Master in Management + Analytics

Modelo Predictivo de mortalidad en neonatos de muy bajo peso en la red Neocosur (Sudamérica)

Autoría: Camerlingo, Sebastián
Año: 2024

UNIVERSIDAD TORCUATO DI TELLA

Master in Management + Analytics

Modelo predictivo de mortalidad en neonatos de muy bajo peso en la red Neocosur (Sudamérica)

Sebastián Camerlingo

Diciembre 2024

Tutor: Ramiro Gálvez

Resumen

Introducción: La utilización de scores de mortalidad y morbilidades mayores es una práctica frecuente en Neonatología, proporcionando una herramienta valiosa para la toma de decisiones clínicas y la estratificación del riesgo. Debido a los cambios temporales en los factores de riesgo, la incorporación de nuevas técnicas metodológicas y los avances en la práctica clínica, los modelos pronósticos requieren de una actualización periódica para mantener su relevancia y eficacia. Este trabajo tiene como objetivo actualizar el score pronóstico de la red Neocosur, utilizando herramientas analíticas modernas y evaluando su impacto en la performance y su utilidad clínica.

Métodos: Se construyeron y evaluaron tres modelos pronósticos: el modelo actual utilizado como benchmark, un modelo de regresión logística con penalización (Lasso) y un modelo basado en Light Gradient Boosting Machine (LGBM). Los modelos fueron desarrollados utilizando una base de datos multicéntrica de pacientes neonatales pertenecientes a la red Neocosur. Para evaluar la performance, se analizaron la discriminación (AUROC), la calibración (parámetros de Intercept y Slope) y la equidad algorítmica en diferentes subgrupos poblacionales. Además, se realizó un análisis de utilidad clínica mediante curvas de beneficio neto, considerando diferentes umbrales de riesgo predicho.

Resultados: El modelo LGBM demostró una superioridad significativa en términos de discriminación en comparación con el modelo benchmark (AUROC = 0.89 vs. 0.87, respectivamente). Asimismo, el análisis de utilidad clínica mostró un beneficio neto mayor para el modelo LGBM en todo el rango de umbrales evaluados. Sin embargo, se observó un leve deterioro en la calibración (Intercept y Slope) del modelo LGBM en comparación con el benchmark (Intercept: -0.17 vs. 0.99; Slope: 1.26 vs. 1.00), lo que podría impactar en la interpretación de predicciones individuales. El modelo basado en Lasso presentó un desempeño intermedio entre el benchmark y el LGBM.

Conclusión: El modelo LGBM representa una mejora significativa en la performance y la utilidad clínica frente al modelo original de la red Neocosur. No obstante, la menor calibración observada sugiere la necesidad de abordar estrategias de recalibración antes de su implementación clínica. Esta actualización resalta el potencial de los métodos basados en machine learning para optimizar herramientas pronósticas, aunque también subraya la importancia de evaluar integralmente las dimensiones de performance y su aplicabilidad en la práctica cotidiana.

Índice

Índice de tablas

Índice de figuras

1 Introducción

1.1 Contexto

La utilización de scores de mortalidad y morbilidades mayores es una práctica frecuente en neonatología. Sus principales objetivos son comparar la performance de centros en cuanto a sus resultados y a la vez predecir desenlaces graves con el fin de tomar acciones preventivas. La población estándar que se evalúa es la de los recién nacidos de peso menor a 1500 g o Recién Nacidos de Muy Bajo Peso al Nacer (RNMBP). Se denomina a esta población como centinela, ya que la calidad de un centro neonatal puede medirse con este grupo de pacientes que globalmente tienen un comportamiento similar en todo el mundo. Además, la vigilancia de resultados en cada centro expresa no solo una medida de calidad, sino una oportunidad de implementar políticas de mejora (benchmarking) en aquellos centros en los que la morbimortalidad es superior a la esperada.

La red Neocosur es una red que recolecta desde 2002 datos de los RNMBP definidos como peso al nacimiento (PN) menor a 1500 g nacidos en 36 centros de nivel III públicos y privados de 5 países de Sudamérica: Argentina, Chile, Paraguay, Perú y Uruguay. Actualmente, cuenta con 34,000 registros. Se trata de una red de centros muy heterogénea con neonatos y recursos técnicos y profesionales de características muy dispares. Para poder comparar estos centros entre sí año a año, evaluar tendencias y determinar riesgos y posibles mejoras, la red Neocosur desarrolló en 2005 un score predictivo de mortalidad propio [Marshall et al., 2005].

Uno de los motores fundamentales de la red para desarrollar el score, fue contar con una herramienta que pudiera reflejar la realidad de los RNMBP nacidos en países del cono sur americano que son, en su mayoría, países de medianos a bajos ingresos. Hasta ese momento, los scores de riesgo perinatal utilizados provenían de países desarrollados [Fenton and Kim, 2013, Villar et al., 2015, Olsen et al., 2010], con tasas de mortalidad neonatal más bajas y características fetales y maternas diferentes.

De manera consistente, una revisión sistemática realizada en 2011 [Medlock et al., 2011] identificó que la mayoría de los modelos de predicción de mortalidad en RNMBP incluían variables postnatales no siempre disponibles en el momento del nacimiento, y reportaban una evaluación limitada de su calibración, dificultando su generalización a otros contextos.

Por este motivo, la validez externa de estos scores era relativa para los RNMBP de la región sudamericana. Un score de riesgo construido exclusivamente con neonatos originarios de países del cono sur y nacidos en centros neonatales con características diferentes a los de los otros países podría constituir una herramienta de suma utilidad, no solo para las unidades pertenecientes a la red Neocosur, sino también para otras unidades de países sudamericanos.

Una vez construido este score, sus aplicaciones han sido diversas a lo largo de los últimos 18 años. Una de las principales utilidades ha sido la estimación de la mortalidad observada sobre esperada o mortalidad ajustada para cada centro de la red [Marshall et al., 2012, Fernández et al., 2014]. Cada año se calcula el score de riesgo de cada neonato en cada centro y se suman todos los scores, estimando el valor total esperado del centro. Luego se le asigna un puntaje de 0 a cada neonato sobreviviente y 1 punto a cada fallecido y se suman para conformar el valor observado. Finalmente, se calcula la mortalidad ajustada como el cociente entre el valor observado y el esperado. Un valor de 1 implica que la

mortalidad de ese centro coincide con la esperada según el modelo predictivo de la red, un valor superior a 1 implica que ese centro presenta un exceso de mortalidad observada respecto a la esperada para la red y, por el contrario, uno inferior a 1 implica que el centro presenta una mortalidad menor a la esperada para la red.

El seguimiento año a año de la mortalidad ajustada en cada centro de la Red ha permitido detectar aquellos centros que requieren intervenciones para mejorar sus resultados y monitorear la performance de cada uno a lo largo del tiempo. Así mismo, al tratarse de un score de riesgo basado fundamentalmente en factores prenatales, permite atribuir el exceso de mortalidad al desempeño de cada centro cuando las condiciones basales son las mismas para todos los centros [Fernández et al., 2014, Marshall et al., 2012].

A continuación se definen los problemas clínicos y metodológicos a ser abordados en el presente trabajo.

1.2 Problema

Pese a la gran utilidad que este score ha tenido desde su creación, el modelo original presenta algunas limitaciones desde los puntos de vista tanto clínicos como metodológicos. Las mismas se explican a continuación.

1.2.1 Aspectos clínicos

El modelo predictivo actual que utiliza la red consiste en un modelo de regresión logística multivariado que incluye las variables edad gestacional en semanas completas, peso de nacimiento en gramos, score de Apgar al minuto (riesgo de mortalidad tomado al minuto de nacer; escala de 0-10, donde 10 refleja la ausencia de riesgo)[Simon et al., 2024], sexo (femenino/masculino), uso de esteroides prenatales (indicado para la maduración pulmonar en pacientes pretérmino) y diagnóstico de malformaciones congénitas mayores [Marshall et al., 2005]. Si bien las variables incluidas fueron las que mayor peso mostraron en el análisis multivariado original, existen otras variables relevantes que quedaron fuera del score construido en el año 2005.

Un caso particular es la condición de pequeño para edad gestacional (PEG). Esta variable, si bien mostró capacidad predictiva en el modelo original bivariado, luego fue descartada en el contexto multivariado. El punto crítico al respecto es que para clasificar a los RNMBP como PEG se utilizó como punto de corte al percentil 10 de la curva de crecimiento chilena [Alarcón et al., 2008]. Dadas las diferencias regionales, la utilización de este instrumento podría no tener la misma representatividad para neonatos de los otros países que forman parte de la red. Aplicando otras curvas de crecimiento que fueron construidas con una muestra más amplia y diversa, y cuyo uso es internacional [Fenton and Kim, 2013, Villar et al., 2015, Olsen et al., 2010], es posible que tanto los coeficientes como los intervalos de confianza para esta variable se modifiquen en contexto y sumen capacidad predictiva en un nuevo modelo. Asimismo, es importante destacar que el punto de corte en el percentil 10 ha mostrado una baja capacidad predictiva para mortalidad en varios estudios de redes multicentro, incluyendo uno reciente del grupo de trabajo de la misma Red Neocosur [García-Muñoz Rodrigo et al., 2021]. En este último, utilizando como referencia las curvas de crecimiento de Fenton e Intergrowth 21st [Villar et al., 2015], el punto de corte en el percentil 3 resultó ser un mejor predictor de mortalidad con ambas curvas de crecimiento.

Otro predictor potencialmente relevante es la presencia de infecciones prenatales (Ej. Corioamnionitis, requerimiento de tratamiento antibiótico, etc.) [Baguiya et al., 2021]. Sin embargo, esta condición no fue registrada inicialmente, y por lo tanto no hubo oportunidad de incluirla en el modelo original construido por la red Neocosur.

Por otra parte, es necesario destacar que no se tuvo en cuenta en el análisis la variabilidad atribuida a los centros. Este aspecto es de particular relevancia ya que los centros que forman parte de la red presentan desempeños muy heterogéneos en cuanto a resultados en mortalidad y morbilidades mayores [García-Muñoz Rodrigo et al., 2021].

Por último, la presencia de malformaciones congénitas mayores es un factor que ha demostrado aumentar significativamente la mortalidad [Lona Reyes et al., 2018]. Sin embargo, debido a la escasez de centros especializados que pueden atender estos casos, los pacientes a menudo son referidos a otras instituciones después de haber sido tratados inicialmente en un centro de menor complejidad. Esto provoca un aumento en la mortalidad observada en los centros receptores de alta complejidad, que, sin embargo, no se encuentra atribuida a la calidad del servicio brindado por dichas instituciones. La incorporación de pacientes con malformaciones mayores en los modelos predictivos es un tema que se encuentra actualmente en discusión en la red Neocosur.

Con respecto a los predictores previamente mencionados, se comenzó a incluirlos en el registro de la red a partir del año 2010.

1.2.2 Aspectos metodológicos

Los modelos de predicción clínica emplean combinaciones de variables para estimar el riesgo de resultados específicos a nivel individual. Evaluar rigurosamente el rendimiento de estos modelos resulta esencial, ya que un desarrollo inadecuado podría no solo comprometer su utilidad clínica, sino también exacerbar inequidades en la provisión de cuidados de salud o en los desenlaces posteriores [Collins et al., 2024]. La evaluación del rendimiento debe realizarse en conjuntos de datos representativos de las poblaciones objetivo previstas para su aplicación. Es frecuente observar que el desempeño predictivo de un modelo resulta óptimo en la cohorte de desarrollo, pero disminuye considerablemente al ser evaluado en un conjunto de validación independiente, incluso si este proviene de la misma población [Riley et al., 2024b].

Respecto a la validación de modelos predictivos, diversos aspectos metodológicos deben ser considerados desde la etapa de desarrollo [Collins et al., 2024]. En primer lugar, la utilización de muestras aleatorias como se realizó en el modelo original— puede no ser la estrategia más adecuada. La división en conjuntos de entrenamiento y validación sin un criterio racional (por ejemplo, segmentar por temporalidad o regionalidad en lugar de de manera aleatoria) introduce un riesgo potencial de sobreajuste, comprometiendo la validez externa del modelo [Riley et al., 2022].

Adicionalmente, la evaluación de la calibración adquiere un rol central en la validación del desempeño. Las recomendaciones más recientes [Riley et al., 2024a] enfatizan la importancia de visualizar curvas suavizadas que representen la relación entre los valores predichos y los observados, así como de reportar métricas específicas tales como la calibration-in-the-large" (calibración global) y la calibration slope" (pendiente de calibración) (véase Sección 3). No obstante, la calibración en el modelo original fue realizada mediante el test de Hosmer–Lemeshow, el cual tiene ciertas desventajas tales como que

está basado en el agrupamiento artificial de pacientes dentro de estratos de riesgo, brindando un p valor que no es informativo con respecto al tipo y magnitud del error y tiene poco poder estadístico, especialmente al incorporar no linealidades e interacciones [Harrell, 2016, Calster et al., 2019].

Otro tema de suma importancia a considerar durante la construcción del modelo es el del manejo de los datos faltantes [Collins et al., 2024]. En el modelo inicial de la red Neocosur, se descartaron los pacientes con datos faltantes (sólo utilizaron datos completos), lo cual, además de la pérdida de información resultante, conlleva el riesgo de sesgo en las estimaciones cuando la pérdida de información está relacionada con la variable resultado [Schafer and Yucel, 2002].

Por último, existe un incremento en el interés de evaluar la estabilidad de las predicciones generadas por los modelos en poblaciones con distintas características o particularidades [Riley and Collins, 2023]. Es importante resaltar que, incluso si las predicciones de un modelo calibran correctamente, aún puede haber variabilidad en las predicciones para subgrupos definidos por una covariable particular (que puede o no ser uno de los predictores incluidos en el modelo), lo cual es de suma relevancia para la práctica clínica. Por consiguiente, cualquier par de modelos de ejemplo puede diferir considerablemente en su riesgo estimado para este subgrupo, limitando la aplicación y alejándose del precepto de equidad ("Fairness") algorítmica.

En la última revisión sistemática realizada [Medlock et al., 2011] de modelos de predicción de mortalidad en neonatos prematuros extremos, se incluyeron 41 modelos publicados entre 1980 y 2010. Una limitación recurrente fue que muchos modelos utilizaron variables postnatales, como intervenciones terapéuticas o evolución clínica temprana, que no siempre están disponibles al momento de la toma inicial de decisiones. Esto restringe su aplicabilidad en contextos donde se requiere una predicción inmediata basada únicamente en variables perinatales. En cuanto a la evaluación del desempeño, solo 22 de los 41 estudios reportaron algún tipo de análisis de calibración. De estos, 11 utilizaron la prueba de Hosmer-Lemeshow. Sin embargo, la interpretación de estos resultados es limitada, ya que pocos estudios reportaron parámetros completos de calibración. Respecto a la discriminación, el AUROC promedio fue de 0.84, con un rango entre 0.68 y 0.93. En términos de calibración, dado que no se reportaron sistemáticamente los valores, no se pudo calcular un resumen cuantitativo confiable de estos parámetros. Por último, ninguno de los modelos reportó utilidad clínica mediante curvas de decisión o cuantificación del beneficio neto (ver Sección 3).

