UNIVERSIDAD NACIONAL DE SAN MARTIN

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

Original article
Artículo original
Jul-Dec, 2026

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

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

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

Received: 31 Dec. 2025 | **Accepted:** 10 Jun. 2026 | **Published:** 20 Jul. 2026

Corresponding author*: eoscuvilca@unjfsc.edu.pe

How to cite this article: 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

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

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

1. INTRODUCTION

Low birth weight remains a relevant public health problem because of its association with newborn survival, growth, and subsequent development. The World Health Organization defines this condition as a birth weight below 2,500 g, regardless of gestational age. Globally, it was estimated that in 2020 approximately 19.8 million newborns, equivalent to 14.7% of all births, had low birth weight, while progress toward reducing its prevalence remained below the established international targets. In addition to increasing vulnerability during the neonatal period, this condition has been associated with higher morbidity, impaired growth and development, and an increased risk of chronic diseases later in life (WHO & UNICEF, 2023).

In Peru, national records have shown that low birth weight presents important temporal, geographic, and social differences. Carrillo-Larco et al. (2021), based on 2,927,761 births registered between 2012 and 2019, observed that the national prevalence decreased from 6.9% to 6.2% during this period; however, subnational differences were considerably greater. In 2019, for example, within the department of Áncash, prevalences of 1.0% in Huarmey and 11.4% in Antonio Raymondi were reported. Likewise, birth weight indicators showed poorer outcomes in high-altitude Andean areas and among mothers with lower educational attainment. Subsequent studies including 3,841,531 births between 2012 and 2021 confirmed the persistence of geographic and socioeconomic inequalities in vulnerable neonatal phenotypes across the country (Cajachagua-Torres et al., 2024).

The availability of administrative birth records provides an opportunity to expand this type of analysis. The Live Birth Certificate database of the Peruvian Ministry of Health contains individual-level information on birth characteristics, maternal history, gestational duration, type and condition of delivery, place of residence, and the health facility where care was provided. Its integration with complementary sources, such as the National Registry of Health Service Provider Institutions and official geographic location codes, makes it possible to incorporate territorial and institutional information on a large scale. Previous studies have demonstrated the usefulness of Peru's national birth registry for describing birth weight patterns and subnational inequalities; however, its use has focused mainly on descriptive epidemiological analyses, creating an opportunity to move toward predictive approaches capable of simultaneously leveraging multiple individual and contextual characteristics.

The incorporation of machine learning into the healthcare sector has extended to various applications related to prevention, health promotion, and the analysis of factors influencing health outcomes, highlighting the growing interest in these techniques within biomedical research (Dominguez Miranda & Rodriguez Aguilar, 2024). In this context, machine learning offers tools for identifying complex and nonlinear relationships among multiple maternal, obstetric, neonatal, and healthcare-related characteristics. In Brazil, Victor et al. (2025) compared Random Forest, XGBoost, CatBoost, and LightGBM for predicting low birth weight using a cohort of pregnant women and additionally incorporated Shapley values to interpret the contribution of individual variables. In China, Wu et al. (2025) evaluated logistic regression and four machine learning algorithms in 10,227 births, finding that gestational age, maternal age, and several clinical characteristics contributed substantially to model predictions; they also used SHAP to explain model behavior. These studies show that machine learning methods can complement conventional statistical approaches by capturing interactions among predictors and generating individualized risk estimates.

Nevertheless, predictive performance alone is insufficient when models are applied to health data. Estimated probabilities should correspond adequately to the observed frequency of the outcome, since an algorithm may show good discriminative ability while simultaneously producing poorly calibrated probabilities (Van Calster et al., 2019). Similarly, tree-based ensemble models may achieve high predictive performance at the expense of reduced transparency regarding the process leading to each prediction. SHAP values partially address this limitation by quantifying the contribution of each feature to both

individual and global model predictions (Lundberg & Lee, 2017). In studies based on longitudinal records, it is also important to determine whether model performance is maintained in periods subsequent to those used for model development rather than relying exclusively on random splits from the same cohort.

Despite these advances, national studies on low birth weight in Peru have mainly focused on describing trends, territorial differences, and vulnerable neonatal phenotypes, whereas the use of national birth records to develop interpretable predictive models and evaluate their performance in subsequent time periods remains less explored. In response to this gap, the objective of the present study was to develop, compare, and interpret machine learning models for predicting low birth weight using records from the Peruvian Ministry of Health. To this end, national records from 2015 to 2025 were integrated; CatBoost, LightGBM, and XGBoost were compared; their performance was statistically evaluated; and a final model was selected, subsequently calibrated, and temporally validated. SHAP values were also applied to identify the characteristics contributing most strongly to model predictions. The proposed approach seeks to provide reproducible evidence for large-scale analysis of low birth weight and to explore the potential of administrative records as a data source for developing predictive tools in maternal and neonatal health.

2. MATERIALS AND METHODS

2.1 Study design and scope

This was an applied study with a quantitative approach and an observational, non-experimental, and retrospective design, aimed at developing, comparing, and interpreting machine learning models for predicting low birth weight. The unit of analysis corresponded to each individual birth record contained in the Live Birth Certificate database of the Peruvian Ministry of Health (CNV-MINSA).

The study included records from 2015 to 2025 and adopted a temporal validation strategy. Data from 2015 to 2022 were used for historical model development; 2023 records were used for comparison through stratified partitions and subsequently for calibration and definition of the operating threshold of the selected model; 2024 constituted the independent temporal evaluation period; and the available 2025 records were reserved to assess subsequent performance stability.

This chronological separation made it possible to evaluate the model's ability to generalize to observations occurring after those used for development, thereby reducing the risk of obtaining overly optimistic estimates derived exclusively from random data partitions.

2.2 Methodological procedure

The procedure included structural and semantic auditing of the data sources, correction of inconsistencies in the main file, normalization of variables and integration keys, construction of the target variable, and linkage with territorial and healthcare-facility information.

