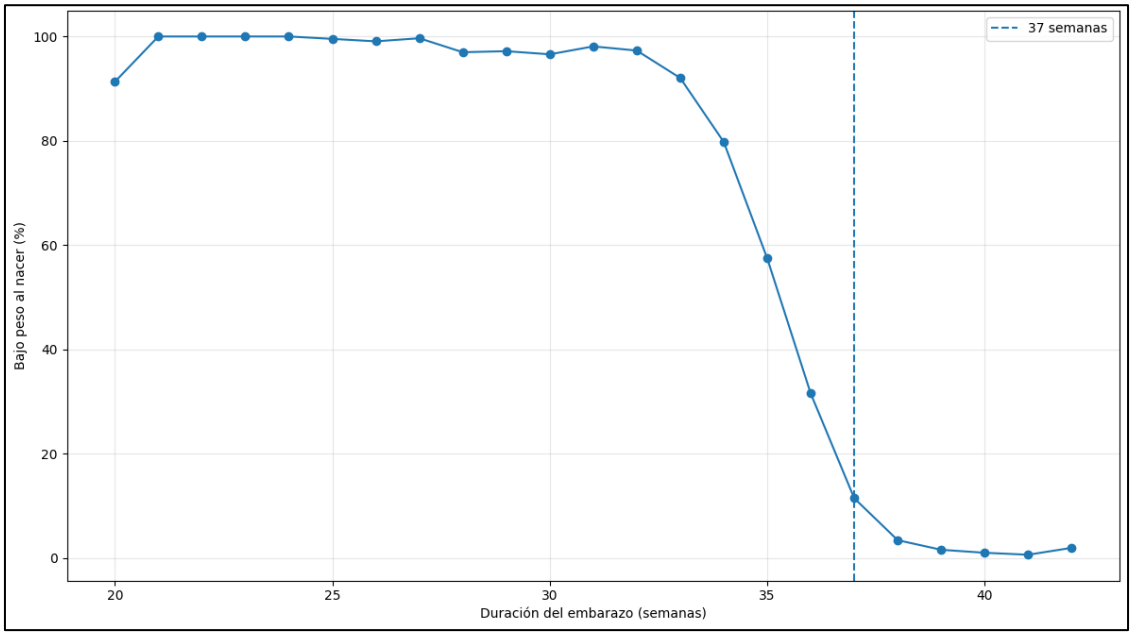

Figura 6. Relación observada entre la duración gestacional y la prevalencia de bajo peso al nacer en 2023

3.6 Análisis de sensibilidad de la duración gestacional

Debido al predominio de la duración gestacional en la explicación del modelo, se realizó un análisis de sensibilidad eliminando esta variable y manteniendo sin cambios los demás predictores y los hiperparámetros del CatBoost definitivo.

Como se presenta en la Tabla 4, la exclusión de la duración gestacional produjo una reducción sustancial del desempeño en los tres periodos evaluados. En 2024, el AP disminuyó de 0,681 a 0,295, mientras que el ROC-AUC pasó de 0,913 a 0,765. El mismo comportamiento se observó en 2025, cuando el AP disminuyó de 0,685 a 0,281 y el ROC-AUC de 0,915 a 0,761.

Tabla 4. Análisis de sensibilidad del modelo con y sin duración gestacional

Periodo	AP con gestación	AP sin gestación	Diferencia AP	ROC-AUC con gestación	ROC-AUC sin gestación	Diferencia ROC-AUC
2023	0,6806	0,2969	0,3837	0,9104	0,7610	0,1494
2024	0,6811	0,2949	0,3862	0,9130	0,7653	0,1477
2025	0,6847	0,2807	0,4040	0,9152	0,7610	0,1542

La magnitud de esta reducción confirmó que la duración gestacional constituyó la principal fuente de información predictiva del modelo. No obstante, el desempeño conservado después de su eliminación, particularmente un ROC-AUC cercano a 0,76, mostró que las demás características mantuvieron capacidad para discriminar parcialmente los nacimientos con bajo peso.