Más recientemente, se han desarrollado nuevos modelos con el objetivo de actualizar las herramientas de predicción de mortalidad neonatal en poblaciones específicas. Un ejemplo de ello es el modelo español propuesto en 2020 para RNMBP [Iriondo et al., 2020]. Este modelo empleó una base de datos multicéntrica de más de 10,000 neonatos e incorporó variables fácilmente disponibles en las primeras horas de vida. Si bien logró una discriminación adecuada (AUROC de 0.85 en la cohorte de validación), presentó limitaciones importantes en su aplicabilidad clínica, entre ellas, el uso de variables clínicas postnatales como requisito para completar la predicción y la falta de evaluación formal de la calibración más allá de la prueba de Hosmer-Lemeshow. Además, este modelo tampoco reportó mediciones de utilidad clínica formales.

Estos hallazgos refuerzan la necesidad de modelos adaptados a poblaciones y sistemas de salud específicos, que utilicen variables disponibles tempranamente y que reporten

adecuadamente medidas de desempeño tanto de discriminación como de calibración.

De los problemas tanto clínicos como metodológicos citados, se desprende la necesidad de actualizar el score pronóstico para la red, lo cual permitirá una mejora en la calidad de las predicciones y su consiguiente valor en la toma de decisiones.

1.3 Objetivos

El objetivo primario de este estudio es desarrollar un modelo de pronóstico de mortalidad individual al egreso en RNMBP evaluando un pool ampliado de variables independientes y comparar su capacidad predictiva (AUROC) y utilidad clínica (beneficio neto del modelo) con el modelo de regresión logística existente.

Como objetivos secundarios se plantean la evaluación de la calibración y fairness algorítmico de los modelos propuestos [Steyerberg and Vergouwe, 2014].

Se reporta la capacidad predictiva de los modelos en subgrupos preespecificados de pacientes pertenecientes a los distintos países y tipos de centro (público o privado).

2 Datos

Se utilizó la base de datos de la red Neocosur, que contiene información del período de tiempo comprendido entre enero de 2002 y agosto de 2022. Para la construcción del modelo se utilizó una muestra que corresponde al período comprendido entre agosto de 2010 y agosto del 2020, dado que a partir de esta fecha se comenzaron a recolectar otros predictores de relevancia para la tarea. La base completa cuenta con 24,000 observaciones y 273 variables tanto del neonato como de la madre en el período prenatal. Contiene características demográficas, clínicas, intervenciones e identificadores de centro de atención. Se realizó un análisis exploratorio a fin de identificar la cantidad de eventos, linealidad de variables continuas e interacciones relevantes dado el conocimiento científico del tema. La selección de variables fue predefinida de acuerdo a la literatura científica y el conocimiento de médicos expertos en el tema [Fenton and Kim, 2013, Villar et al., 2015, Olsen et al., 2010].

El evento de interés se especifica como la presencia de muerte antes del alta hospitalaria. Se utiliza la variable resultado como binaria, dado que el período entre el nacimiento y el evento es relativamente corto [Toso et al., 2022].

Tabla 1. Variables seleccionadas para el análisis

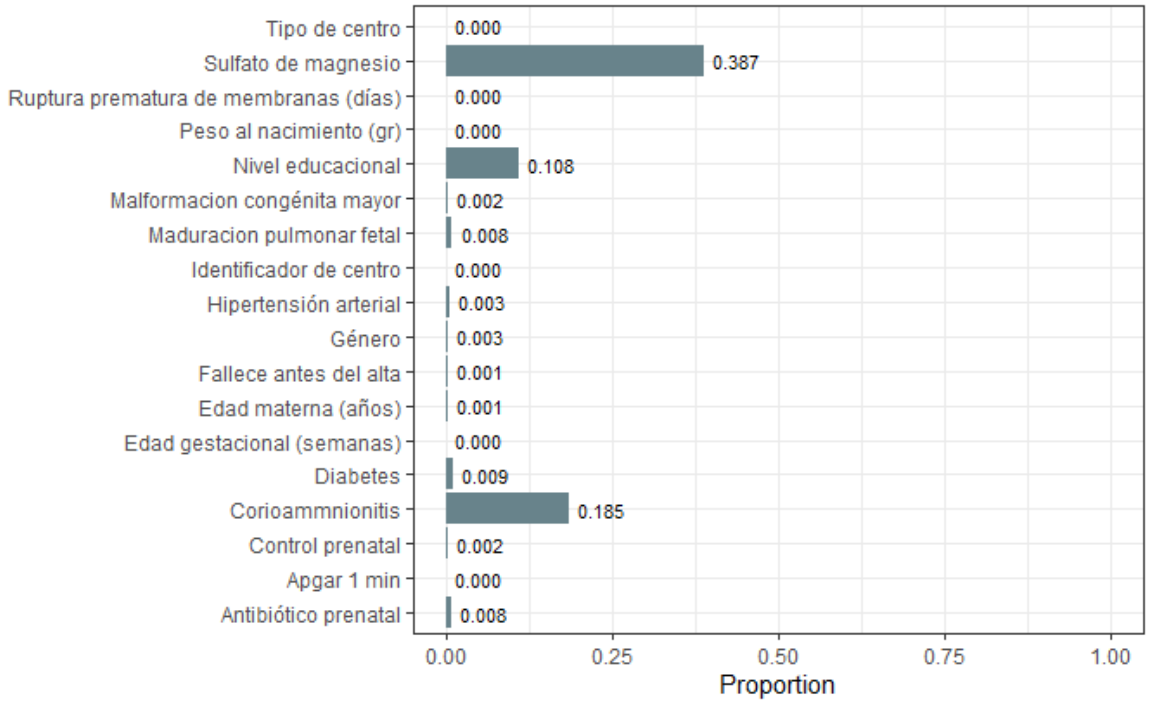
Variable	**Unidad de medición**	**Categorías (si aplica)**
Fallece antes del alta	Binaria	Sí / No (Outcome)
Edad gestacional	Continua	Semanas
Peso al nacimiento	Continua	Gramos
Apgar al minuto	Contínua	Score
Edad materna	Continua	Años
Malformación congénita mayor	Binaria	Presente / Ausente
Maduración pulmonar fetal	Binaria	Sí / No
Ruptura prematura de membranas	Continua	Días
Corticoides prenatales	Binaria	Sí / No
Género	Categórica	Masculino / Femenino
Nivel educacional materno	Categórica	Primaria incompleta / Secundaria incompleta / Terciaria / Universitaria
Hipertensión arterial materna	Binaria	Sí / No
Diabetes materna	Binaria	Sí / No
Corioamnionitis fetal	Binaria	Sí / No
Recibe antibiótico prenatal	Binaria	Sí / No
Recibe sulfato de magnesio	Binaria	Sí / No
Tipo de centro	Categórica	Público / Privado
Identificador de centro	Categórica	Valor nominal por centro

El análisis exploratorio de los datos tuvo como objetivo caracterizar tanto la magnitud del problema de datos faltantes como la variabilidad en la mortalidad observada en función de diferentes niveles de agrupamiento (países, centros y tipo de centro) y su evolución temporal. De esta manera, se pretende ayudar a la comprensión inicial del comportamiento de los datos y justificar las decisiones metodológicas posteriores en el desarrollo y validación de los modelos predictivos.

En primer lugar, se analizó la frecuencia de datos faltantes por variable [Figura 1], observándose un rango amplio que va desde la ausencia total hasta un 38 % de datos faltantes, en el caso del predictor Sulfato de Magnesio. Cabe señalar que el modelo tradicional utilizado como benchmark se basa mayormente en variables con bajo porcentaje de valores faltantes, lo cual plantea un desafío adicional para el desarrollo de modelos más complejos que podrían beneficiarse de la incorporación de nuevas variables, aunque estas presenten mayores niveles de incompletitud. Esta situación resalta la importancia de explorar estrategias de imputación que logren equilibrar precisión y viabilidad práctica

en contextos clínicos reales.

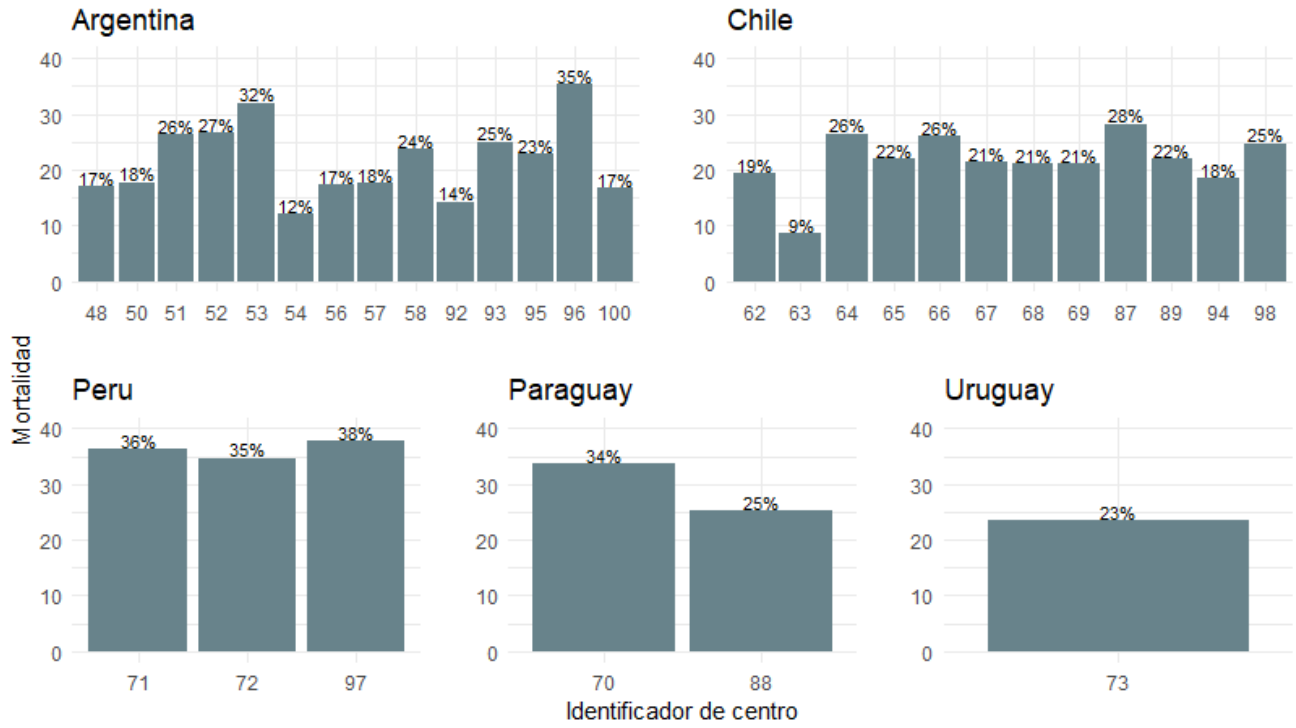
Figura 1. Distribución de datos faltantes



Fuente: elaboración propia a partir de la base de datos.

En cuanto a la variabilidad en la mortalidad observada entre los centros, se evidencia una marcada heterogeneidad, incluso dentro de los mismos países, y a pesar de la limitada contribución de algunos centros individuales [Figura 2]. Argentina y Chile concentran la mayor cantidad de centros y observaciones, mientras que Uruguay participa únicamente con un centro. Esta distribución asimétrica refleja diferencias estructurales y operativas relevantes entre países y entre centros dentro de un mismo país, lo que tiene implicancias metodológicas importantes. En este sentido, para análisis predictivos que busquen generalizar sus hallazgos a nivel de la red, resulta más adecuado considerar al centro como predictor en lugar del país, ya que permite capturar con mayor precisión la variabilidad observada en los desenlaces clínicos, como la mortalidad.

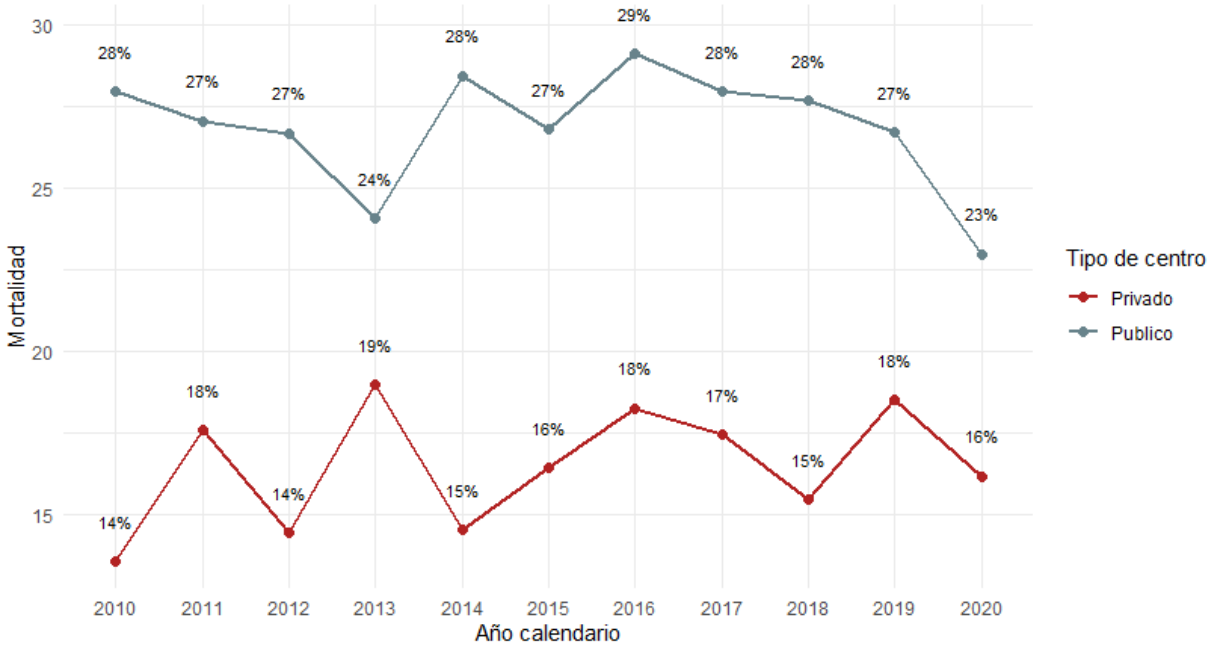
Figura 2. Distribución mortalidad por centro y país



Fuente: elaboración propia a partir de la base de datos.