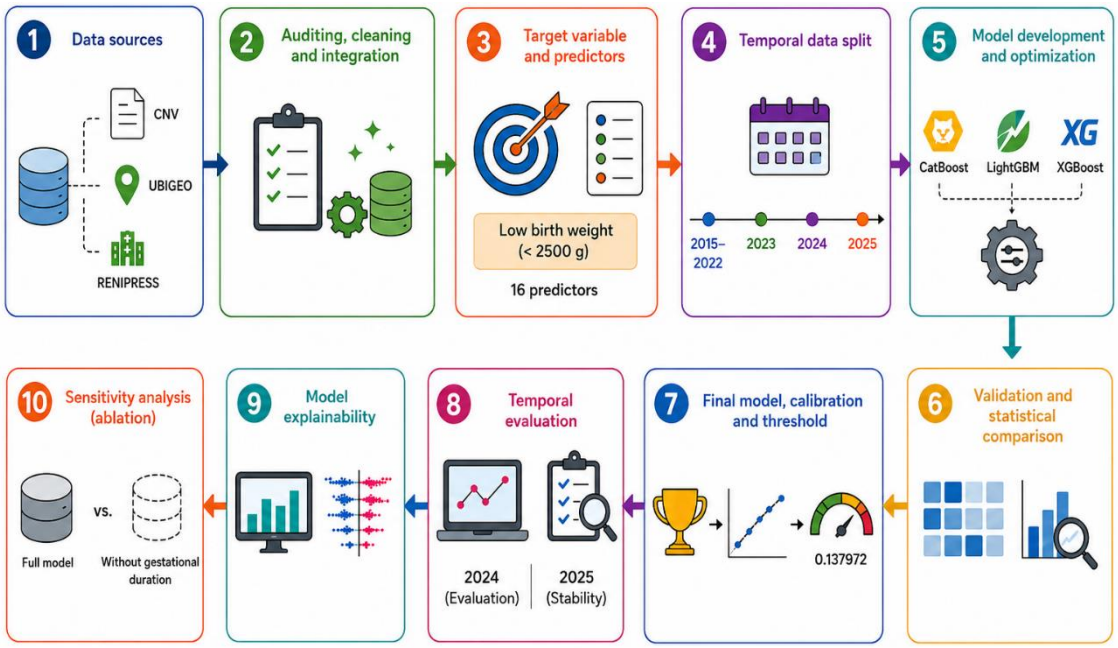

Figure 1. Methodological flowchart for the development and evaluation of the low birth weight predictive model

Subsequently, the records were chronologically partitioned, predictors were prepared, hyperparameters were optimized, and three gradient boosting algorithms were compared: CatBoost, LightGBM, and XGBoost. All three models were evaluated using the same predictor set, the same validation partitions, and a common primary metric.

Algorithm selection was complemented by paired statistical procedures. Finally, the selected model was retrained using the complete 2015–2022 historical dataset, its probabilities were calibrated using 2023 records, and an operating threshold based on a predefined minimum sensitivity was established. The model remained unchanged during its evaluation in 2024 and the stability analysis in 2025.

The procedure was complemented by explainability analyses using SHAP values, calibration assessment, estimation of confidence intervals through bootstrap resampling, and a sensitivity analysis based on ablation of gestational duration.

Data sources and analytical environment

The primary source consisted of records from the Live Birth Certificate database of the Peruvian Ministry of Health, contained in a file with a cutoff date of November 30, 2025. The initial dataset had an approximate size of 793.79 MB and contained 4,874,510 records.

Complementary sources included the CNV data dictionary, the Geographic Location catalog of the National Institute of Statistics and Informatics, and the National Registry of Health Service Provider Institutions (RENIPRESS). The UBIGEO catalog was used to incorporate department, province, and district information, whereas RENIPRESS provided information related to the institution, category, and geographic location of the healthcare facility.

Data processing was performed in Google Colaboratory using Python. pandas and NumPy were used for data manipulation and integration; scikit-learn for preprocessing, partitioning, metrics, and calibration; Optuna for hyperparameter optimization; CatBoost, LightGBM, and XGBoost for supervised learning; SciPy and Statsmodels for statistical analysis; and Matplotlib and Seaborn for graphical representation.

The evaluated algorithms correspond to gradient boosting implementations widely used for tabular-data problems. XGBoost incorporates regularization and optimizations for efficient tree construction (Chen & Guestrin, 2016); LightGBM introduces strategies aimed at improving computational efficiency during

training (Ke et al., 2017); and CatBoost incorporates specific procedures for handling categorical variables and reducing bias during their encoding (Prokhorenkova et al., 2019).

Data auditing, cleaning, and integration

Initially, a structural audit was performed to identify encoding, delimiter, dimensions, available columns, special values, and the overall consistency of the files. CSV files were processed using UTF-8-SIG encoding and a semicolon as the delimiter.

During this stage, additional separators were detected within the field corresponding to the description of maternal occupation. Therefore, these records were repaired by reconstructing the original column structure before further processing.

A semantic audit of numeric, categorical, temporal, and identification variables was then performed. Ranges, categories, special values, integration coverage with UBIGEO and RENIPRESS, and potential duplicate records were examined.

Final data cleaning was performed in chunks of 200,000 records to reduce memory consumption. Birth weights between 300 and 6,000 g, maternal ages between 10 and 59 years, and gestational durations between 20 and 45 weeks were considered valid.

Records with birth weight values outside the established criteria were excluded because this variable was directly used to define the outcome of interest. In contrast, maternal age and gestational duration values outside their respective ranges were converted to missing values and subsequently handled during modeling.

Of the initial 4,874,510 records, 1,364 were excluded because of inconsistencies in birth weight. Of these, 1,315 had non-positive values, two had positive weights below 300 g, and 47 had values above 6,000 g. As a result, the analytical dataset consisted of 4,873,146 births.

Categorical variables were normalized by removing additional spaces, converting text to uppercase, and unifying different representations of missingness. UBIGEO codes were standardized to six digits. For integration with RENIPRESS, IPRESS codes were normalized before linkage with the main dataset.

Counts related to maternal history were converted to numeric format. For the variables corresponding to number of pregnancies and number of living children, the category “5 or more” was represented by the value 5. For other maternal-history variables, the category “none” was represented by zero and “11 or more” by the value 11.

Maternal occupation was deterministically grouped into occupational categories using previously defined text-based rules.

A total of 974 potentially duplicated records were retained because the source did not contain an individual identifier that would allow confirmation that they corresponded to the same birth. Likewise, the variable related to the mother’s number of deceased children was excluded because approximately 96.97% of its records corresponded to ignored or unavailable values.

Definition of the target variable and predictors

Low birth weight was defined as a birth weight below 2,500 g, an internationally used criterion for identifying this neonatal outcome (Blencowe et al., 2019).

The target variable was coded as 1 for births with weights between 300 and less than 2,500 g and as 0 for births with weights between 2,500 and 6,000 g.

The newborn weight variable was used exclusively to construct the outcome and was subsequently removed from the predictor set. Newborn length was also excluded because of its direct proximity to the

neonatal outcome and the risk of introducing information excessively related to the variable being predicted.