3.7 Discusión

Los resultados mostraron que los registros nacionales del Ministerio de Salud del Perú permiten desarrollar modelos de aprendizaje automático con capacidad para discriminar el bajo peso al nacer a gran escala. CatBoost presentó el mejor desempeño entre los algoritmos comparados y mantuvo valores de ROC-AUC superiores a 0,91 y Average Precision cercanos a 0,68 entre 2023 y 2025. Aunque la diferencia absoluta frente a LightGBM y XGBoost fue pequeña, CatBoost obtuvo mejores resultados de manera consistente en nueve de los diez folds, lo que explica la significancia observada en las comparaciones

pareadas. Este comportamiento es coherente con estudios recientes que han encontrado un rendimiento competitivo de los algoritmos de boosting en la predicción del bajo peso al nacer. Ranjbar et al. (2023), utilizando 8 853 nacimientos, identificaron a XGBoost entre los modelos con mejor desempeño y obtuvieron una sensibilidad de 0,69 y un F1-score de 0,77. De manera similar, Victor et al. (2025) compararon XGBoost, CatBoost, LightGBM y Random Forest y observaron valores de ROC-AUC cercanos a 0,94 para los principales algoritmos. En conjunto, estos resultados sugieren que las diferencias entre métodos de boosting pueden ser reducidas y que la elección del modelo debe considerar no solo el valor máximo de una métrica, sino también su estabilidad y comportamiento frente a datos posteriores.

Uno de los principales hallazgos del presente estudio fue precisamente la estabilidad temporal del modelo. Después de entrenarse con los registros de 2015-2022 y utilizar 2023 para la calibración y definición del punto operativo, el desempeño no disminuyó en los periodos posteriores. El ROC-AUC pasó de 0,910 en 2023 a 0,913 en 2024 y 0,915 en 2025, mientras que el Average Precision se mantuvo entre 0,681 y 0,685. La sensibilidad, especificidad y Brier score también presentaron variaciones reducidas. Este resultado fortalece la evidencia de generalización dentro del sistema nacional de registros, porque el rendimiento fue evaluado en cohortes cronológicamente posteriores y no únicamente mediante una partición aleatoria de la misma población. No obstante, esta estabilidad temporal no debe considerarse equivalente a una validación externa, ya que todos los periodos procedieron del mismo sistema nacional de información. En consecuencia, una aplicación futura requerirá comprobar la transportabilidad del modelo en otras poblaciones o escenarios asistenciales.

La selección de un umbral orientado a alcanzar una sensibilidad mínima de 70 % permitió identificar aproximadamente siete de cada diez nacimientos con bajo peso en 2024 y 2025, manteniendo una especificidad superior al 94 %. Sin embargo, la precisión permaneció alrededor del 48 %, lo que implica una proporción importante de falsas alertas. Este comportamiento responde a una decisión deliberada de favorecer la detección de casos y reducir los falsos negativos, en lugar de maximizar exclusivamente métricas como el F1-score. Por ello, el umbral de 0,137972 debe interpretarse como un punto operativo establecido para este experimento y no como un criterio clínico universal. Antes de una eventual utilización en servicios de salud sería necesario valorar las consecuencias de las falsas alertas, los recursos requeridos para su revisión y el beneficio clínico de intervenir sobre los casos identificados. Asimismo, su implementación requeriría que los profesionales de la salud comprendan las probabilidades generadas, las limitaciones del modelo y la evidencia científica que sustenta su utilización. Esta consideración es relevante porque las competencias para buscar, evaluar y emplear información científica constituyen un componente importante de la práctica basada en evidencia entre los profesionales médicos (Valles-Coral et al., 2025).

El hallazgo interpretativo más importante fue el predominio de la duración gestacional. Esta variable presentó una contribución considerablemente superior a la del resto de predictores y mostró una relación inversa muy marcada con las estimaciones de bajo peso. El patrón fue consistente con los datos observados, pues la prevalencia disminuyó desde 57,47 % a las 35 semanas hasta 1,03 % a las 40 semanas. La literatura reciente presenta resultados concordantes. Ranjbar et al. (2023) identificaron la edad gestacional y el antecedente de bajo peso como los principales predictores de su modelo. Victor et al. (2025) encontraron igualmente que la edad gestacional fue la característica más influyente en los modelos de boosting, mientras que Wu et al. (2025) señalaron la gestación inferior a 37 semanas entre los factores más relevantes identificados mediante XGBoost y SHAP. En Bangladesh, Sultana et al. (2025) encontraron que la duración del embarazo fue también el predictor con mayor influencia en la clasificación del bajo peso. La coincidencia entre contextos distintos respalda la coherencia del patrón encontrado en los registros peruanos.