La exploración temporal de la mortalidad según el tipo de centro (público o privado) [Figura 3] revela tendencias variables, pero en paralelo. En los centros públicos se observa una leve disminución de la mortalidad en los años recientes, aunque sin una trayectoria sostenida ni consistente. Por el contrario, en los centros privados se advierte un leve aumento en el mismo período, manteniéndose siempre por debajo de los niveles observados en el sistema público. Esta dinámica refuerza la hipótesis de diferencias estructurales persistentes entre tipos de centro y plantea interrogantes sobre los factores subyacentes que podrían explicar estas brechas.

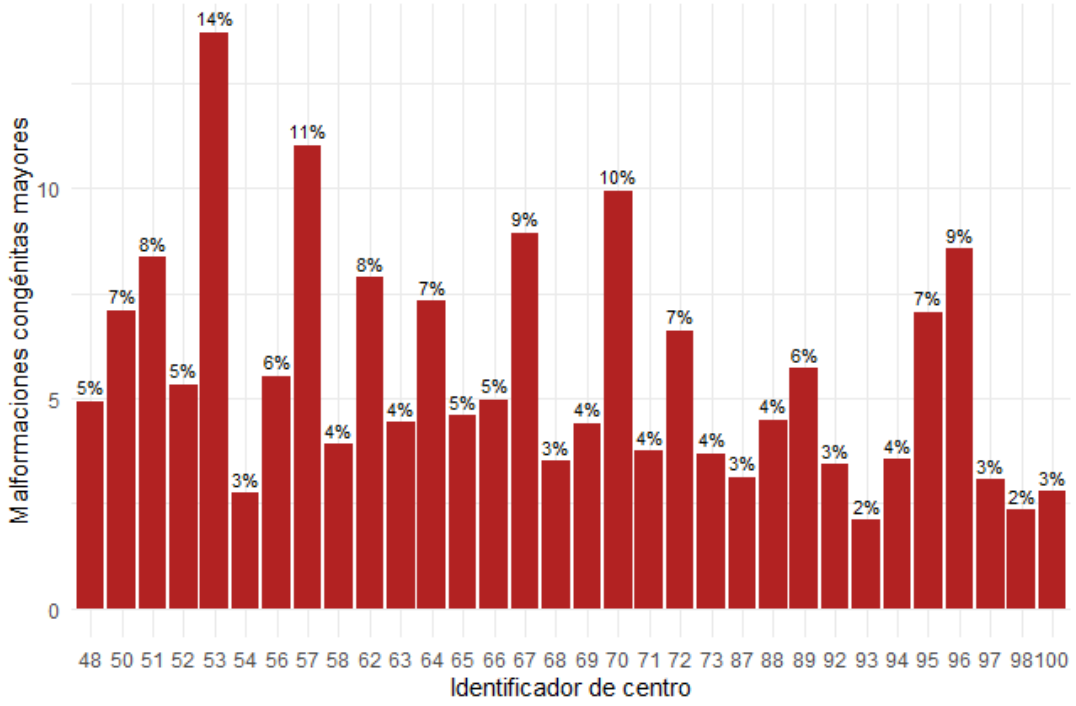
Figura 3. Distribución mortalidad por tipo centro y año



Fuente: elaboración propia a partir de la base de datos.

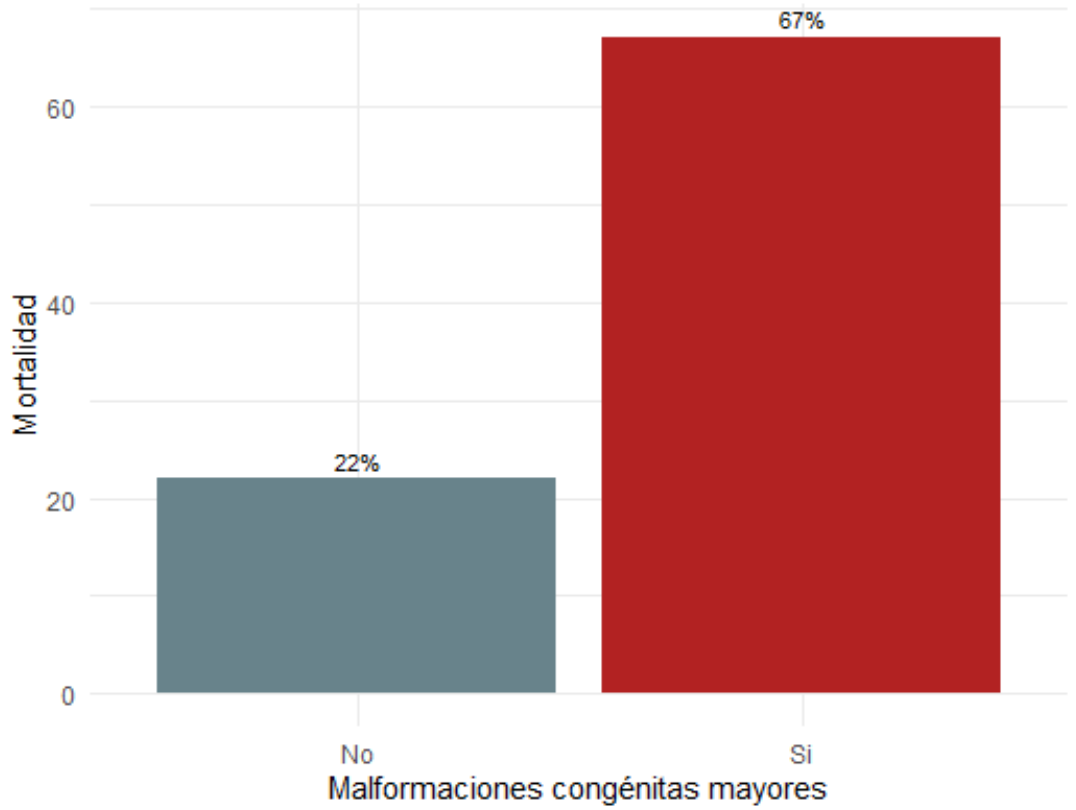
En este contexto, se analizó también la distribución de malformaciones congénitas mayores (MCM) entre centros [Figura 4], encontrándose diferencias de hasta 11 % en la prevalencia de estas condiciones. Dado que las MCM están fuertemente asociadas a un mayor riesgo de mortalidad, tal como lo ilustra la Figura 5, su consideración se vuelve fundamental al momento de construir modelos ajustados por riesgo que pretendan extrapolarse a diferentes entornos.

Figura 4. Distribución malformaciones congénitas mayores por centro



Fuente: elaboración propia a partir de la base de datos.

Figura 5. Mortalidad atribuída a malformaciones congénitas mayores



Fuente: elaboración propia a partir de la base de datos.

En suma, las diferencias observadas en los patrones de mortalidad según país, centro, tipo de institución y presencia de comorbilidades relevantes como las malformaciones congénitas mayores, subrayan la necesidad de incorporar esta heterogeneidad en los modelos predictivos. Ignorar estas fuentes de variabilidad podría limitar la capacidad del modelo para generalizar adecuadamente sus predicciones en distintos contextos clínicos y geográficos, afectando su utilidad práctica y validez externa.

3 Metodología

3.1 Conjunto de entrenamiento y tamaño muestral

Dada la necesidad de contar con conjuntos de entrenamiento, validación y testeo con la muestra disponible, es importante evaluar si el tamaño de la muestra es suficiente para determinar la mejor estrategia. En caso de realizar una división del conjunto completo (split-validation) se crean dos conjuntos de datos más pequeños, lo cual en ocasiones puede afectar el rendimiento en caso de no poseer un tamaño muestral adecuado. Tener un conjunto de datos que sea demasiado pequeño para desarrollar el modelo aumenta la probabilidad de sobreajuste y produce un modelo poco confiable al ocasionar una pérdida de información valiosa [Riley et al., 2022, Archer et al., 2021, Snell et al., 2021]. De este modo, el cálculo del tamaño muestral permite anticipar la necesidad de evitar el split-validation en caso de ser necesario y utilizar métodos de re-muestreo (Bootstrap) o validación cruzada (cross-validation) para la validación interna [Steyerberg et al., 2001].

El adecuado tamaño muestral es un requisito indispensable, que contribuye a la estabilidad del riesgo promedio estimado en el modelo desarrollado. Cuanto más pequeño sea el conjunto de datos de desarrollo, mayor será la inestabilidad en el riesgo promedio estimado de un modelo y la consecuencia probable será la inadecuada calibración entre el riesgo promedio estimado y el riesgo promedio observado en la población (inadecuada calibración global). Por esta razón, resulta esencial garantizar que el tamaño de muestra mínimo para el desarrollo del modelo permita estimar el riesgo general con una precisión adecuada [Riley et al., 2020].

Se realizó el cálculo del tamaño muestral necesario para la cantidad de predictores candidatos en el modelo [Riley et al., 2022]. Asumimos de antemano que al menos el 24 % de las admisiones hospitalarias tendrían un evento dado lo reportado en estudios previos [Marshall et al., 2005].

- *Criterio 1: Tamaño muestral para una estimación precisa del riesgo global*

  Para el outcome binario, basado en el riesgo global anticipado del 24 % ($\hat{p}$) y un margen de error absoluto objetivo (E), que corresponde a un intervalo de confianza del 95 % con una amplitud de $2E$, el tamaño muestral mínimo requerido (N) para estimar el riesgo global con precisión se calcula como:

$$N = \left(\frac{1{,}96}{E}\right)^2 \cdot \hat{p} \cdot (1 - \hat{p})$$

  Se recomienda $E \leq 0{,}05$, lo que resulta en un intervalo de confianza con una amplitud menor o igual a 0.1.

- *Criterio 2: Tamaño muestral para minimizar el sobreajuste de los efectos de los predictores:*

$$N = \frac{P}{(S-1) \cdot \ln(1 - R^2_{cs})}$$

P Representa el número de predictores candidatos reconocidos como importantes en el campo de investigación. El R^2_{cs} (estadístico de Cox y Snell) es un coeficiente de determinación generalizado que se utiliza para estimar la proporción de varianza de la variable dependiente explicada por las variables predictoras. Su valor fluctúa entre 0 y 1, pero en la práctica no llega a 1 [Cox, 2018].

El valor de R^2_{cs} se puede derivar de estudios previos o, en ausencia de información, asumir que R^2_{cs} corresponde al 15 % de la varianza total explicada. Esto puede aproximarse como:

$$R^2_{cs} = 0{,}15 \cdot \max(R^2_{cs})$$

Donde:

$$\max(R^2_{cs}) = 1 - \frac{1}{p \cdot (1-p) \cdot \ln(2)^2}$$

siendo p la proporción de eventos en la población.

Alternativamente, si se dispone de un AUROC, esta puede usarse para estimar R^2_{cs}.

- *Criterio 3: Tamaño muestral para un optimismo reducido en el ajuste aparente del modelo*

El tercer criterio se enfoca en limitar la diferencia entre los valores aparentes y ajustados por optimismo de $R^2_{Nagelkerke}$, que es una transformación del R^2_{cs}, es una fundamental del ajuste global del modelo [Cox, 2018].

- El $R^2_{Nagelkerke}$ aparente es el desempeño observado del modelo en los mismos datos utilizados para desarrollarlo.

- El $R^2_{Nagelkerke}$ ajustado por optimismo proporciona una estimación más realista y aproximadamente imparcial del ajuste del modelo en la población objetivo.

El factor de penalización o *shrinkage* que corresponde a un optimismo esperado (O) en $R^2_{Nagelkerke}$ se calcula como:

$$S = \frac{R^2_{cs}}{R^2_{cs} + O \cdot \max(R^2_{cs})}$$

Donde:

- S: Factor de *shrinkage*.

- R^2_{cs}: Estadística de Cox-Snell del modelo.

- $\max(R^2_{cs})$: Valor máximo posible de R^2_{cs}, determinado por la proporción de eventos y el tamaño muestral.

- O: Optimismo permitido (se recomienda $O \leq 0{,}05$).

El valor obtenido de S puede sustituirse en la fórmula del criterio 2 para calcular el tamaño muestral mínimo (N):

$$N = \frac{P}{(S-1) \cdot \ln\left(1 - R^2_{cs}\right)}$$

Basándonos en una AUROC conservadora de 0.84 (la reportada por el benchmark), se calcula un R2 de Cox-Snell de 0.27 [Riley et al., 2019]. Con 28 parámetros predictores candidatos (incluyendo potenciales transformaciones e interacciones clínica y biológicamente relevantes) y un $O \leq 0{,}05$ el tamaño de muestra mínimo requerido para el desarrollo de un nuevo modelo se estimó en un mínimo de 765 observaciones y 184 eventos.

Dado la abundante cantidad de observaciones disponibles en la base de datos utilizada (ampliamente superior al tamaño muestral requerido), se optó por emplear la técnica de split-validation para la construcción y evaluación de los modelos predictivos. Dicha estrategia determinó 2 conjuntos de 13180 (entrenamiento) y 4396 (validación) observaciones, utilizados para la construcción del modelo. Esto permitió disponer de 3683 observaciones para el conjunto de testeo.

3.2 Construcción de los modelos

Previo a la construcción de los modelos, es fundamental considerar las diferencias algorítmicas en el abordaje de los datos faltantes. El manejo de los mismos para los predictores en el modelo de LGBM se realizó mediante la implementación automática incorporada en el algoritmo, mientras que para los modelos benchmark y Lasso el análisis se realizó en los casos completos (se descartan los casos en que alguno de los predictores seleccionados está ausente).

Debido a que la variable resultado presentaba solo 17 observaciones faltantes en el conjunto de entrenamiento, se decidió eliminar esos casos previo a la construcción de los modelos.

Se construyeron 3 modelos mediante regresión logística (benchmark), regresión logística penalizada (Lasso) y LGBM. Se utilizaron todas las variables preseleccionadas para los modelos de LGBM y Lasso. Para la optimización de hiperparámetros se utilizó optimización bayesiana con validación cruzada (10-fold cross-validation). En el modelo benchmark, los predictores utilizados fueron los mismos reportados en el modelo original (peso al nacer, edad gestacional, malformaciones congénitas mayores, maduración pulmonar fetal y score de Apgar al minuto), seleccionados a través de backwards elimination.