The comparative experiment used a common set of 16 predictors: gestational duration, IPRESS category, type of delivery, IPRESS district, province of residence, IPRESS province, IPRESS department, maternal educational level, number of living children, delivery financing source, delivery condition, number of maternal pregnancies, IPRESS institution, professional or personnel attending the delivery, district of residence, and newborn sex.

Gestational duration, number of living children, and number of maternal pregnancies were treated as numerical variables, whereas the remaining 13 predictors were processed as categorical variables. The same predictor set was used across all three algorithms to maintain comparable conditions during the experimental evaluation.

Temporal partitioning and preparation for modeling

After data cleaning, records were chronologically divided into four datasets. The 2015–2022 period included 3,726,200 births, of which 233,235 corresponded to low birth weight, representing a prevalence of 6.2593%. The year 2023 comprised 410,422 records, with a prevalence of 7.0013%; 2024 included 390,047 records, with a prevalence of 6.9020%; and 2025 consisted of 346,477 records available up to the cutoff date, with a prevalence of 6.7699%.

To reduce computational cost during the initial algorithm comparison, a sample of 600,000 observations was obtained from the 2015–2022 historical dataset. Sampling was simultaneously stratified by year and target class, using a random seed of 42, in order to preserve the temporal structure and proportion of low-birth-weight cases.

Missing numerical variables were imputed using the median calculated exclusively from the records assigned to training at each stage. Missing categorical values were represented using an explicit unknown category. Parameters derived from the training data were subsequently applied to the validation datasets without refitting.

Because of class imbalance, weighting was incorporated during training. The weight assigned to low-birth-weight cases was calculated from the ratio between the number of observations in the majority class and the number of observations in the minority class. This procedure increased the importance of errors involving low-birth-weight cases without artificially modifying the data distribution through oversampling.

Model development and optimization

CatBoost, LightGBM, and XGBoost were compared, all of which are gradient boosting algorithms based on decision trees. In general, these methods sequentially construct multiple trees so that each new iteration attempts to correct part of the errors produced by previous trees.

CatBoost incorporates specific procedures for handling categorical variables and ordered boosting (Prokhorenkova et al., 2019); LightGBM introduces strategies designed to increase computational efficiency during learning (Ke et al., 2017); and XGBoost uses regularization and different optimizations aimed at efficient tree training (Chen & Guestrin, 2016).

Hyperparameter optimization was performed using Optuna with the Tree-structured Parzen Estimator and 15 trials per algorithm. For this stage, historical records from the sample were divided into training data corresponding to 2015–2021 and validation data corresponding to 2022. Each trial used a maximum of 150,000 observations for training and 50,000 for validation, obtained through stratified sampling.

For CatBoost, the search space included 300–650 iterations, depths between 5 and 7, learning rates between 0.035 and 0.09, and L2 regularization between 3 and 10. Other parameters related to categorical-variable handling, boosting type, bootstrap procedure, and subsampling proportion were kept constant.

For LightGBM, learning rates between 0.02 and 0.08, 31–127 leaves, depths between 5 and 9, a minimum number of observations per leaf between 30 and 100, and L2 regularization between 1 and 10 were evaluated. The maximum number of estimators was set at 1,500, and the effective number was determined through early stopping.

For XGBoost, learning rates between 0.02 and 0.08, depths between 5 and 8, minimum child weights between 2 and 10, L2 regularization between 1 and 10, and gamma values between 0 and 1 were evaluated. A maximum of 1,800 estimators was established, together with observation and feature subsampling proportions of 0.90.

Optimization aimed to maximize Average Precision for the low-birth-weight class. After the 15 trials, the best hyperparameters were confirmed using the available 2015–2021 and 2022 records within the historical sample. The final number of trees was determined through early stopping using 2022 as the validation set and subsequently remained fixed during the algorithm comparison.

Cross-validation and statistical comparison

Algorithm comparison was performed using ten stratified partitions of the 2023 records. Observation assignments to folds were generated once using shuffling and a random seed of 42. The same partitions were reused for CatBoost, LightGBM, and XGBoost to ensure paired comparisons.

In each iteration, the training set consisted of the historical 2015–2022 records plus nine-tenths of the 2023 data, while the remaining tenth of the 2023 data served as the validation set. In this way, each 2023 observation received a single out-of-fold prediction.

The primary metric was Average Precision because of class imbalance and the specific interest in evaluating the ability to identify low-birth-weight cases. Metrics based on precision and sensitivity are particularly informative when the positive class has a low prevalence (Saito & Rehmsmeier, 2015).

Complementary metrics included ROC-AUC, log loss, Brier score, precision, sensitivity, specificity, F1-score, balanced accuracy, Matthews correlation coefficient, and negative predictive value.

For each algorithm, results obtained across the ten folds were summarized using mean, standard deviation, median, interquartile range, minimum and maximum values, and 95% confidence intervals.

The distribution of Average Precision was examined using the Shapiro–Wilk test. Global comparison among the three algorithms was performed using the nonparametric Friedman test, and effect size was estimated using Kendall's W coefficient.

When subsequent paired comparisons were required, the Wilcoxon signed-rank test was applied. Significance values were adjusted using the Holm procedure to control for multiple comparisons, with a significance level of 0.05. These procedures are commonly used in comparative studies of algorithms evaluated under common data partitions (Demšar, 2006; García et al., 2010).

Because the ten results were derived from partitions of the same temporal cohort rather than from ten completely independent datasets, the statistical tests were interpreted as complementary comparative evidence within the experimental design and not as a substitute for independent temporal evaluation.

Final model, calibration, and threshold selection

Following the experimental and statistical comparison, CatBoost was selected as the final algorithm. The model was retrained using the complete set of 3,726,200 records from the 2015–2022 period. This stage again used the entire historical dataset rather than the 600,000-observation sample employed during initial development.

Hyperparameters were fixed before evaluation on subsequent temporal periods. The final model used 630 iterations, a depth of 7, a learning rate of 0.0422639, L2 regularization of 6.59964, random strength of 1,

64 discretization borders, a maximum categorical interaction complexity of 1, Bernoulli bootstrap, and a subsampling proportion of 0.80.

Missing numerical variables were imputed using medians obtained exclusively from the 2015–2022 records. The resulting values were 39 weeks for gestational duration, two for the number of living children, and two for the number of maternal pregnancies.

The weight assigned to the positive class was recalculated using the entire historical dataset and reached 14.9762.

Probabilities produced by CatBoost were subsequently calibrated using the 410,422 records from 2023 through Platt scaling. This procedure fitted a logistic regression model on transformed model probabilities in order to improve the correspondence between estimated probabilities and the observed event frequency.