El análisis de ablación reforzó esta interpretación. Al retirar únicamente la duración gestacional, el Average Precision disminuyó de 0,681 a 0,295 en 2024 y de 0,685 a 0,281 en 2025; el ROC-AUC cayó aproximadamente de 0,91 a 0,76. Por tanto, una parte sustancial de la capacidad predictiva del modelo

depende de esta variable. Este hallazgo representa simultáneamente una fortaleza y una limitación. Por un lado, confirma que el modelo identifica una señal clínica estrechamente relacionada con el bajo peso; por otro, impide considerarlo una herramienta de predicción prenatal temprana, debido a que la duración final del embarazo solo se conoce en etapas cercanas al nacimiento. En consecuencia, el modelo resulta más apropiado como herramienta de estratificación e interpretación del riesgo en registros perinatales que como sistema de alerta durante las primeras etapas de la gestación. Un modelo dirigido a predicción temprana requeriría restringir los predictores a información disponible antes del desenlace.

Finalmente, el estudio presenta como fortalezas el uso de más de 4,8 millones de registros nacionales, la comparación de algoritmos bajo las mismas particiones, la evaluación temporal en dos periodos posteriores, la calibración probabilística y la incorporación de SHAP y análisis de ablación. Sin embargo, sus resultados dependen de la calidad de los registros administrativos y no incorporan algunas variables clínicas, nutricionales o conductuales utilizadas en investigaciones recientes, como índice de masa corporal materno, hipertensión gestacional, tabaquismo o características detalladas del control prenatal. Además, las contribuciones SHAP deben interpretarse como señales predictivas y no como efectos causales. En conjunto, los hallazgos muestran que el aprendizaje automático puede aprovechar registros rutinarios del MINSA para discriminar el bajo peso al nacer con un comportamiento temporal estable y proporcionar información interpretable sobre los predictores empleados. Su principal aporte radica en combinar escala nacional, comparación algorítmica, validación temporal y explicabilidad, aunque su eventual aplicación requerirá validación externa y una definición más precisa del momento clínico en que se pretende utilizar.

CONCLUSIONES

La integración de los registros nacionales de nacimientos del Ministerio de Salud del Perú permitió desarrollar y validar un modelo explicable de aprendizaje automático para la predicción del bajo peso al nacer. CatBoost fue seleccionado por presentar el mejor desempeño comparativo frente a LightGBM y XGBoost y mantuvo resultados estables durante la evaluación temporal en 2024 y 2025, lo que evidenció una adecuada capacidad de generalización dentro del sistema nacional de registros. La duración gestacional constituyó el principal predictor del modelo, hallazgo confirmado tanto por el análisis SHAP como por la reducción del rendimiento observada al retirar esta variable. No obstante, su elevada contribución indica que el modelo resulta más apropiado para la estratificación e interpretación del riesgo en registros perinatales que para una predicción prenatal temprana. En consecuencia, el enfoque propuesto demuestra el potencial de los datos administrativos del MINSA para apoyar la identificación y análisis del bajo peso al nacer a gran escala, siempre que sus predicciones se interpreten como apoyo a la vigilancia y al análisis sanitario y no como sustituto de la evaluación clínica.

FINANCIAMIENTO

Los autores no recibieron ningún patrocinio para llevar a cabo este estudio-artículo.

CONFLICTO DE INTERESES

Los autores declaran que no existe ningún tipo de conflicto de interés relacionado con la materia del trabajo.

CONTRIBUCIÓN DE AUTORES

Conceptualización: Andrade-Girón, D.; Neri-Ayala, A. Curación de datos: Andrade-Girón, D.; Luna-Victoria, M. A.; Oscuvicla-Tapia, E. Análisis formal: Andrade-Girón, D.; Neri-Ayala, A.; Luna-Victoria, M. A. Investigación: Andrade-Girón, D.; Neri-Ayala, A.; Oscuvicla-Tapia, E.; Peña, A. Metodología: Andrade-Girón, D.; Neri-Ayala, A.; Cuevas-Huari, E. Administración del proyecto: Andrade-Girón, D.; Cuevas-Huari, E.

Software: Andrade-Girón, D.; Luna-Victoria, M. A. Supervisión: Peña, A.; Cuevas-Huari, E. Validación: Neri-Ayala, A.; Oscuvicla-Tapia, E.; Peña, A. Visualización: Andrade-Girón, D.; Luna-Victoria, M. A. Redacción – borrador original: Andrade-Girón, D.; Neri-Ayala, A. Redacción – revisión y edición: Andrade-Girón, D.; Neri-Ayala, A.; Luna-Victoria, M. A.; Oscuvicla-Tapia, E.; Peña, A.; Cuevas-Huari, E.