Con el objetivo de explorar potenciales ventajas de los modelos de aprendizaje automático, se comparan las predicciones en 2 subconjuntos de entrenamiento:

1. Se construyen los modelos en un conjunto conformado únicamente por observaciones completas (sin datos faltantes) de los predictores de interés utilizados en el modelo benchmark. Esto permite una comparación justa ya que elimina la ventaja de la imputación automática del modelo de LGBM, la cual aumenta el tamaño muestral en relación al que utiliza el modelo benchmark (dado que solo utiliza los datos completos), y permite de esta manera resaltar las potenciales ventajas de los algoritmos de machine learning con respecto a la identificación de interacciones y no linealidades de forma automática.

2. Luego se construyen los modelos en el conjunto que incluye datos faltantes, a modo de explorar la ventaja mencionada sobre la imputación mediante LGBM en comparación con el resto de los modelos.

3.2.1 Regresión logística (benchmark)

La regresión logística (RL) se introduce en este contexto como el modelo de elección clasificación binaria dado el extenso uso [Nick and Campbell, 2007]. Aunque los modelos lineales generalizados no siempre presentan el mejor rendimiento con grandes volúmenes de datos, su aplicación en este estudio sirve como un punto de referencia significativo para evaluar y comparar la eficacia de otros modelos de clasificación. La interpretación intuitiva de sus coeficientes la convierte en una herramienta valiosa para comprender la dinámica subyacente de los datos, proporcionando así una base para el análisis y la evaluación de otras técnicas de modelado más avanzadas.

La estimación de los coeficientes se realiza de la siguiente manera:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

donde:

- y_i: Variable dependiente o respuesta para la observación i.
- β_0: Intercepto o término constante del modelo.
- β_j: Coeficiente asociado a la variable x_{ij}.
- x_{ij}: Valor de la variable independiente j para la observación i.
- n: Número total de observaciones.
- p: Número de variables independientes en el modelo.

3.2.2 Regresión logística penalizada tipo Lasso

La regresión penalizada con operador de contracción y selección de menor valor absoluto (Lasso) [Tibshirani, 1996] es una técnica popular para la selección de modelos y la estimación en modelos de regresión logística. Al modelo de regresión generalizado tradicional se le agrega el término de penalización mencionado:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{\text{penalización}}$$

siendo λ un parámetro de ajuste no negativo que controla la cantidad de contracción, con mayor contracción para valores más altos (penalización). El λ óptimo basado en algún criterio, por ejemplo, el error cuadrático medio (ECM), puede estimarse en un procedimiento de validación cruzada generalizada o mediante el método de bootstrap [Steyerberg, 2010].

En estudios pronósticos, el Lasso es particularmente atractivo por su capacidad para contraer los coeficientes de regresión y realizar automáticamente la selección de variables al establecer algunos coeficientes en cero. Esto mejora el rendimiento predictivo, la parsimonia y mantiene la oportunidad de interpretar los coeficientes de cada predictor de manera sencilla.

Los modelos con menos variables predictoras suelen ser más fáciles de implementar en la práctica y, por lo tanto, en ocasiones existe la preferencia de sacrificar parte del rendimiento predictivo en pos de la simplificación del esfuerzo. Por ejemplo, los médicos en general estarán poco dispuestos a usar modelos que requieran la recopilación de un número elevado de variables, dado que la posibilidad de que existan datos faltantes se incrementa en consideración.

3.2.3 Light gradient boosting machine (LGBM)

LGBM es un algoritmo de aprendizaje supervisado basado en el método de boosting, un tipo de "ensemble learning". El ensemble se refiere a la combinación de varios modelos (en este caso, árboles de decisión) para mejorar la capacidad predictiva global [Zhang et al., 2019].

El boosting funciona entrenando los modelos de manera secuencial. Cada modelo se entrena para corregir los errores cometidos por los modelos anteriores, dándole mayor peso a las observaciones mal predichas. LGBM optimiza este proceso mediante el uso de técnicas avanzadas como el "histogram-based learning" y el "leaf-wise growth", que aceleran el entrenamiento y permiten manejar grandes cantidades de datos con menos recursos de memoria.

Al igual que otros algoritmos de aprendizaje automático, LGBM tiene varios hiperparámetros que se pueden optimizar para mejorar el rendimiento del modelo. Los hiperparámetros optimizados en el presente trabajo fueron:

- Tasa de aprendizaje (learning rate): Determina la contribución de cada árbol al modelo final. Una tasa de aprendizaje más baja puede ayudar a prevenir el sobreajuste al hacer que los árboles contribuyan menos al modelo.
- Número de características consideradas en cada división (feature fraction): Controla la proporción de features consideradas al buscar la mejor división en cada nodo del árbol.
- Cantidad minima de observaciones en una hoja (min data in leaf): Este parámetro especifica el número mínimo de observaciones requeridos en cada hoja del árbol. De esta forma un número mas alto disminuye el sobreajuste pero uno muy elevado puede aumentar el sesgo en las predicciones.
- Número máximo de hojas por árbol (num leaves): Mientras mas hojas por árbol, mayor la capacidad del modelo de capturar patrones, pero un numero muy elevado incrementara en exceso el apego del modelo a los datos de entrenamiento redundando en sobreajuste.

Los modelos de caja negra tales como el LGBM, aunque presentan mayores inconvenientes a la hora de la interpretación, tienen también la ventaja de incorporar interacciones y no linealidades no predefinidas e imputar internamente los datos faltantes. Esto se tra-

duce en un incremento en la información para las predicciones orientada a través de los datos, más allá del conocimiento clínico.

3.2.4 Optimización Bayesiana de hiperparámetros

La optimización bayesiana tiene como objetivo encontrar las combinaciones de hiperparámetros que maximicen el rendimiento del problema original [Snoek et al., 2012]. Define su función objetivo de la misma manera que el problema original, y estima iterativamente la distribución posterior de la función objetivo, así como los subespacios de los hiperparámetros que probablemente logren la función objetivo optimizada. Para ser específicos, en cada iteración, la optimización bayesiana primero utiliza la distribución uniforme para muestrear hiperparámetros de un rango dado y adopta el proceso gaussiano para estimar la distribución posterior de la función objetivo para cada combinación de hiperparámetros muestreados. Luego identifica la región de hiperparámetros que probablemente optimicen la función objetivo utilizando la función de adquisición, que equilibra entre la media posterior $\mu(\mathrm{x})$ y la varianza $\sigma(\mathrm{x})$ de la función objetivo utilizando el parámetro de ajuste κ.

$$AC(x) = \mu(x) + \kappa\sigma(x) \tag{1}$$

Al estimar iterativamente la distribución posterior de la función objetivo y actualizar adaptativamente el subespacio de hiperparámetros, la optimización bayesiana puede aprovechar la información de las evaluaciones previas y concentrarse únicamente en los subespacios prometedores, lo que no solo simplifica el proceso de búsqueda de hiperparámetros, sino que también mejora la estabilidad y tiende a proporcionar configuraciones óptimas de los mismos [Yang et al., 2024].

3.3 Performance de los modelos

3.3.1 Validación interna

Dada la hipótesis de que existe una parte importante de la variabilidad explicada por los centros de atención, se realizó una división en conjuntos de entrenamiento y validación interna basada en el criterio de temporalidad (años calendario comprendidos entre 2010-2018 y 2019-2020, respectivamente). Esto además resultó en una contribución diferencial de centros únicos para cada conjunto, ya que algunos centros se incorporaron a la red de forma más tardía (5 centros, representandoe el 15.5 % del total del conjunto de validación).

3.3.2 Calibración

La calibración mide el grado de acuerdo entre los resultados observados y los riesgos estimados por el modelo. Esta se evalúa mediante la construcción de una curva de calibración flexible suavizada (para detectar potenciales no linealidades) en los datos individuales, representada en un gráfico donde el riesgo estimado corresponde al eje x y el resultado observado al eje y.

La visualización de la curva de calibración se enfoca en la detección de la desviación de la pendiente de la curva de calibración con respecto a la línea de 45° (acuerdo perfecto, lo cual implica una pendiente igual a 1) [Van Calster et al., 2016, Calster et al., 2019].

La calibración también puede cuantificarse numéricamente a través de la pendiente de calibración (valor ideal 1) y la calibración global (valor ideal 0).

La pendiente de calibración (representada a través del parámetro Slope) cuantifica la dispersión de los riesgos estimados del modelo en relación con los resultados observados. Una pendiente menor a 1 sugiere que la dispersión de los riesgos estimados es demasiado extrema (es decir, demasiado alta para las personas con alto riesgo y demasiado baja para aquellas con bajo riesgo). Una pendiente mayor a 1 sugiere que la dispersión de los riesgos estimados es demasiado estrecha.

La calibración global evalúa la calibración media y cuantifica cualquier sobreestimación o subestimación sistemática del riesgo, comparando el número promedio de resultados predichos y el número promedio de resultados observados. Esta se refleja en el valor de la ordenada al origen (Intercept) y cobra mayor relevancia en el conjunto de testeo.

3.3.3 Discriminación

La discriminación evalúa qué tan bien las predicciones del modelo diferencian entre aquellos con y sin el resultado [Steyerberg and Vergouwe, 2014] y se cuantifica típicamente mediante el AUROC (también conocido como el área bajo la curva para resultados binarios).

Un valor de 0.5 indica que el modelo no es mejor que el azar, y un valor de 1 denota una discriminación perfecta (es decir, todas las personas con el resultado tienen riesgos estimados más altos que todas las personas sin el resultado). Lo que define un buen valor del AUROC es específico para el problema y el contexto. Las consecuencias de anticipar un adecuado tamaño muestral también se evidencian en la performance observada en el AUROC.

Es de reciente interés tener en cuenta que incluso si las predicciones de un modelo son estables en el modelo global, aún puede existir una calibración inadecuada en las predicciones para algunos subgrupos definidos por una covariable particular (que puede o no ser uno de los predictores incluidos en el modelo). En consecuencia, es recomendable evaluar la estabilidad de las predicciones en subgrupos de interés, concepto reportado en la literatura como equidad ("Fairness") algorítmica [Riley et al., 2024a].

El modelo con mejor capacidad de discriminación (LGBM) se comparó con el benchmark en los subgrupos de interés prespecificados, enfocándose en una variable predictora incluida en el modelo (Tipo de centro = público y privado) y otra no seleccionada (país). Debido al escaso aporte de los Paraguay, Uruguay y Perú, se decidió agruparlos (Resto), mientras que los países Argentina (Arg) y Chile se presentan por separado.

3.3.4 Validación externa

El conjunto utilizado para la validación externa también cumple con el criterio de temporalidad, reflejando los 2 últimos años del período recolectado (período 2021-2022). Aplicando este criterio, al igual que en el conjunto de validación, existen centros que fueron excluidos en la construcción del modelo en el conjunto el que se se desarrolló. El conjunto de testeo contiene 3 centros (12.6 % del total) que no fueron incluidos en los conjuntos de entrenamiento o de validación.

Las métricas de performance descriptas para la validación interna se calcularon en el conjunto de testeo para los modelos construidos [Riley et al., 2024b]. Sin embargo, las métricas propuestas para la evaluación de la performance de los modelos no proporcionan una respuesta definitiva con respecto a la utilidad de los modelos en la práctica clínica. Por ejemplo, no está claro qué tan alta debe ser el área bajo la curva para justificar su uso clínicoc, o qué grado de falla de calibración sugeriría que un modelo de predicción no debe ser utilizado. Mayores dificultades se plantean al tener que elegir entre dos modelos que presenten resultados contrapuestos entre calibración y discriminación.

El análisis mediante curvas de decisión [Steyerberg and Vergouwe, 2014] permite evaluar la utilidad clínica al cuantificar el equilibrio entre identificar correctamente verdaderos positivos (unidad de medida del beneficio neto) e identificar incorrectamente falsos positivos, ponderados según la probabilidad umbral. La probabilidad umbral representa el punto de corte de riesgo por encima del cual se podría considerar cualquier tratamiento o intervención, y refleja la relación riesgo-beneficio percibida para la intervención.

El beneficio neto (NB) se define como:

$$\text{NB} = \frac{\text{VP}}{n} - \frac{\text{FP}}{n}\left(\frac{pt}{1-pt}\right)$$

donde:

- VP es el número de verdaderos positivos.
- n es el tamaño total de la muestra.
- FP es el número de falsos positivos.
- pt es el umbral de probabilidad.

Incorporando los pt en forma de escala continua se logra construir la curva de decisiones para todos los umbrales, permitiendo luego seleccionar los de interés y reportar el NB acorde.

Se utilizó el análisis de curva de decisiones en la validación externa para cuantificar el NB de implementar el modelo en la práctica clínica, utilizando las estrategias descriptas en la literatura [Steyerberg and Vergouwe, 2014, Vickers et al., 2008]:

- Un enfoque de tratar a todos ("Treat all"), en el que se asume un resultado hipotético para todos los individuos independientemente del riesgo (todos presentarán el outcome). .
- Un enfoque de no tratar a ninguno ("Treat none"), en el que se asume que nadie presentará el evento de interés.
- Un enfoque que utiliza los modelos construídos para identificar el NB para compararlos entre ellos y además con las estrategias hipotéticas previamente descriptas.

En el contexto del análisis de beneficio neto aplicado a modelos predictivos de riesgo de mortalidad, las estrategias "Treat all" y "Treat none" representan dos enfoques extremos frente a los cuales se evalúa el rendimiento clínico de un modelo como herramienta de apoyo para la toma de decisiones.

La estrategia "Treat none" asume que ningún paciente tiene riesgo significativo de morir; por lo tanto, no se indicaría ninguna intervención o acción médica adicional basada en el modelo. Esta estrategia, al no identificar ningún caso en riesgo, tiene un beneficio neto igual a cero. Por el contrario, la estrategia "Treat all" supone que todos los pacientes están en riesgo y, por ende, se les aplicaría la intervención preventiva a todos, sin discriminar entre quienes efectivamente tienen riesgo de morir y quienes no. Esta estrategia, aunque puede captar a todos los verdaderos positivos, también conlleva un alto número de falsos positivos, lo que penaliza su beneficio neto.

El análisis de beneficio neto compara estas estrategias con el rendimiento del modelo predictivo, considerando cuántos verdaderos positivos (pacientes correctamente identificados como en riesgo de mortalidad) se detectan y cuántos falsos positivos se generan. El beneficio neto se interpreta como el número de verdaderos positivos netos por cada 100 pacientes evaluados, ajustado por el umbral de riesgo y el impacto negativo de tratar (o intervenir) innecesariamente a quienes no están en riesgo. Por lo tanto, un modelo con un beneficio neto superior al de "treat all" y "treat none" en un rango clínicamente relevante de umbrales se considera útil, ya que permite una mejor discriminación del riesgo de mortalidad, orientando de manera más eficiente las decisiones clínicas o preventivas.

Por lo previamente explicado, se presentan los valores de NB a lo largo de un rango de umbrales, permitiendo, de esta manera, la comparación de utilidades entre los distintos modelos.