Calibration is particularly important when model probabilities are intended to be interpreted as risk measures rather than merely as values used to rank observations (Niculescu-Mizil & Caruana, 2005; Van Calster et al., 2019).

The classification threshold was also not conventionally set at 0.50. Because the objective was to reduce false negatives, a minimum sensitivity of 70% was predefined. Among all thresholds meeting this criterion in the 2023 records, the threshold with the highest specificity was selected. This procedure resulted in a final threshold of 0.137972.

Both the calibrator and the operating threshold were determined exclusively using 2023 information and were subsequently applied without refitting to the 2024 and 2025 records.

Temporal evaluation and model stability

Final model performance was initially evaluated using the 2024 records. These data were not involved in hyperparameter optimization, final training, calibration, or threshold selection.

Average Precision, ROC-AUC, log loss, and Brier score were calculated on this dataset as threshold-independent metrics. Using the previously established threshold, precision, sensitivity, specificity, F1-score, balanced accuracy, Matthews correlation coefficient, negative predictive value, and the confusion matrix were also calculated.

To quantify uncertainty associated with the metrics, 95% confidence intervals were estimated using 300 stratified bootstrap resamples. In each repetition, class proportions were preserved and the corresponding metrics were recalculated.

Subsequently, the same model, calibration procedure, and threshold were applied without any refitting to the 2025 records available through November 30. This second evaluation made it possible to assess the temporal stability of discrimination, calibration, and classification performance in a period subsequent to the main test dataset.

Model explainability

The final model was interpreted using global importance measures and SHAP values. This procedure is based on Shapley values and makes it possible to distribute among the features the difference between an individual prediction and the model's expected value (Lundberg & Lee, 2017).

To avoid performing interpretation exclusively on observations used during training, SHAP values were calculated using a stratified sample of 5,000 births from the independent 2024 temporal dataset.

The global importance of each predictor was obtained from the mean absolute value of its SHAP contributions. Thus, variables with higher mean values were considered to have greater participation in the predictions generated by the model.

Positive SHAP values were interpreted as contributions shifting predictions toward a higher probability of low birth weight, whereas negative values represented contributions shifting predictions toward a lower probability of the event.

For numerical predictors, Spearman correlations were additionally calculated between observed feature values and their corresponding SHAP contributions. For categorical variables, contributions were summarized by category to identify patterns associated with increases or decreases in predicted risk.

Explainability findings were interpreted exclusively in predictive terms. SHAP values make it possible to examine how variables contribute to model predictions but do not demonstrate causal relationships between predictors and low birth weight.

Sensitivity analysis

Because of the high predictive importance observed for gestational duration, a complementary descriptive analysis of low-birth-weight prevalence by gestational week was conducted using the 2023 records.

For each gestational week, the percentage of low-birth-weight births was calculated relative to the total number of births recorded during that same week. This analysis allowed direct examination of the observed behavior of the outcome across different gestational durations, independently of the explanations produced by the model.

Finally, a sensitivity analysis was performed by ablating gestational duration. A second CatBoost model was trained after removing only this variable while retaining all other predictors and the same hyperparameters used in the primary model. No additional hyperparameter optimization was performed.

The performance of the reduced model was compared with that of the complete model using Average Precision and ROC-AUC. The reduction observed in these metrics was used to estimate the extent to which predictive performance depended on the information provided by gestational duration. This analysis was used solely as a sensitivity procedure and not to present the reduced model as an alternative to the final classifier.

3. RESULTS

3.1 Construction of the Analytical Dataset

The initial dataset consisted of 4,874,510 birth records. During data cleaning, 1,364 observations were excluded due to inconsistencies in birth weight: 1,315 had non-positive values, two recorded weights below 300 g, and 47 exceeded 6,000 g. Consequently, the final analytical dataset comprised 4,873,146 births corresponding to the 2015–2025 period.

As shown in Table 1, the historical 2015–2022 dataset included 3,726,200 records, of which 233,235 corresponded to low birth weight, with a prevalence of 6.26%. In 2023, prevalence increased to 7.00%, while in 2024 and 2025 it reached 6.90% and 6.77%, respectively. These proportions indicated a relatively stable frequency of the outcome across the periods used for validation and temporal evaluation.

As shown in Table 1, the historical 2015–2022 dataset included 3,726,200 records, of which 233,235 corresponded to low birth weight, with a prevalence of 6.26%. In 2023, prevalence increased to 7.00%, while in 2024 and 2025 it reached 6.90% and 6.77%, respectively. These proportions indicated a relatively stable frequency of the outcome across the periods used for validation and temporal evaluation.

Table 1. General characteristics of the analytical dataset

Partition	Period	Records	Low birth weight	Weight ≥ 2,500 g	Low-birth-weight prevalence

Historical development	2015–2022	3,726,200	233,235	3,492,965	6.26%
Operational validation	2023	410,422	28,735	381,687	7.00%
Temporal evaluation	2024	390,047	26,921	363,126	6.90%
Temporal stability	2025	346,477	23,456	323,021	6.77%
Total	**2015–2025**	**4,873,146**	**312,347**	**4,560,799**	**6.41%**

3.2 Evaluation of Machine Learning Models

The initial optimization produced similar results across the three algorithms. During confirmation using 2022 data, CatBoost achieved an Average Precision (AP) of 0.655, followed by LightGBM at 0.650 and XGBoost at 0.647. The models were subsequently compared using the same ten partitions of the 2023 records.

CatBoost achieved the highest mean AP, at 0.680 ± 0.005, followed by XGBoost at 0.678 ± 0.006 and LightGBM at 0.678 ± 0.005. CatBoost also achieved the highest mean ROC-AUC, at 0.910. Although the absolute AP differences were small, CatBoost achieved the best result in nine of the ten folds compared with each of the other algorithms and obtained the lowest mean rank, as shown in Table 2.

Table 2. Algorithm performance in ten-fold validation on 2023 data

Algorithm	Mean AP	SD	95% CI for AP	Mean ROC-AUC	Mean rank
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

The Friedman test identified statistically significant overall differences among the algorithms ($\chi^2 = 9.60$; $p = 0.0082$), with a Kendall's W of 0.48, corresponding to a moderate effect size. In pairwise comparisons, CatBoost showed significantly higher AP than both LightGBM and XGBoost after Holm correction (adjusted $p = 0.0117$ for both comparisons). In contrast, no significant differences were found between LightGBM and XGBoost (adjusted $p = 0.5566$). Based on these results, CatBoost was selected as the final algorithm.