REFERENCIAS

Blencowe, H., Krasevec, J., de Onis, M., Black, R. E., An, X., Stevens, G. A., Borghi, E., Hayashi, C., Estevez, D., Cegolon, L., Shiekh, S., Ponce Hardy, V., Lawn, J. E., & Cousens, S. (2019). National, regional, and worldwide estimates of low birthweight in 2015, with trends from 2000: a systematic analysis. *The Lancet Global Health*, *7*(7), e849–e860. https://doi.org/10.1016/S2214-109X(18)30565-5

Cajachagua-Torres, K. N., Quezada-Pinedo, H. G., Guzman-Vilca, W. C., Tarazona-Meza, C., Carrillo-Larco, R. M., & Huicho, L. (2024). Vulnerable newborn phenotypes in Peru: a population-based study of 3,841,531 births at national and subnational levels from 2012 to 2021. *The Lancet Regional Health - Americas*, *31*, 100695. https://doi.org/10.1016/j.lana.2024.100695

Carrillo-Larco, R. M., Cajachagua-Torres, K. N., Guzman-Vilca, W. C., Quezada-Pinedo, H. G., Tarazona-Meza, C., & Huicho, L. (2021). National and subnational trends of birthweight in Peru: Pooled analysis of 2,927,761 births between 2012 and 2019 from the national birth registry. *The Lancet Regional Health - Americas*, *1*, 100017. https://doi.org/10.1016/j.lana.2021.100017

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. https://doi.org/10.1145/2939672.2939785

Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, *7*(1), 1–30. http://jmlr.org/papers/v7/demsar06a.html

Dominguez Miranda, S. A., & Rodriguez Aguilar, R. (2024). Machine learning models in health prevention and promotion and labor productivity: A co-word analysis. *Iberoamerican Journal of Science Measurement and Communication*, *4*(1), 1–16. https://doi.org/10.47909/ijsmc.85

García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, *180*(10), 2044–2064. https://doi.org/10.1016/j.ins.2009.12.010

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Conference: Advances in Neural Information Processing Systems 30 (NIPS 2017)*. https://papers.nips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. http://arxiv.org/abs/1705.07874

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning - ICML '05*, 625–632. https://doi.org/10.1145/1102351.1102430

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019). *CatBoost: unbiased boosting with categorical features*. http://arxiv.org/abs/1706.09516

Ranjbar, A., Montazeri, F., Farashah, M. V., Mehrnoush, V., Darsareh, F., & 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

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Sultana, N., Afia, Z., Sifat, I. K., Zoha, S., Jisa, T. A., & Kibria, M. K. (2025). Machine learning based prediction of low birth weight and its associated risk factors: Insights from the Bangladesh Demographic and

Health Survey 2022. *PLOS Global Public Health*, *5*(9), e0005187. https://doi.org/10.1371/journal.pgph.0005187

Valles-Coral, M., Pinedo, L., Navarro-Cabrera, J. R., Valverde-Iparraguirre, J., Injante, R., Saavedra, S., Quintanilla-Morales, L. K., & Almeida-Espinosa, A. (2025). Skills in searching for and using scientific information among physicians in the context of evidence-based practice: A descriptive study. *Iberoamerican Journal of Science Measurement and Communication*, *5*(1), 1–9. https://doi.org/10.47909/ijsmc.181

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, *17*(1), 230. https://doi.org/10.1186/s12916-019-1466-7

Victor, A., Almeida, F., Xavier, S. P., & Rondó, P. H. C. (2025). Predicting low birth weight risks in pregnant women in Brazil using machine learning algorithms: data from the Araraquara cohort study. *BMC Pregnancy and Childbirth*, *25*(1), 320. https://doi.org/10.1186/s12884-025-07351-3

WHO, & UNICEF. (2023). *Nutrition and Food Safety*. Word Health Organization. https://www.who.int/teams/nutrition-and-food-safety/monitoring-nutritional-status-and-food-safety-and-events/joint-low-birthweight-estimates

Wu, X., Zhao, Q., Gao, Y., Zhang, Y., Xu, L., Cong, X., Sun, N., Shi, F., & Wang, S. (2025). Interpretable machine learning model for predicting low birth weight in singleton pregnancies: a retrospective cohort study. *BMC Pregnancy and Childbirth*, *25*(1), 1159. https://doi.org/10.1186/s12884-025-08318-0