4 Resultados

4.1 Análisis primario

La muestra con la cual se entrenó el modelo comprende 17.566 RNMBP vivos en el período comprendido entre mayo del año 2010 a mayo del 2020 (conjuntos de entrenamiento y validación interna). Las características demográficas y clínicas neonatales y maternas se muestran estratificadas por resultado [Tabla 2]. Se desconocían los resultados de desenlace de solo 17 participantes, los cuales fueron excluidos.

La mortalidad en promedio en la red es del 24.7 %, el 6.1 % presenta malformaciones congénitas mayores y de estas el 67 % de los neonatos fallecen. Se destaca que Argentina y Chile son los países que más contribuyen con pacientes en la red, la cual a su vez está representada mayoritariamente por centros públicos.

El conjunto de testeo incluye 3683 RNMBP y no presenta grandes diferencias en la distribución de variables [Sección 7, Tabla 6].

Tabla 2. Características demográficas del conjunto de entrenamiento

	Vive (N=13211)	Fallece (N=4355)	Total (N=17566)
Peso al nacimiento (gr)			
Mean (SD)	1160 (243)	860 (274)	1090 (283)
Median [Min, Max]	1200 [400, 1500]	800 [400, 1500]	1120 [400, 1500]
Tipo de centro			
Privado	2963 (22.4 %)	584 (13.4 %)	3547 (20.2 %)

	Vive (N=13211)	Fallece (N=4355)	Total (N=17566)
Público	10248 (77.6 %)	3771 (86.6 %)	14019 (79.8 %)
País			
Argentina	5297 (40.1 %)	1652 (37.9 %)	6949 (39.6 %)
Chile	5236 (39.6 %)	1512 (34.7 %)	6748 (38.4 %)
Paraguay	788 (6.0 %)	319 (7.3 %)	1107 (6.3 %)
Perú	1208 (9.1 %)	663 (15.2 %)	1871 (10.7 %)
Uruguay	682 (5.2 %)	209 (4.8 %)	891 (5.1 %)
Apgar 1 minuto			
Mean (SD)	6.43 (2.17)	3.92 (2.46)	5.81 (2.50)
Median [Min, Max]	7.00 [0, 10.0]	4.00 [0, 9.00]	7.00 [0, 10.0]
Edad gestacional (semanas)			
Mean (SD)	29.4 (2.48)	26.6 (2.82)	28.7 (2.85)
Median [Min, Max]	29.0 [23.0, 36.0]	26.0 [23.0, 36.0]	29.0 [23.0, 36.0]
Ruptura prematura de membranas (días)			
Mean (SD)	1.40 (6.43)	2.35 (10.1)	1.64 (7.52)
Median [Min, Max]	0 [0, 99.0]	0 [0, 99.0]	0 [0, 99.0]
Corioamnionitis			
No	9574 (72.5 %)	2997 (68.8 %)	12571 (71.6 %)
Sí	1212 (9.2 %)	533 (12.2 %)	1745 (9.9 %)
Missing	2425 (18.4 %)	825 (18.9 %)	3250 (18.5 %)
Malformación congénita mayor			
No	12840 (97.2 %)	3624 (83.2 %)	16464 (93.7 %)
Sí	350 (2.6 %)	716 (16.4 %)	1066 (6.1 %)
Missing	21 (0.2 %)	15 (0.3 %)	36 (0.2 %)
Maduración pulmonar fetal			
No recibe	1759 (13.3 %)	1404 (32.2 %)	3163 (18.0 %)
Recibe	11363 (86.0 %)	2897 (66.5 %)	14260 (81.2 %)
Missing	89 (0.7 %)	54 (1.2 %)	143 (0.8 %)
Sulfato de Mg			
No recibe	4491 (34.0 %)	1884 (43.3 %)	6375 (36.3 %)
Tratamiento	3586 (27.1 %)	812 (18.6 %)	4398 (25.0 %)
Missing	5134 (38.9 %)	1659 (38.1 %)	6793 (38.7 %)
Antibiótico prenatal			
No	8660 (65.6 %)	2852 (65.5 %)	11512 (65.5 %)
Sí	4464 (33.8 %)	1454 (33.4 %)	5918 (33.7 %)
Missing	87 (0.7 %)	49 (1.1 %)	136 (0.8 %)
Nivel educacional			
Primaria incompleta	286 (2.2 %)	97 (2.2 %)	383 (2.2 %)
Secundaria incompleta	2143 (16.2 %)	808 (18.6 %)	2951 (16.8 %)
Técnica	4730 (35.8 %)	1532 (35.2 %)	6262 (35.6 %)
Universitaria	4599 (34.8 %)	1480 (34.0 %)	6079 (34.6 %)
Missing	1453 (11.0 %)	438 (10.1 %)	1891 (10.8 %)
Edad materna (años)			
Mean (SD)	28.9 (7.24)	27.9 (7.37)	28.6 (7.29)
Median [Min, Max]	29.0 [12.0, 60.0]	27.0 [12.0, 52.0]	29.0 [12.0, 60.0]
Missing	9 (0.1 %)	6 (0.1 %)	15 (0.1 %)

	Vive (N=13211)	Fallece (N=4355)	Total (N=17566)
Sexo			
Femenino	6472 (49.0 %)	1849 (42.5 %)	8321 (47.4 %)
Masculino	6725 (50.9 %)	2475 (56.8 %)	9200 (52.4 %)
Missing	14 (0.1 %)	31 (0.7 %)	45 (0.3 %)
Diabetes materna			
No	12189 (92.3 %)	4098 (94.1 %)	16287 (92.7 %)
Sí	919 (7.0 %)	203 (4.7 %)	1122 (6.4 %)
Missing	103 (0.8 %)	54 (1.2 %)	157 (0.9 %)
Hipertensión arterial materna			
No	8823 (66.8 %)	3483 (80.0 %)	12306 (70.1 %)
Sí	4353 (32.9 %)	848 (19.5 %)	5201 (29.6 %)
Missing	35 (0.3 %)	24 (0.6 %)	59 (0.3 %)
Control Prenatal			
Completo	12129 (91.8 %)	3719 (85.4 %)	15848 (90.2 %)
Incompleto	1082 (8.2 %)	636 (14.6 %)	1718 (9.8 %)
Tabaquismo			
No	11697 (88.5 %)	3748 (86.1 %)	15445 (88.0 %)
Sí	1484 (11.2 %)	575 (13.2 %)	2059 (11.7 %)
Missing	30 (0.2 %)	32 (0.7 %)	62 (0.4 %)

Distribución de valores de la población. Las variables numéricas se expresan en media[SD] y mediana[min-max]. Las variables categóricas se expresan en porcentaje.

Las métricas de discriminación y calibración se presentan para los 3 modelos generados para la construcción del modelo (Original), luego de la validación interna y en el conjunto de testeo [Tabla 3]. En base a las mismas, el modelo utilizando LGBM supera la performance de los modelos de Lasso y LR (AUROC de 0.89, 0.87 y 0.87 respectivamente), presentando una peor, aunque aceptable calibración representada en los parámetros de Intercept y Slope.

Cabe destacar que el modelo LGBM fue comparado en el conjunto de testeo que no presenta datos faltantes en las variables utilizadas por el modelo benchmark y los resultados no arrojaron diferencias (los datos faltantes representan menos del 1.2 % del conjunto de entrenamiento y 1.3 % en el conjunto de testeo). Sin embargo, a diferencia del modelo Benchmark (de menor complejidad debido a la menor cantidad de predictores incluidos), el modelo Lasso solo utiliza el 52 % de los datos del conjunto de entrenamiento (9142 RNMBP).

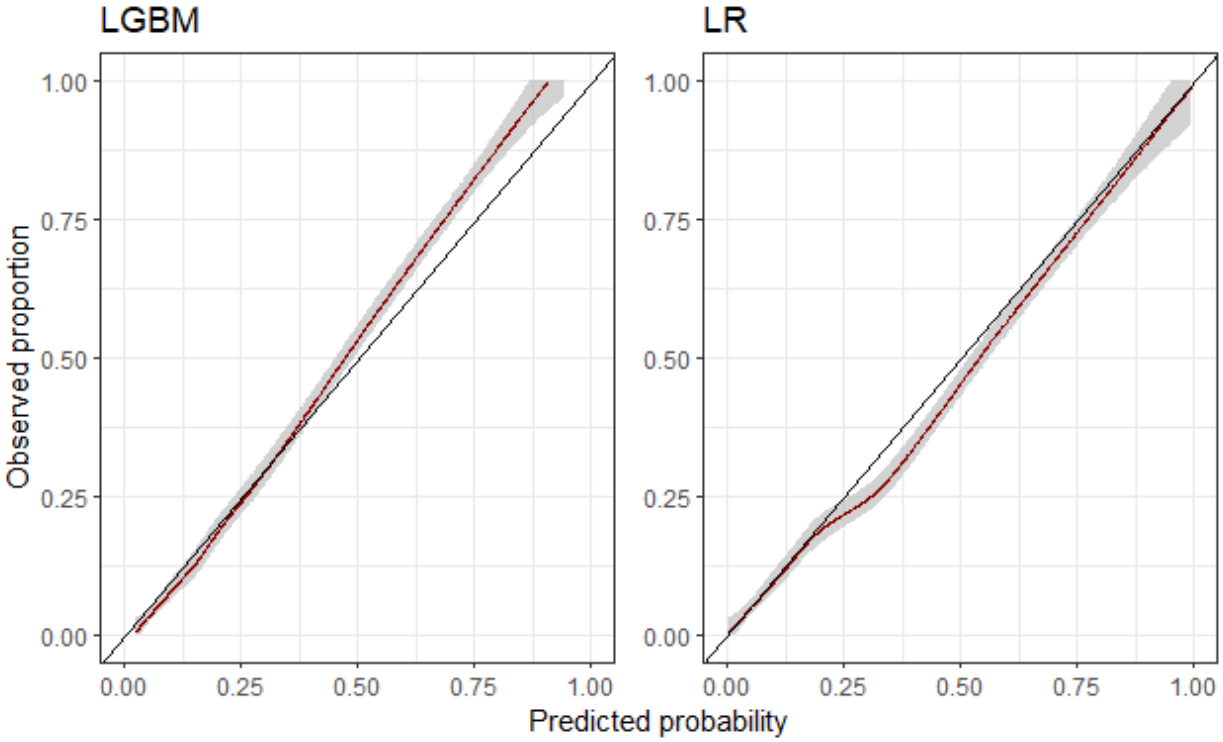
Con respecto a la visualización de la calibración, se presentan las curvas suavizadas para los modelos de RL y de LGBM [Figura 6] las cuales demuestran ser apropiadas a la inspección.

Tabla 3. Resultados de los modelos

Modelo	Métrica	Entrenamiento	Testeo
LGBM	**AUROC**	0.90	0.89
	Intercept	0.00	-0.17
	Slope	1.31	1.26
Lasso	**AUROC**	0.88	0.87
	Intercept	0.00	-0.14
	Slope	1.12	1.11
LR	**AUROC**	0.87	0.87
	Intercept	0.00	-0.15
	Slope	1.00	0.99

Entrenamiento: Validación interna. *Testeo:* Validación externa.

Figura 6. Calibración de los modelos



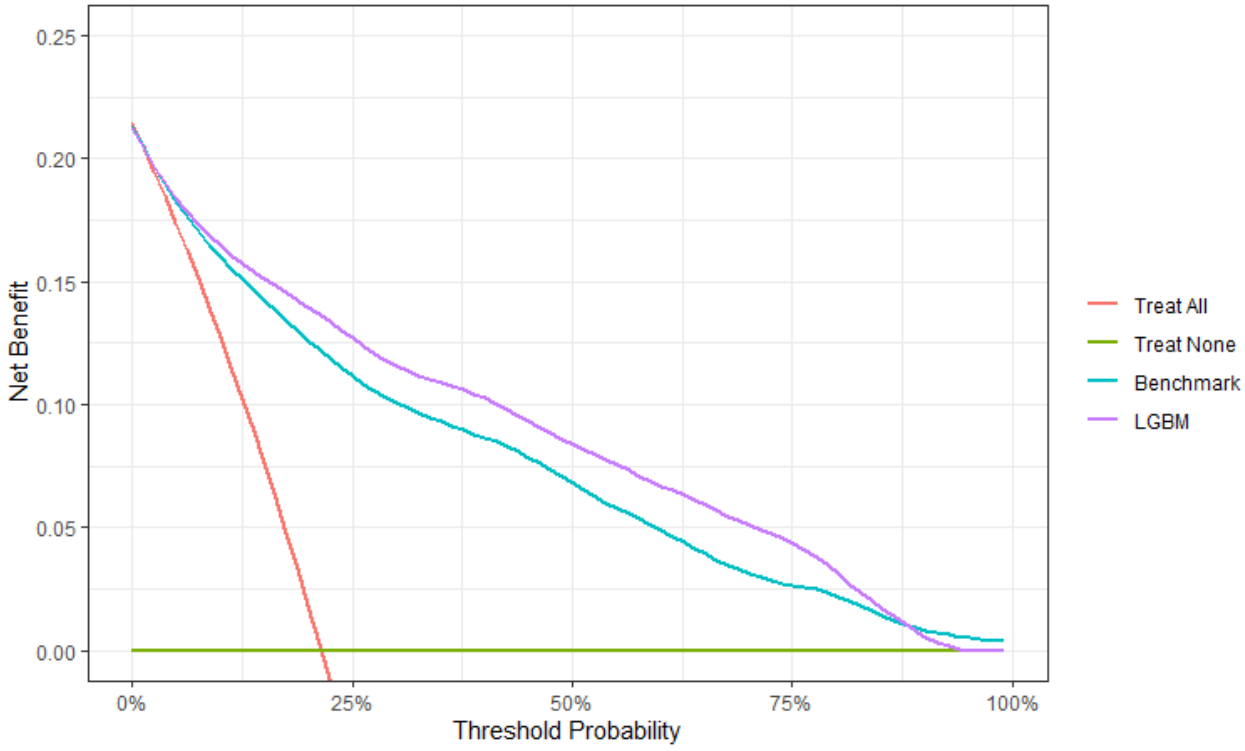
Riesgo observado vs predicho para cada modelo. Negro: Riesgo observado
Rojo: Riesgo predicho

El análisis de la curva de decisiones en el conjunto reservado para el testeo permite examinar la utilidad clínica para ambos modelos pronósticos. Los modelos mostraron un beneficio neto superior a los enfoques de tratar a todos (asumir que todos van a presentar el evento de interés) y no tratar a ninguno (nadie presentará el evento de interés) en un rango de probabilidades umbral [Figura 7].

La utilidad de los modelos pronósticos construídos es evidente al ser comparada con la estrategia que prescinde de ellos. Sin embargo, el modelo de LGBM es superior al

benchmark para la gran mayoría de los umbrales de riesgo, tal como se presenta en el material suplementario [Tabla 7].