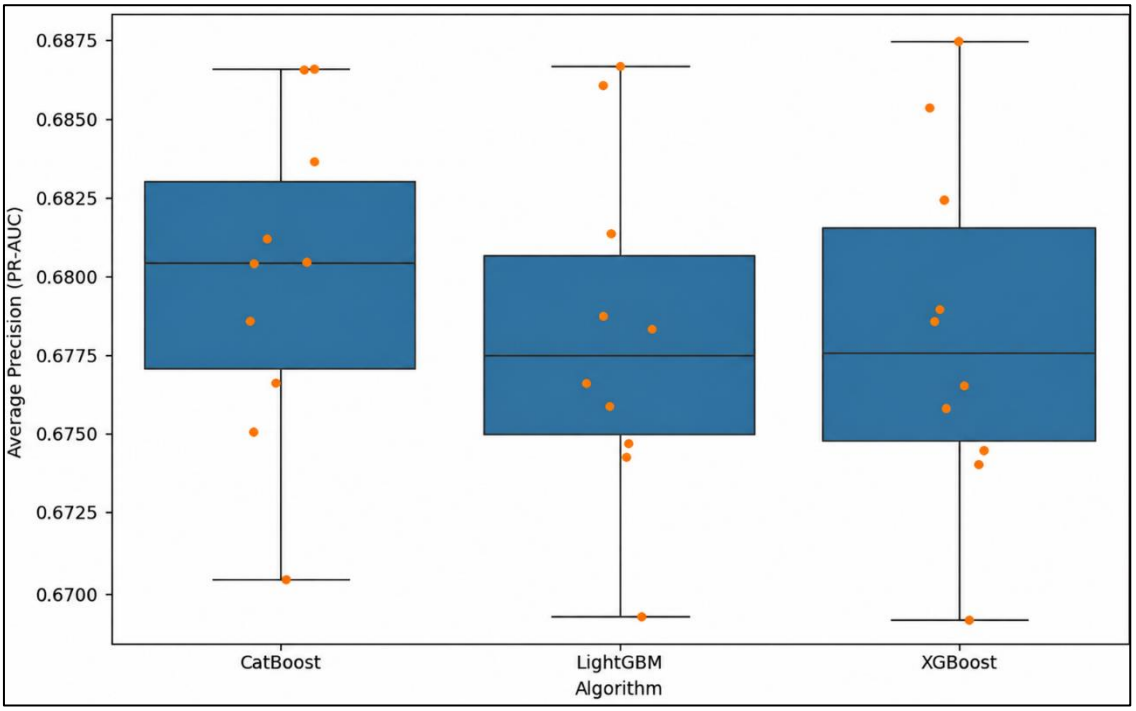

Figure 2. Distribution of Average Precision obtained by CatBoost, LightGBM, and XGBoost across the ten validation folds

3.3 Final Model and Operating Threshold Selection

The selected CatBoost model was retrained using all 3,726,200 births from the 2015–2022 period. The final configuration used 630 iterations, a depth of 7, and a learning rate of 0.0423. Its probabilities were subsequently calibrated through Platt scaling using the 410,422 records from 2023.

Threshold selection prioritized a minimum sensitivity of 70%. Under this criterion, the final threshold was 0.137972, with a sensitivity of 70.00%, specificity of 94.66%, precision of 49.69%, and F1-score of 0.581 in 2023. Although a threshold of 0.321889 yielded a higher F1-score of 0.632, sensitivity decreased to 58.52%; therefore, it was not selected as the operating threshold.

Calibration modified the probability scale without altering the discriminatory ranking produced by the model. The correspondence between estimated probabilities and observed frequencies approached the reference line after calibration, and this pattern was maintained during subsequent temporal evaluation.

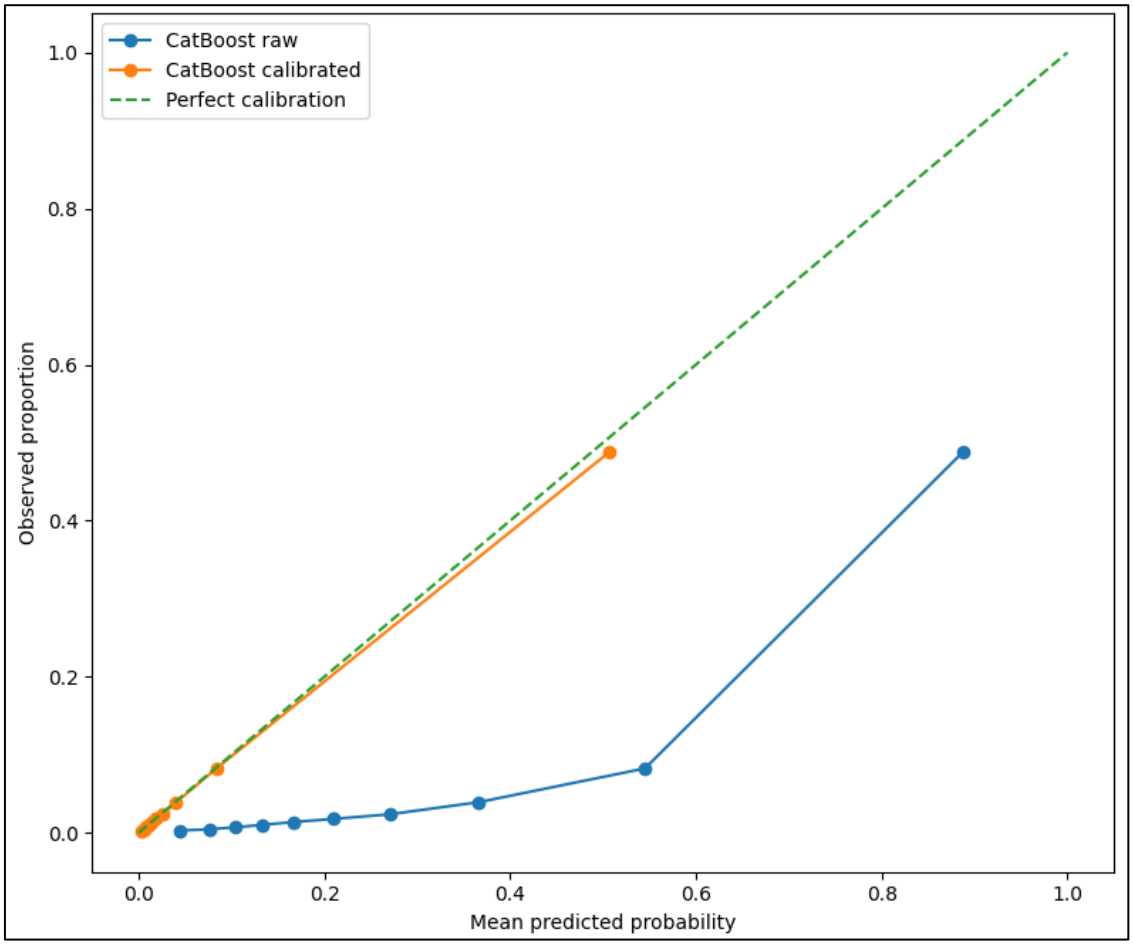

Figure 3. Calibration of CatBoost model probabilities in the 2024 temporal evaluation