Figura 7. Curvas de análisis de decisión



Beneficio neto por umbral de riesgo para cada modelo. *Probabilidad umbral (eje X) Beneficio neto (eje Y)*

4.2 Análisis secundario por subgrupos

El análisis por subgrupos también demuestra la superioridad en la performance del modelo de LGBM en los subgrupos por región (predictor no incluido en el modelo) y tipo de centro (predictor incluido), cumpliendo con el requerimiento de equidad algorítmica.

4.2.1 Análisis por países

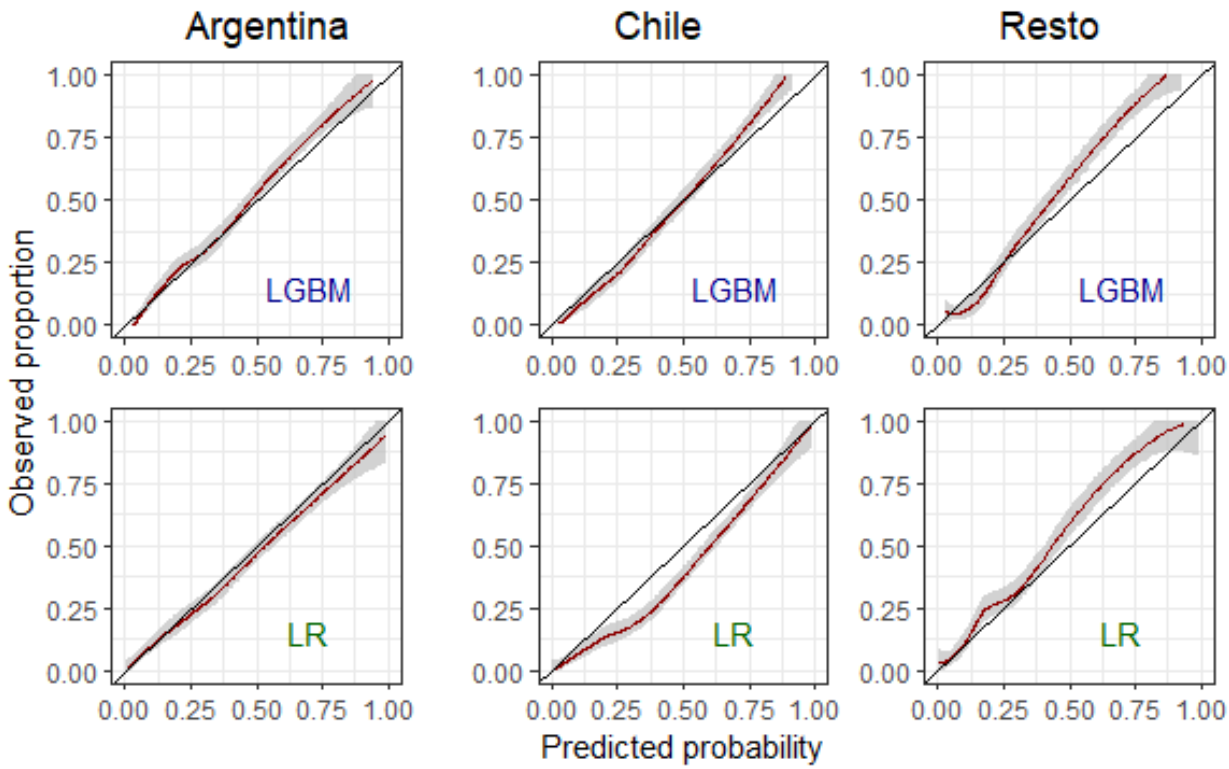
El AUROC fue superior para cada una de las regiones analizadas con moderado impacto en la calibración (Slope) y mayor riesgo de sobreajuste [Tabla 4, Figura 8].

Tabla 4. Comparación de modelos LGBM y LR para los diferentes países

Modelo		Arg	Chile	Resto
LGBM	**AUROC**	0.90	0.91	0.88
	Intercept	0.01	-0.14	0.10
	Slope	1.20	1.27	1.34
LR	**AUROC**	0.86	0.88	0.87
	Intercept	-0.12	-0.49	0.28
	Slope	0.95	1.04	1.13

Valores de performance en el conjunto de testeo

Figura 8. Calibración de los modelos por país



Riesgo observado vs predicho para cada modelo *Negro: Riesgo observado*
Rojo: Riesgo predicho

4.2.2 Análisis por tipo de centro

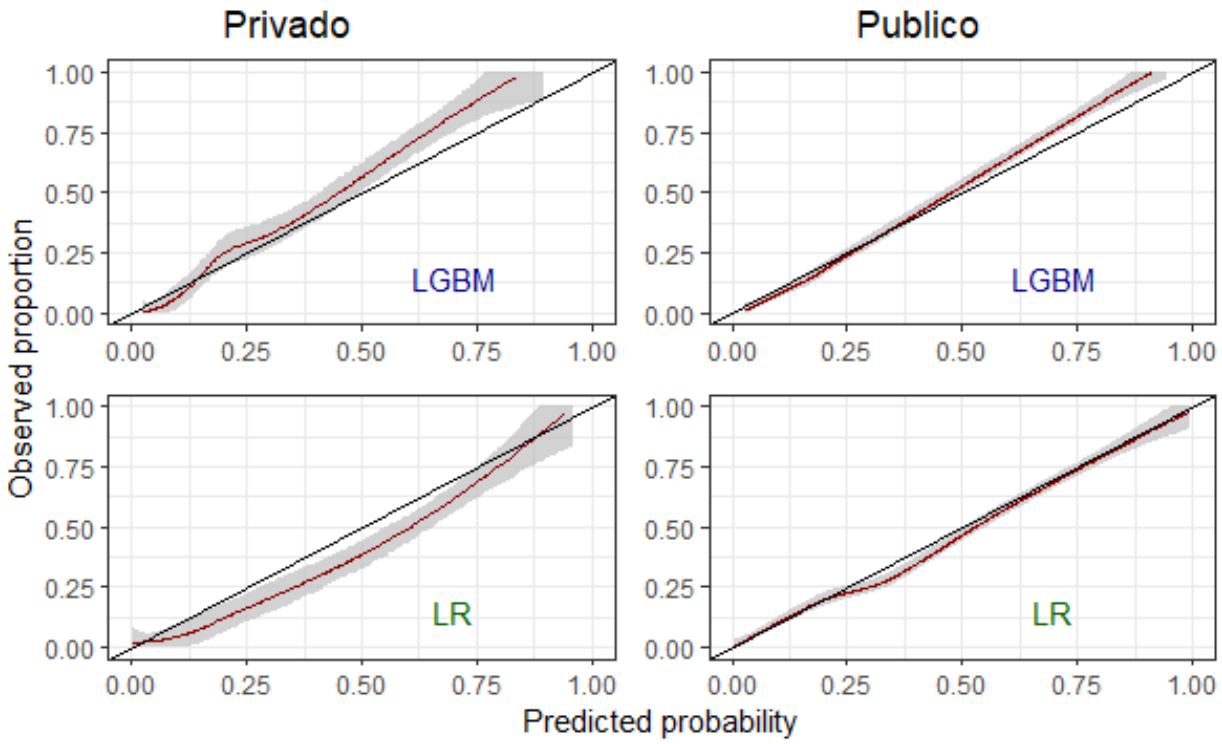
Similar al análisis regional, el correspondiente enfocado a centros públicos o privados, también demuestra la superioridad del modelo elegido [Tabla 5, Figura 9].

Tabla 5. Comparación de modelos LGBM y LR para datos privados y públicos

Modelo		Privado	Público
LGBM	**AUROC**	0.92	0.89
	Intercept	-0.03	-0.01
	Slope	1.45	1.23
LR	**AUROC**	0.88	0.87
	Intercept	-0.48	-0.10
	Slope	1.04	0.99

Valores de performance en el conjunto de testeo

Figura 9. Calibración de los modelos por tipo de centro



Riesgo observado vs predicho para cada modelo *Negro: Riesgo observado*
Rojo: Riesgo predicho

Los resultados considerando al tipo de centro público/privado muestran una mejora en la performance del modelo de LGBM en relación al benchmark, aunque con un leve empeoramiento en la calibración.

Al igual que en el análisis de subgrupo por región, existe un aumento en el riesgo de sobreajuste.

5 Discusión

En este estudio, la aplicación de modelos de aprendizaje supervisado automático permitió superar la performance del modelo logístico ajustado previamente por la Red Neocosur. Asimismo, se logró una mejora en la robustez metodológica con la incorporación de técnicas de evaluación de calibración (Intercept, Slope y visualización a través de curvas suavizadas) y de beneficio neto (NB; curvas de análisis de decisión).

Aunque la comparación entre modelos no muestra diferencias extremas, el modelo LGBM fue el de mejor performance. Entre las ventajas del nuevo modelo propuesto, más allá de la incorporación de una mayor cantidad de predictores y la mejoría en el AUROC, se destaca la mejora en el NB. Para ilustrar, imaginemos una intervención hipotética que podría prevenir una muerte inminente si se administra a tiempo. Considerando un umbral de decisión de 0.2 (es decir, se decide intervenir si el modelo estima una probabilidad de muerte mayor al 20 %), el modelo LGBM alcanza un NB de 0.14, mientras que el modelo comparador (benchmark) alcanza un NB de 0.12. Esto implica que, por cada 100 pacientes evaluados, el modelo LGBM permite identificar correctamente a 2 pacientes más que realmente están en riesgo de morir y que podrían beneficiarse de la intervención (14 vs. 12), en comparación con el benchmark. Sin embargo, ambos modelos superan ampliamente a la estrategia de "Treat all", que para el mismo umbral solo lograría identificar correctamente a 2 pacientes de cada 100 (NB = 0.02), debido a la penalización por el alto número de falsos positivos. Esto resalta el valor agregado de utilizar un modelo predictivo calibrado frente a decisiones clínicas no estratificadas por riesgo.

Un aspecto relevante de los modelos de aprendizaje supervisado es su capacidad para capturar interacciones complejas y relaciones no lineales entre las variables predictoras. Tradicionalmente, este tipo de relación ha sido abordado de forma simplificada mediante la construcción de curvas de percentilos que relacionan el peso al nacer con la edad gestacional, dando lugar a la variable bajo peso para la edad gestacional (BPEG). Sin embargo, como se planteó en la introducción, las curvas disponibles suelen ser específicas para cada país y difícilmente extrapolables a otras poblaciones. Además, desde una perspectiva estadística, la dicotomización de variables continuas (como ocurre al definir el BPEG como un valor sí/no) implica una pérdida sustancial de información y poder predictivo, lo cual ha sido ampliamente documentado [Harrell, 2016].

El análisis del ranking de importancia de variables (Sección 7) revela diferencias sustanciales entre las capacidades mencionadas en el apartado anterior. Mientras que el modelo de regresión logística [Figura 11] identifica la presencia de malformaciones congénitas mayores como el predictor de mayor peso, el modelo LGBM [Figura 12] otorga una importancia preponderante a variables continuas como el peso al nacimiento, el puntaje de Apgar y la edad gestacional, lo cual sugiere la presencia de interacciones no lineales relevantes que son adecuadamente modeladas mediante técnicas de aprendizaje automático.

No obstante, los modelos de tipo çaja negra"presentan riesgos inherentes relacionados con el sobreajuste o subajuste. En el presente estudio, se identificó una tendencia del modelo LGBM hacia una calibración optimista, tanto en el conjunto global de prueba como en los análisis de subgrupos, evidenciando una sobreestimación del riesgo predicho. Aunque la incorporación de técnicas de validación cruzada interna permitió atenuar esta limitación y mejorar la estabilidad de los resultados, el modelo más parsimonioso alcanzó una calibración superior, lo que refuerza la importancia de considerar la complejidad como

un factor crítico en la construcción de modelos predictivos clínicos.

A pesar de su superior capacidad discriminativa, el modelo con mejor desempeño presenta una mayor complejidad, lo que conlleva desafíos prácticos en términos de aplicabilidad e interpretabilidad clínica. La necesidad de recolectar un mayor número de predictores podría comprometer la calidad y completitud de los datos en contextos de práctica clínica habitual, en contraste con modelos más parsimoniosos que emplean variables recolectadas de forma rutinaria. No obstante, cabe destacar que el algoritmo LGBM incorpora mecanismos internos de imputación, lo que le permite mantener un rendimiento robusto aun en presencia de datos faltantes. Esta característica representa una ventaja particular al considerar modelos con múltiples predictores, dado que la pérdida de información asociada a datos incompletos puede constituir una limitación significativa. El modelo construido mediante penalización Lasso ilustra este riesgo potencial de pérdida de información; sin embargo, en este caso específico, el tamaño muestral fue suficiente para mitigar su impacto.

Respecto a la interpretabilidad, los modelos de aprendizaje automático presentan limitaciones inherentes, dado que no ofrecen coeficientes o parámetros fácilmente interpretables para cada variable predictora. Sin embargo, considerando que el objetivo de este trabajo se centra exclusivamente en la predicción del riesgo y no en la inferencia causal, la falta de interpretabilidad no constituye un obstáculo metodológico significativo. Por el contrario, evita el riesgo de otorgar interpretaciones causales erróneas a coeficientes que no fueron estimados bajo un diseño inferencial [Westreich and Greenland, 2013].

Finalmente, considerando la equidad algorítmica, se observa que el modelo seleccionado también supera en performance al incluir subgrupos relevantes tales como el tipo de centro y el país. Esto es de particular interés ya que el beneficio observado al incluir la variabilidad por centros se refleja en los países (a pesar de no haber sido estos últimos incluidos como predictores en el modelo) y refuerza el concepto de que el modelo es estable en las predicciones para distintos subgrupos, y por lo tanto tiene potencial validez para la implementación.

Si bien los resultados obtenidos son prometedores, su aplicación en la práctica clínica requerirá validación adicional mediante estudios de implementación. Al igual que el modelo benchmark, el modelo desarrollado en este trabajo fue concebido con un enfoque generalizable, buscando favorecer su aplicabilidad en diversos contextos sanitarios en lugar de una implementación única y específica.

No obstante, una limitación importante radica en la heterogeneidad de los sistemas de salud en la región. En particular, Argentina y Chile (que concentran la mayor parte de las observaciones en la base de datos) presentan una atención sanitaria altamente fragmentada, con marcadas diferencias estructurales y funcionales entre el sector público y el privado. En este contexto, no es metodológicamente viable excluir el tipo de centro como predictor, ya que representa realidades asistenciales profundamente distintas. Esta característica introduce una tensión inherente entre construir un modelo regional generalizable y preservar la valiosa información proveniente de los principales contribuyentes. Así, esta limitación, difícil de corregir sin comprometer la validez o la representatividad, debe ser considerada al interpretar los resultados y planificar futuras validaciones externas.

Una potencial aplicación adicional del modelo desarrollado radica en su utilización como herramienta para la evaluación de intervenciones, a través de la comparación entre

los riesgos observados y aquellos predichos a partir del modelo. Esta estrategia permitiría estimar el efecto de las iniciativas de mejora de la calidad, tanto en mortalidad como en complicaciones asociadas a la prematuridad, de forma más ajustada a las características de la población local. En la práctica habitual, ante la implementación de intervenciones destinadas a mejorar los desenlaces clínicos, y en ausencia de parámetros propios de referencia, es frecuente que se recurra a comparaciones con resultados de redes o iniciativas internacionales. Sin embargo, dichas comparaciones presentan limitaciones intrínsecas, dado que los estudios utilizados como referencia suelen estar basados en poblaciones distintas a la de interés, generando un riesgo de sesgo al momento de evaluar el impacto real de las intervenciones.

Para complementar, es importante considerar que la replicación de estudios prospectivos experimentales en cada contexto específico no siempre es factible, debido a restricciones de tiempo, de recursos económicos y a consideraciones éticas. En este escenario, el desarrollo de modelos predictivos ajustados a la población blanco ofrece una alternativa metodológicamente robusta, permitiendo evaluar las mejoras observadas mediante la comparación de los riesgos predichos con los resultados efectivamente observados, evitando la necesidad de recurrir a comparaciones externas potencialmente inadecuadas.

Entre las líneas futuras de trabajo, destacan dos áreas prioritarias: la optimización del manejo de datos faltantes y el fortalecimiento de la interpretabilidad de los modelos. Si bien existen métodos de imputación múltiples con eficacia demostrada en diversos contextos [White et al., 2011], su aplicación en escenarios de alta dimensionalidad, como el presente, presenta desafíos importantes relacionados con el costo computacional y la estabilidad de las imputaciones. En este sentido, se propone como objetivo futuro la exploración de estrategias que logren un equilibrio adecuado entre precisión, eficiencia y facilidad de implementación en la práctica clínica. En paralelo, la interpretabilidad de los modelos de aprendizaje supervisado constituye un aspecto crítico para su adopción efectiva. Tal como se ha discutido, los modelos parsimoniosos ofrecen ventajas sustanciales no solo por su simplicidad técnica, sino también por su capacidad de generar confianza entre los profesionales de la salud, lo cual resulta clave para promover su utilización en entornos asistenciales reales.

En conclusión, el presente trabajo ofrece un modelo pronóstico de mortalidad neonatal en RNMBP que supera al modelo benchmark en términos de desempeño predictivo. Este modelo representa una contribución relevante y actualizada a las necesidades identificadas en el ámbito de la red Neocosur, proporcionando una herramienta potencialmente útil para la mejora continua de la calidad asistencial en esta población vulnerable.

6 Bibliografía

Referencias

[Alarcón et al., 2008] Alarcón, J., Alarcón, Y., Hering, E., and Buccioni, R. (2008). Curvas antropométricas de recién nacidos chilenos. *Revista Chilena de Pediatría*, 79(4):364–372.

[Archer et al., 2021] Archer, L., Snell, K. I. E., Ensor, J., Hudda, M. T., Collins, G. S., and Riley, R. D. (2021). Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Statistics in Medicine*, 40:133–146.

[Baguiya et al., 2021] Baguiya, A., Bonet, M., Cecatti, J., Brizuela, V., Curteanu, A., Minkauskiene, M., Jayaratne, K., Ribeiro-do Valle, C., Budianu, M., Souza, J., Kouanda, S., Group, W. G. M. S. S. G. R., and research group, G. (2021). Perinatal outcomes among births to women with infection during pregnancy. *Archives of Disease in Childhood*, 106(10):946–953.

[Calster et al., 2019] Calster, B. V., McLernon, D. J., van Smeden, M., et al. (2019). Calibration: the achilles heel of predictive analytics. *BMC Medicine*, 17:230.

[Collins et al., 2024] Collins, G., Dhiman, P., Ma, J., Schlussel, M., Archer, L., Van Calster, B., Harrell Jr, F., Martin, G., Moons, K., van Smeden, M., Sperrin, M., Bullock, G., and Riley, R. (2024). Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ*, 384:e074819. PMID: 38191193; PMCID: PMC10772854.

[Cox, 2018] Cox, D. R. (2018). *Analysis of binary data*. Routledge.

[Fenton and Kim, 2013] Fenton, T. R. and Kim, J. H. (2013). A systematic review and meta-analysis to revise the fenton growth chart for preterm infants. *BMC Pediatrics*, 13:59.

[Fernández et al., 2014] Fernández, R., D'Apremont, I., Domínguez, A., Tapia, J. L., and Neocosur, R. N. (2014). Survival and morbidity of very low birth weight infant in a south american neonatal network. *Archivos argentinos de pediatria*, 112(5):405–412.

[García-Muñoz Rodrigo et al., 2021] García-Muñoz Rodrigo, F., Fabres, J., Tapia, J. L., D'Apremont, I., San Feliciano, L., Zozaya Nieto, C., Figueras-Aloy, J., Mariani, G., Musante, G., Silvera, F., Zegarra, J., and Vento, M. (2021). Factors associated with survival and survival without major morbidity in very preterm infants in two neonatal networks: Sen1500 and neocosur. *Neonatology*, 118(3):289–296.

[Harrell, 2016] Harrell, F. E., J. (2016). *Regression modeling strategies*. Springer International Publishing.

[Iriondo et al., 2020] Iriondo, M., Thio, M., Del Río, R., Baucells, B. J., Bosio, M., and Figueras-Aloy, J. (2020). Prediction of mortality in very low birth weight neonates in spain. *PLOS ONE*, 15(7):e0235794.

[Lona Reyes et al., 2018] Lona Reyes, J. C., Pérez Ramírez, R. O., Llamas Ramos, L., Gómez Ruiz, L. M., Benítez Vázquez, E. A., and Rodríguez Patiño, V. (2018). Neonatal

mortality and associated factors in newborn infants admitted to a neonatal care unit. *Archivos Argentinos de Pediatría*, 116(1):42–48.

[Marshall et al., 2012] Marshall, G., Luque, M. J., Gonzalez, A., D'Apremont, I., Musante, G., and Tapia, J. L. (2012). Center variability in risk of adjusted length of stay for very low birth weight infants in the neocosur south american network. *Jornal de pediatria*, 88(6):524–530.

[Marshall et al., 2005] Marshall, G., Tapia, J. L., D'Apremont, I., Grandi, C., Barros, C., Alegria, A., Standen, J., Panizza, R., Roldan, L., Musante, G., Bancalari, A., Bambaren, E., Lacarruba, J., Hubner, M. E., Fabres, J., Decaro, M., Mariani, G., Kurlat, I., Gonzalez, A., and NEOCOSUR, G. C. (2005). A new score for predicting neonatal very low birth weight mortality risk in the neocosur south american network. *Journal of perinatology : official journal of the California Perinatal Association*, 25(9):577–582.

[Medlock et al., 2011] Medlock, S., Ravelli, A. C., Tamminga, P., Mol, B. W., and Abu-Hanna, A. (2011). Prediction of mortality in very premature infants: a systematic review of prediction models. *PLOS ONE*, 6(9):e23441.

[Nick and Campbell, 2007] Nick, T. and Campbell, K. (2007). Logistic regression. *Methods in Molecular Biology*, 404:273–301.

[Olsen et al., 2010] Olsen, I. E., Groveman, S. A., Lawson, M. L., Clark, R. H., and Zemel, B. S. (2010). New intrauterine growth curves based on united states data. *Pediatrics*, 125(2):e214–e224.

[Riley et al., 2024a] Riley, R. D., Archer, L., Snell, K. I. E., Ensor, J., Dhiman, P., Martin, G. P., Bonnett, L. J., and Collins, G. S. (2024a). Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ*, 384:e074820. PMID: 38224968; PMCID: PMC10788734.

[Riley and Collins, 2023] Riley, R. D. and Collins, G. S. (2023). Stability of clinical prediction models developed using statistical or machine learning methods. *Biom J*, 65(8):e2200302. Epub 2023 Jul 19.

[Riley et al., 2022] Riley, R. D., Collins, G. S., Ensor, J., and et al. (2022). Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. *Statistics in Medicine*, 41:1280–1295.

[Riley et al., 2020] Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E. J., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., and van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, 368:m441.

[Riley et al., 2024b] Riley, R. D., Snell, K. I. E., Archer, L., Ensor, J., Debray, T. P. A., van Calster, B., van Smeden, M., and Collins, G. S. (2024b). Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. *BMJ*, 384:e074821. PMID: 38253388.

[Riley et al., 2019] Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Harrell, F. E. J., Moons, K. G. M., and Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: Part ii - binary and time-to-event outcomes. *Statistics in Medicine*, 38(7):1276–1296. Erratum in: Stat Med. 2019 Dec 30;38(30):5672. doi: 10.1002/sim.8409.

[Schafer and Yucel, 2002] Schafer, J. L. and Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, 11(2):437–457.

[Simon et al., 2024] Simon, L. V., Shah, M., and Bragg, B. N. (2024). Apgar score. StatPearls [Internet]. Treasure Island (FL).

[Snell et al., 2021] Snell, K. I. E., Archer, L., Ensor, J., and et al. (2021). External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. *Journal of Clinical Epidemiology*, 135:79–89.

[Snoek et al., 2012] Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 25.

[Steyerberg, 2010] Steyerberg, E. W. (2010). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer, New York.

[Steyerberg et al., 2001] Steyerberg, E. W., Harrell Jr, F. E., Borsboom, G. J. J. M., Eijkemans, M. J. C., Vergouwe, Y., and Habbema, J. D. F. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, 54:774–781.

[Steyerberg and Vergouwe, 2014] Steyerberg, E. W. and Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an abcd for validation. *European Heart Journal*, 35(29):1925–1931. Epub 2014 Jun 4.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[Toso et al., 2022] Toso, A., Vaz Ferreira, C., Herrera, T., Villarroel, L., Brusadin, M., Escalante, M. J., Masoli, D., D'Apremont, I., Mariani, G., and Tapia, J. L. R. N. N. (2022). Mortality in very low birth weight (vlbw) infants in south american neocosur neonatal network: timing and causes. *Archivos Argentinos de Pediatría*, 120(5):296–303. English, Spanish. Epub 2022 Aug.

[Van Calster et al., 2016] Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., and Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*, 74:167–176. Epub 2016 Jan 6.

[Vickers et al., 2008] Vickers, A. J., Cronin, A. M., Elkin, E. B., and Gonen, M. (2008). Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC medical informatics and decision making*, 8:53.

[Villar et al., 2015] Villar, J., Giuliani, F., Bhutta, Z. A., Bertino, E., Ohuma, E. O., Ismail, L. C., et al. (2015). Postnatal growth standards for preterm infants: the preterm postnatal follow-up study of the intergrowth-21(st) project. *Lancet Glob Health*, 3:e681–e691.

[Westreich and Greenland, 2013] Westreich, D. and Greenland, S. (2013). The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*, 177(4):292–298.

[White et al., 2011] White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine*, 30:377–399.

[Yang et al., 2024] Yang, K., Liu, L., and Wen, Y. (2024). The impact of bayesian optimization on feature selection. *Scientific Reports*, 14(1):3948.

[Zhang et al., 2019] Zhang, J., Mucs, D., Norinder, U., and Svensson, F. (2019). Lightgbm: An effective and scalable algorithm for prediction of chemical toxicity-application to the tox21 and mutagenicity data sets. *Journal of Chemical Information and Modeling*, 59(10):4150–4158.

7 Apéndice

7.1 Conjunto de testeo

La mortalidad en promedio en la red es en el conjunto de testeo es del 21.7 %, el 5.1 % presenta malformaciones congénitas mayores y de estas el 41 % de los neonatos fallecen. El aporte de pacientes de cada país es similar al del conjunto de entrenamiento [Tabla 6].

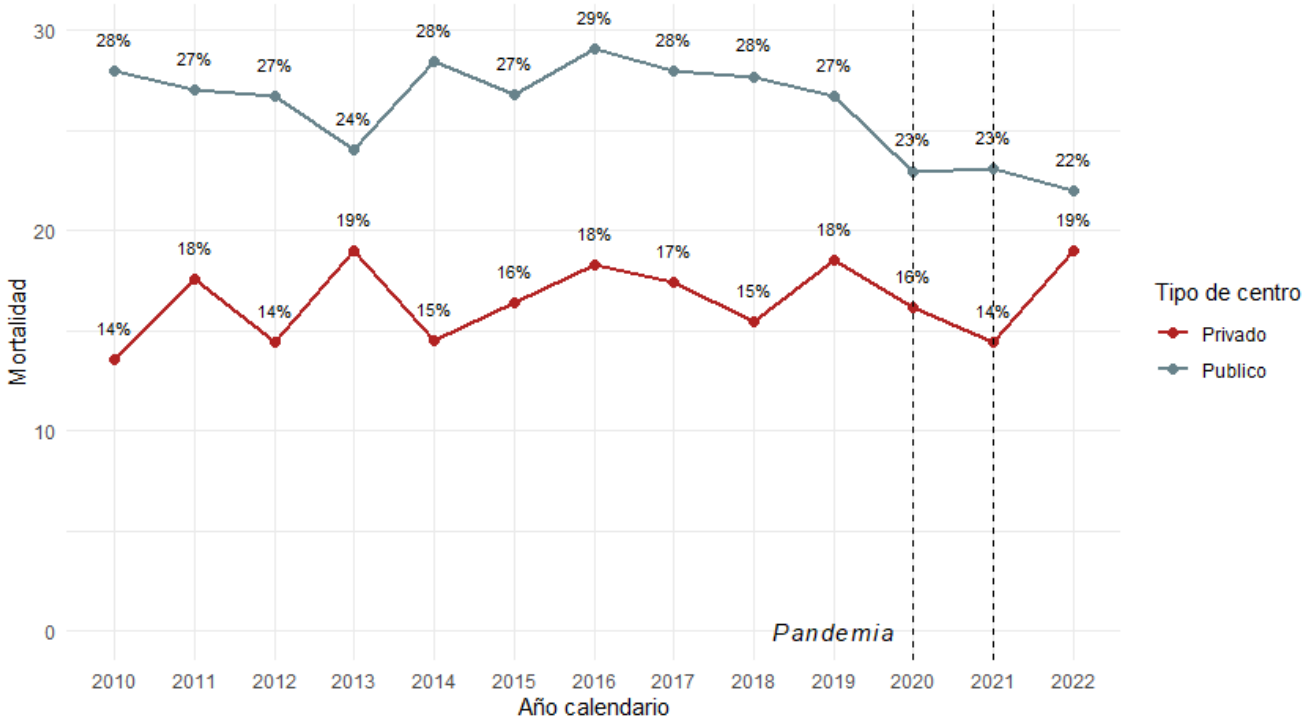
Tabla 6. Características demográficas del conjunto de testeo