3.4 Temporal Evaluation and Model Stability

The final model, calibrator, and previously established threshold were applied without adjustment to the 2024 records. In this independent dataset, CatBoost achieved an AP of 0.681, a ROC-AUC of 0.913, and a balanced accuracy of 0.827. Sensitivity was 71.12% and specificity was 94.31%, while precision and F1-score reached 48.09% and 0.574, respectively.

Of the 26,921 low-birth-weight cases recorded in 2024, the model correctly identified 19,145 and failed to detect 7,776. Among the 363,126 births weighing at least 2,500 g, it correctly classified 342,462 and generated 20,664 false-positive results. The negative predictive value reached 97.78%, indicating a high proportion of correctly identified negative results.

Application of the unchanged model to the 2025 records showed similar behavior. AP increased slightly to 0.685 and ROC-AUC to 0.915; sensitivity reached 71.85%, while specificity remained at 94.21%. The F1-score was 0.571 and balanced accuracy was 0.830. Brier score and log loss also remained stable, as shown in Table 3.

Table 3. Temporal performance of the final CatBoost model

Metric	Validation 2023	Evaluation 2024	Stability 2025
Average Precision	0.6806	0.6811	0.6847
ROC-AUC	0.9104	0.9130	0.9152
Precision	0.4969	0.4809	0.4741

Sensitivity	0.7000	0.7112	0.7185
Specificity	0.9466	0.9431	0.9421
F1-score	0.5812	0.5738	0.5713
Balanced accuracy	0.8233	0.8271	0.8303
MCC	0.5534	0.5478	0.5470
Brier score	0.0357	0.0352	0.0344
Log loss	0.1378	0.1356	0.1326

Overall, the variations observed between 2023, 2024, and 2025 were small. Discrimination and sensitivity increased slightly in subsequent periods, while precision and F1-score showed minor decreases. These findings demonstrated temporal stability of predictive performance under the previously defined operating threshold.

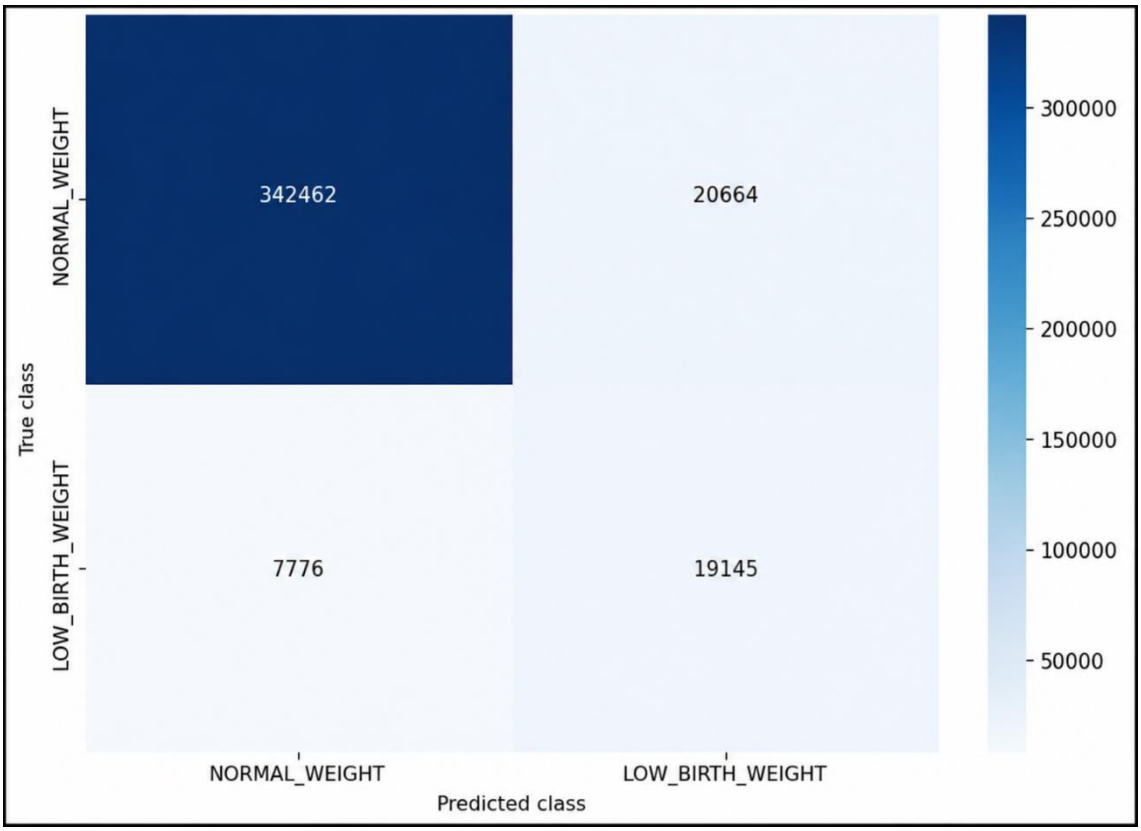

Figure 4. Confusion matrix of the CatBoost model in the 2024 temporal evaluation

3.5 Explainability of Predictions

The SHAP analysis showed a markedly predominant contribution of gestational duration. This variable reached a mean absolute SHAP value of 1.758, considerably higher than that observed for type of delivery (0.170), newborn sex (0.168), IPRESS category (0.138), maternal educational level (0.120), and number of maternal pregnancies (0.119).

CatBoost's native feature importance showed a consistent pattern: gestational duration accounted for 66.07% of total importance, followed at a considerable distance by type of delivery at 6.63%. The remaining

variables had individual contributions below 4%, indicating a strong dependence of the model on information related to gestational duration.