	Fallece (N=797)	Vive (N=2886)	Total (N=3683)
Peso al nacimiento (gr)			
Mean (SD)	980 (308)	1120 (270)	1090 (284)
Median [Min, Max]	950 [405, 1500]	1160 [410, 1500]	1130 [405, 1500]
Tipo de centro			
Privado	121 (15.2 %)	449 (15.6 %)	570 (15.5 %)
Público	676 (84.8 %)	2437 (84.4 %)	3113 (84.5 %)
País			
Argentina	319 (40.0 %)	1124 (38.9 %)	1443 (39.2 %)
Chile	271 (34.0 %)	1080 (37.4 %)	1351 (36.7 %)
Paraguay	81 (10.2 %)	244 (8.5 %)	325 (8.8 %)
Perú	99 (12.4 %)	296 (10.3 %)	395 (10.7 %)
Uruguay	27 (3.4 %)	142 (4.9 %)	169 (4.6 %)
Apgar 1 minuto			
Mean (SD)	5.03 (2.59)	6.15 (2.23)	5.91 (2.36)
Median [Min, Max]	5.00 [0, 9.00]	7.00 [0, 9.00]	7.00 [0, 9.00]
Edad gestacional (semanas)			
Mean (SD)	27.8 (3.11)	29.0 (2.68)	28.7 (2.82)
Median [Min, Max]	27.0 [23.0, 36.0]	29.0 [23.0, 36.0]	29.0 [23.0, 36.0]
Ruptura prematura de membrana (días)			
Mean (SD)	1.56 (6.66)	1.54 (6.91)	1.55 (6.85)
Median [Min, Max]	0 [0, 77.0]	0 [0, 98.0]	0 [0, 98.0]
Corioamnionitis			
No	690 (86.6 %)	2588 (89.7 %)	3278 (89.0 %)
Sí	104 (13.0 %)	294 (10.2 %)	398 (10.8 %)
Unknown	3 (0.4 %)	4 (0.1 %)	7 (0.2 %)
Malformación congénita mayor			
No	719 (90.2 %)	2775 (96.2 %)	3494 (94.9 %)
Sí	77 (9.7 %)	111 (3.8 %)	188 (5.1 %)
Unknown	1 (0.1 %)	0 (0 %)	1 (0.0 %)
Maduración pulmonar fetal			
No recibe	178 (22.3 %)	403 (14.0 %)	581 (15.8 %)
Recibe	609 (76.4 %)	2456 (85.1 %)	3065 (83.2 %)
Unknown	10 (1.3 %)	27 (0.9 %)	37 (1.0 %)
Sulfato de Mg			
No recibe	436 (54.7 %)	1307 (45.3 %)	1743 (47.3 %)
Tratamiento	352 (44.2 %)	1536 (53.2 %)	1888 (51.3 %)
Unknown	9 (1.1 %)	43 (1.5 %)	52 (1.4 %)

	Fallece (N=797)	Vive (N=2886)	Total (N=3683)
Antibiótico prenatal			
No	527 (66.1 %)	1977 (68.5 %)	2504 (68.0 %)
Sí	269 (33.8 %)	893 (30.9 %)	1162 (31.6 %)
Unknown	1 (0.1 %)	16 (0.6 %)	17 (0.5 %)
Nivel educacional			
Primaria incompleta	37 (4.6 %)	157 (5.4 %)	194 (5.3 %)
Secundaria incompleta	193 (24.2 %)	672 (23.3 %)	865 (23.5 %)
Técnica	395 (49.6 %)	1344 (46.6 %)	1739 (47.2 %)
Universitaria	168 (21.1 %)	698 (24.2 %)	866 (23.5 %)
Unknown	4 (0.5 %)	15 (0.5 %)	19 (0.5 %)
Edad materna (años)			
Mean (SD)	29.6 (7.03)	29.5 (7.07)	29.5 (7.06)
Median [Min, Max]	29.0 [13.0, 50.0]	30.0 [13.0, 50.0]	30.0 [13.0, 50.0]
Sexo			
Femenino	343 (43.0 %)	1381 (47.9 %)	1724 (46.8 %)
Masculino	452 (56.7 %)	1502 (52.0 %)	1954 (53.1 %)
Unknown	2 (0.3 %)	3 (0.1 %)	5 (0.1 %)
Diabetes materna			
No	707 (88.7 %)	2574 (89.2 %)	3281 (89.1 %)
Sí	74 (9.3 %)	284 (9.8 %)	358 (9.7 %)
Unknown	16 (2.0 %)	28 (1.0 %)	44 (1.2 %)
Hipertensión arterial materna			
No	547 (68.6 %)	1862 (64.5 %)	2409 (65.4 %)
Sí	248 (31.1 %)	1018 (35.3 %)	1266 (34.4 %)
Unknown	2 (0.3 %)	6 (0.2 %)	8 (0.2 %)
Control prenatal			
Completo	691 (86.7 %)	2599 (90.1 %)	3290 (89.3 %)
Incompleto	95 (11.9 %)	252 (8.7 %)	347 (9.4 %)
Unknown	11 (1.4 %)	35 (1.2 %)	46 (1.2 %)
Tabaquismo			
No	702 (88.1 %)	2580 (89.4 %)	3282 (89.1 %)
Sí	41 (5.1 %)	154 (5.3 %)	195 (5.3 %)
Unknown	54 (6.8 %)	152 (5.3 %)	206 (5.6 %)

Se grafica la tendencia de mortalidad para todo el período de tiempo utilizado con el objetivo de identificar si existe un patrón excepcional durante el período de pandemia. La mortalidad por año (considerando la ocurrencia de la pandemia en el período 2020-2021) no mostró diferencias relevantes.

Figura 10. Tendencias de mortalidad anuales en conjuntos de entrenamiento y testeo combinados

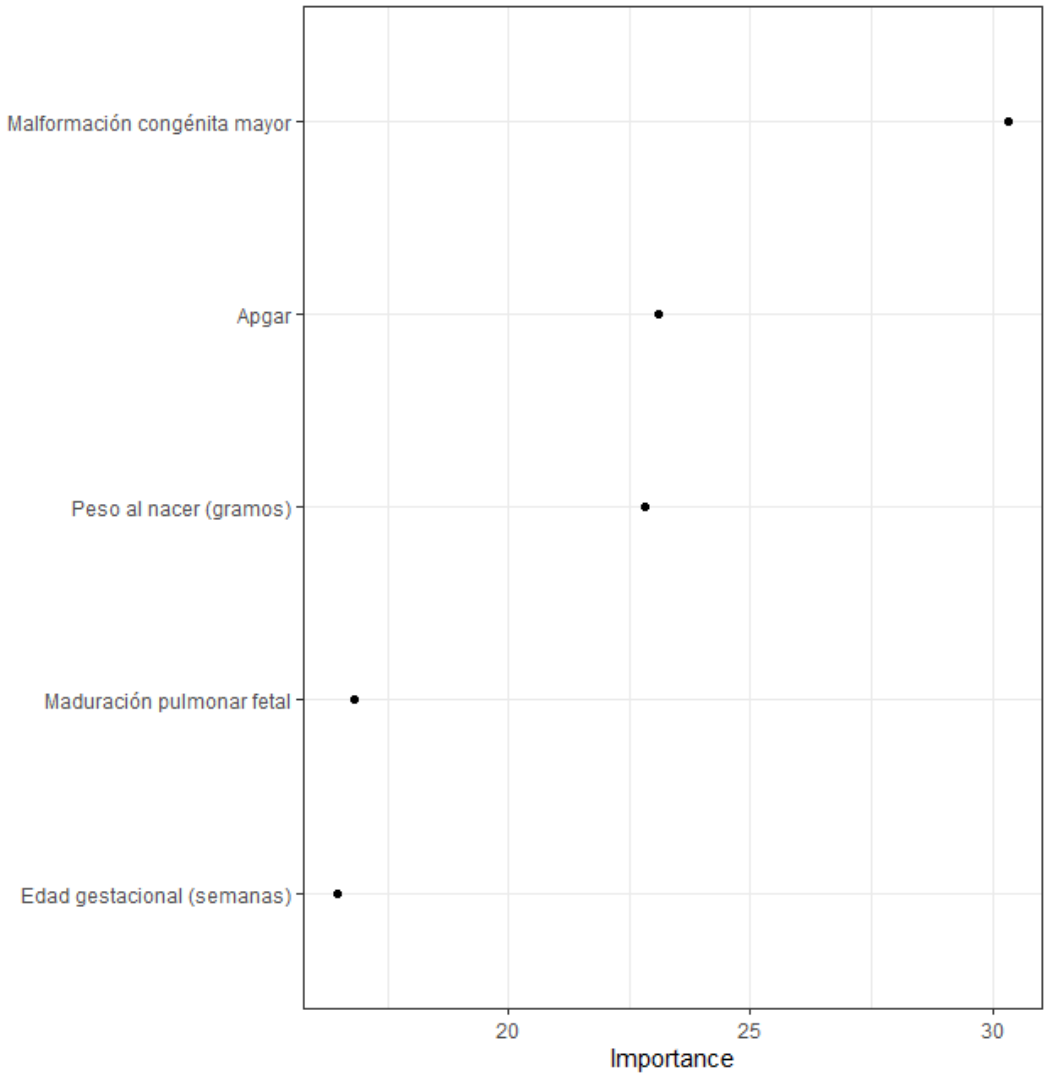


En lineas punteadas se representa el período de la pandemia

7.2 Ranking de importancia de variables

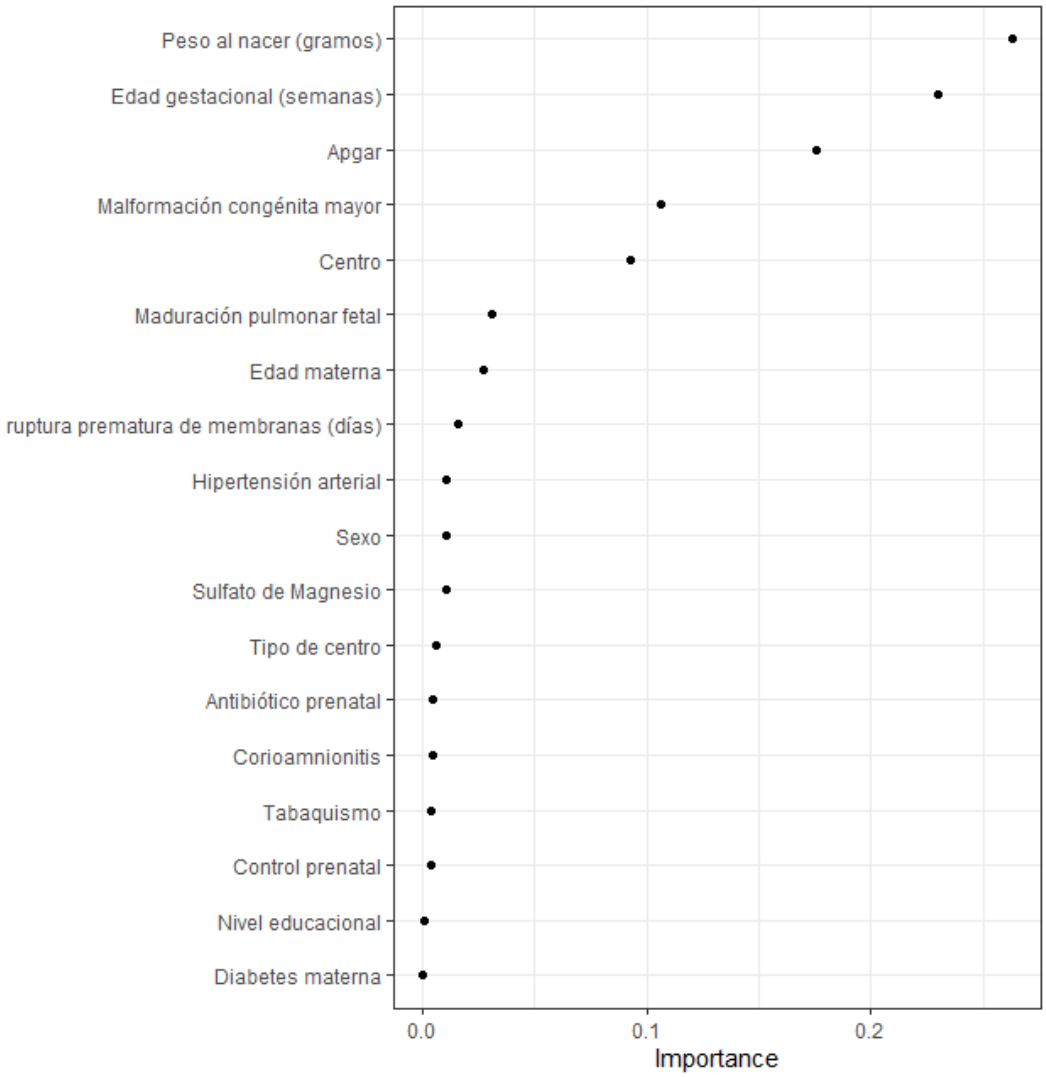
Se observa la mayor contribución de las variables numéricas de peso al nacer, apgar al minuto y edad gestacional en el modelo de LGBM en comparación con el modelo Benchmark. Esto se interpreta como una ventaja del primero a la hora de captar interacciones y no linealidades entre los predictores.

Figura 11. Ranking de importancia: LR



Importancia representada por chi-cuadrado - grados de libertad

Figura 12. Ranking de importancia: LGBM



Importancia expresada en porcentaje

7.3 Beneficio neto

Se observa el incremento del beneficio neto aportado por el modelo LGBM en relación al Benchmark para cada umbral de riesgo, a excepción del umbral de 0.9. Este último umbral no tiene relevancia alguna en la práctica clínica.

Tabla 7. Beneficio neto

Umbral	Treat all	Treat none	benchmark	LGBM
0.1	0.128	0	0.159	0.164
0.2	0.019	0	0.127	0.140
0.3	-0.122	0	0.100	0.116
0.4	-0.309	0	0.087	0.103
0.5	-0.570	0	0.067	0.083
0.6	-0.963	0	0.049	0.068
0.7	-1.617	0	0.032	0.052
0.8	-2.926	0	0.023	0.032
0.9	-6.851	0	0.011	0.005

Valores de beneficio neto por modelo para cada valor de umbral de riesgo.