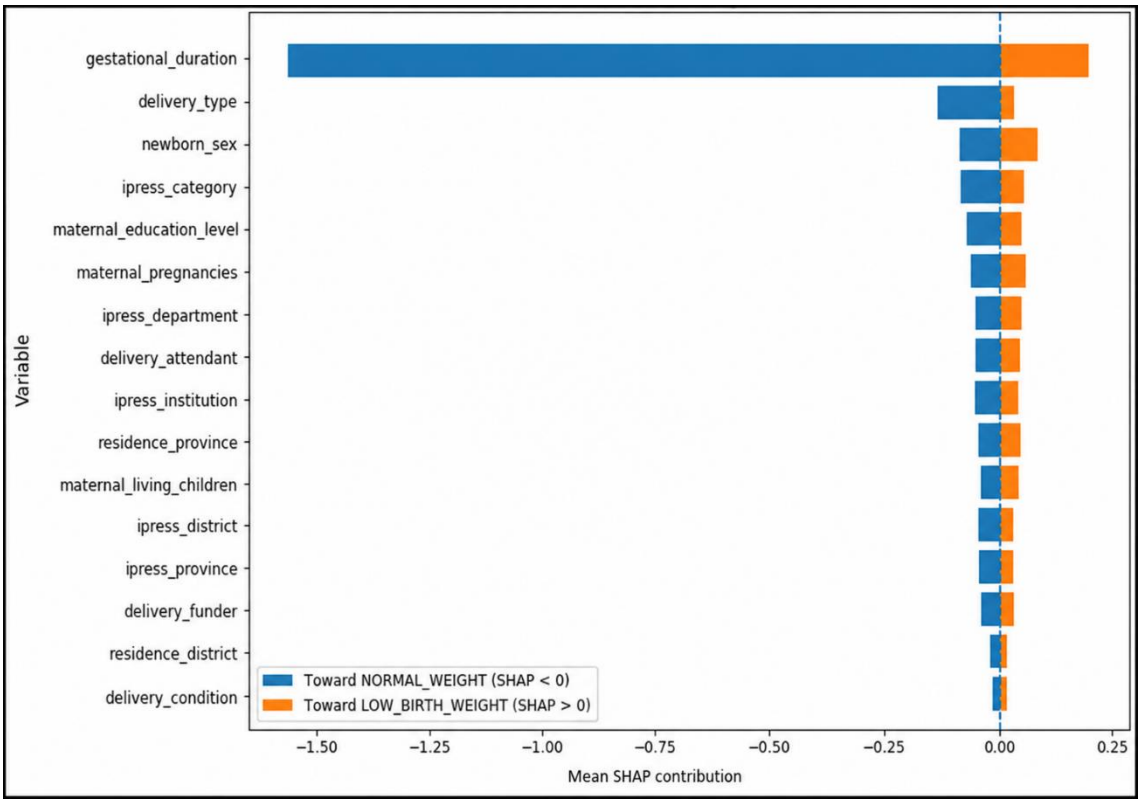

Figure 5. Global contribution of predictors to low-birth-weight estimates using SHAP values

Among numerical variables, gestational duration showed a Spearman correlation of −0.967 between its observed values and SHAP contributions. Negative correlations were also observed for the number of maternal pregnancies ($\rho = -0.904$) and number of living children ($\rho = -0.724$). Within the model, these results indicated that higher values of these characteristics tended to shift predictions toward a lower probability of low birth weight, although these relationships should not be interpreted as causal effects.

The relationship directly observed in the 2023 records was consistent with the pattern identified by the model. Low-birth-weight prevalence reached 57.47% at 35 weeks, decreased to 31.56% at 36 weeks, and to 11.49% at 37 weeks. It subsequently declined to 3.44% at 38 weeks, 1.60% at 39 weeks, and 1.03% at 40 weeks.

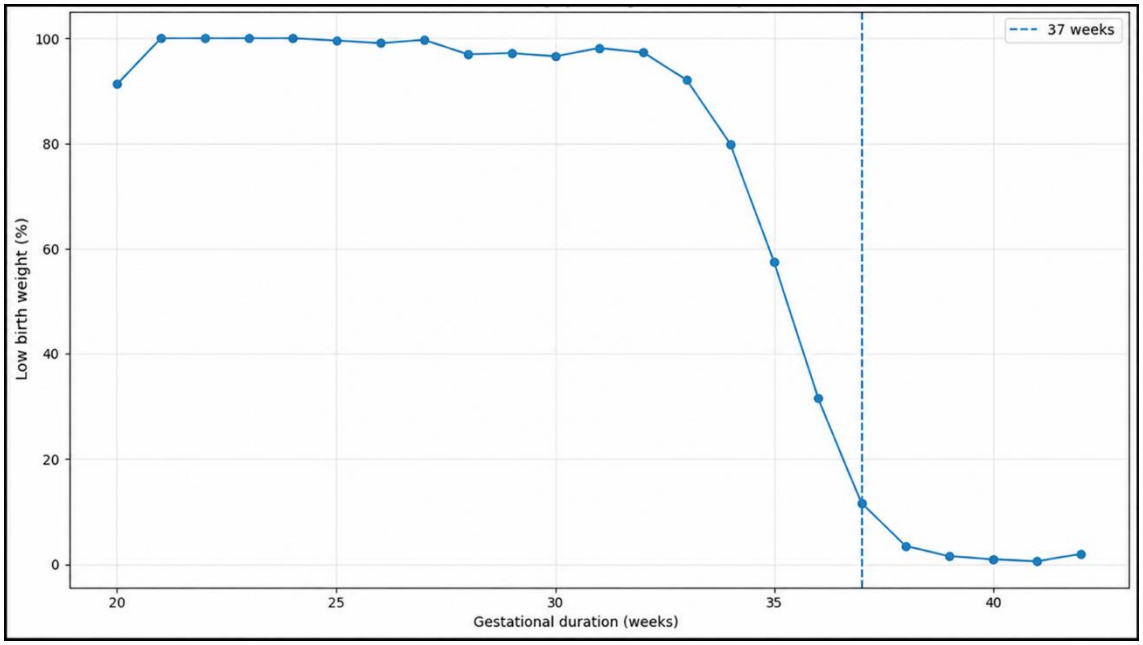

Figure 6. Observed relationship between gestational duration and low-birth-weight prevalence in 2023

3.6 Sensitivity Analysis of Gestational Duration

Because of the predominance of gestational duration in the model explanation, a sensitivity analysis was performed by removing this variable while leaving the remaining predictors and the hyperparameters of the final CatBoost model unchanged.

As shown in Table 4, excluding gestational duration produced a substantial reduction in performance across all three periods evaluated. In 2024, AP decreased from 0.681 to 0.295, while ROC-AUC decreased from 0.913 to 0.765. A similar pattern was observed in 2025, when AP decreased from 0.685 to 0.281 and ROC-AUC from 0.915 to 0.761.

Table 4. Sensitivity analysis of the model with and without gestational duration

Period	AP with gestation	AP without gestation	AP difference	ROC-AUC with gestation	ROC-AUC without gestation	ROC-AUC difference
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

The magnitude of this reduction confirmed that gestational duration constituted the main source of predictive information for the model. Nevertheless, the performance retained after its removal, particularly a ROC-AUC close to 0.76, showed that the remaining characteristics still preserved some ability to discriminate low-birth-weight births.

3.7 Discussion

The results showed that national records from the Peruvian Ministry of Health can be used to develop machine learning models capable of discriminating low birth weight on a large scale. CatBoost achieved the best performance among the algorithms compared and maintained ROC-AUC values above 0.91 and Average Precision values close to 0.68 between 2023 and 2025. Although the absolute difference compared with LightGBM and XGBoost was small, CatBoost consistently achieved better results in nine of the ten folds, which explains the significance observed in the pairwise comparisons. This behavior is consistent

with recent studies that have reported competitive performance of boosting algorithms for low-birth-weight prediction. Ranjbar et al. (2023), using 8,853 births, identified XGBoost among the best-performing models and reported a sensitivity of 0.69 and an F1-score of 0.77. Similarly, Victor et al. (2025) compared XGBoost, CatBoost, LightGBM, and Random Forest and observed ROC-AUC values close to 0.94 for the leading algorithms. Taken together, these findings suggest that differences among boosting methods may be small and that model selection should consider not only the maximum value of a single metric but also stability and performance on subsequent data.

One of the main findings of this study was the model's temporal stability. After being trained using records from 2015–2022 and using 2023 for calibration and operating-threshold definition, performance did not decline in subsequent periods. ROC-AUC increased from 0.910 in 2023 to 0.913 in 2024 and 0.915 in 2025, while Average Precision remained between 0.681 and 0.685. Sensitivity, specificity, and Brier score also showed only minor variation. This result strengthens the evidence of generalization within the national registry system because model performance was assessed in chronologically subsequent cohorts rather than solely through a random split of the same population. Nevertheless, this temporal stability should not be considered equivalent to external validation because all periods originated from the same national information system. Therefore, future implementation would require assessment of model transportability in other populations or healthcare settings.

The selection of a threshold aimed at achieving a minimum sensitivity of 70% made it possible to identify approximately seven out of ten low-birth-weight births in 2024 and 2025 while maintaining specificity above 94%. However, precision remained around 48%, implying a substantial proportion of false alerts. This behavior reflects a deliberate decision to prioritize case detection and reduce false negatives rather than exclusively maximizing metrics such as F1-score. Therefore, the threshold of 0.137972 should be interpreted as an operating threshold established for this experiment rather than as a universal clinical criterion. Before any potential use in healthcare services, the consequences of false alerts, the resources required for their review, and the clinical benefit of intervening in identified cases would need to be evaluated. Furthermore, its implementation would require healthcare professionals to understand the generated probabilities, the model's limitations, and the scientific evidence supporting its use. This consideration is relevant because the skills to search for, evaluate, and utilize scientific information constitute a key component of evidence-based practice among medical professionals (Valles-Coral et al., 2025).

The most important interpretive finding was the predominance of gestational duration. This variable contributed considerably more than the remaining predictors and showed a strong inverse relationship with estimated low-birth-weight risk. The pattern was consistent with the observed data, as prevalence decreased from 57.47% at 35 weeks to 1.03% at 40 weeks. Recent literature reports concordant findings. Ranjbar et al. (2023) identified gestational age and a previous history of low birth weight as the main predictors in their model. Victor et al. (2025) similarly found gestational age to be the most influential characteristic in boosting models, while Wu et al. (2025) identified gestation below 37 weeks among the most relevant factors using XGBoost and SHAP. In Bangladesh, Sultana et al. (2025) also found pregnancy duration to be the predictor with the greatest influence on low-birth-weight classification. The consistency across different settings supports the plausibility of the pattern identified in the Peruvian records.

The ablation analysis reinforced this interpretation. When gestational duration alone was removed, Average Precision decreased from 0.681 to 0.295 in 2024 and from 0.685 to 0.281 in 2025; ROC-AUC fell from approximately 0.91 to 0.76. Therefore, a substantial proportion of the model's predictive capacity depended on this variable. This finding represents both a strength and a limitation. On the one hand, it confirms that the model identifies a clinically meaningful signal closely related to low birth weight; on the other, it prevents the model from being considered an early prenatal prediction tool because final gestational duration is only known near the time of birth. Consequently, the model is more appropriately

interpreted as a tool for risk stratification and interpretation within perinatal records rather than as an early-pregnancy warning system. A model intended for early prediction would require restricting predictors to information available before the outcome occurs.

Finally, the main strengths of this study include the use of more than 4.8 million national records, comparison of algorithms using the same partitions, temporal evaluation across two subsequent periods, probability calibration, and the incorporation of SHAP and ablation analyses. However, the results depend on the quality of the administrative records and do not include several clinical, nutritional, or behavioral variables used in recent studies, such as maternal body mass index, gestational hypertension, smoking, or detailed characteristics of prenatal care. In addition, SHAP contributions should be interpreted as predictive signals rather than causal effects. Overall, the findings show that machine learning can leverage routine MINSA records to discriminate low birth weight with temporally stable performance while providing interpretable information about the predictors used. The principal contribution of this study lies in combining national-scale data, algorithmic comparison, temporal validation, and explainability, although eventual implementation will require external validation and a more precise definition of the clinical point at which the model is intended to be used.

CONCLUSIONS

The integration of national birth records from the Peruvian Ministry of Health enabled the development and validation of an explainable machine learning model for predicting low birth weight. CatBoost was selected because it achieved the best comparative performance against LightGBM and XGBoost and maintained stable results during the temporal evaluations conducted in 2024 and 2025, demonstrating adequate generalization within the national registry system. Gestational duration emerged as the model's most important predictor, a finding supported both by the SHAP analysis and by the reduction in performance observed when this variable was removed. However, its strong contribution indicates that the model is more suitable for risk stratification and interpretation within perinatal records than for early prenatal prediction. Accordingly, the proposed approach demonstrates the potential of MINSA administrative data to support large-scale identification and analysis of low birth weight, provided that model predictions are used as an aid for surveillance and health analysis rather than as a substitute for clinical assessment.

FINANCING

The authors received no funding to conduct this study.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest related to the subject matter of this work.

AUTHORSHIP CONTRIBUTION

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.

REFERENCES

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