



PROYECTO DE TRABAJO FINAL

Seminario de Trabajo Final

C.C.C Lic. en Ciencia de Datos

Título: "Modelado del riesgo neonatal en recién nacidos prematuros mediante aprendizaje automático supervisado utilizando datos abiertos de acceso público"

Autora:

Camila Itati Galvez – N.º DNI 43.825.006

Lugar:

Rosario

Fecha de presentación:

27/07/2026

Índice

PROYECTO DE TRABAJO FINAL.....	0
Seminario de Trabajo Final.....	0
Dedicatorias.....	2
Resumen.....	3
1. Problemática observada.....	3
2. Hipótesis de Investigación.....	3
3. Objetivo General.....	4
4. Objetivos Específicos.....	4
5. Marco teórico.....	4
5. 1. Prematuridad y clasificaciones OMS.....	4
5. 2. Variables perinatales:.....	4
5. 3. Datos abiertos en salud.....	5
5. 4. Aprendizaje automático supervisado.....	5
5. 5. Métricas de evaluación.....	6
5. 6. Metodología CRISP-DM.....	6
5. 7. Mortalidad neonatal y principales complicaciones en prematuros.....	7
6. Metodología.....	8
7. Descripción de Datasets.....	9
7. 1. Dataset 1: SPNeoDeath - Dataset Principal.....	9
7. 2. Dataset 2: Centro Médico Samsung - Dataset Complementario.....	9
7. 3. Dataset 3: Development Preterm Infant - Sección Complementaria.....	9
8. Resultados.....	10
8. 1. Análisis exploratorio y ANOVA.....	10
8. 2. Modelado: Dataset SPNeoDeath.....	12
8. 3. Modelado: Dataset Samsung.....	13
8. 4. Curvas ROC.....	13
8. 5. Importancia de variables.....	14
8. 6. Curva de aprendizaje: Dataset Mendeley.....	15
9. Despliegue: PrematurApp.....	16
10. Discusión.....	18
11. Conclusiones.....	18
12. Limitaciones y trabajo futuro.....	19
12.1. Limitaciones.....	19
12. 2. Trabajo futuro.....	19
13. Recursos disponibles.....	20
14. Impacto esperado.....	20
15. Referencias.....	21
16. Anexos.....	22
Anexo A: Reproducibilidad.....	22
A. 1. Entorno y librerías.....	22
A. 2. Acceso al código.....	22
A. 3. Acceso a los datos.....	22
A. 4. Presentación de defensa.....	23

Dedicatorias

A Dios

A mi familia, por su amor incondicional y por haberme enseñado el valor de no rendirse.

A mis amigos, por hacer más livianos los momentos difíciles y recordarme de lo que soy capaz.

A los médicos que creyeron en mí desde el comienzo y me dieron la oportunidad de dar mis primeros pasos. Este trabajo es, en parte, el reflejo de ese impulso.

Y a quienes hoy ya no están, pero hicieron posible que yo llegara hasta acá. Cada página de este trabajo lleva también un pedacito de ustedes. Dondequiera que estén, gracias.

PROYECTO DE TRABAJO FINAL:
**“Modelado del riesgo neonatal en recién nacidos prematuros
mediante aprendizaje automático supervisado utilizando datos
abiertos de acceso público”**

Resumen

En el presente trabajo se analizó el riesgo de mortalidad neonatal en recién nacidos prematuros mediante técnicas de aprendizaje automático supervisado. Se caracterizaron variables clínicas y perinatales, tales como peso al nacer, edad gestacional, puntaje APGAR y supervivencia neonatal. Los datos fueron adquiridos de repositorios públicos: el dataset SPNeoDeath, con registros de nacimientos y mortalidad neonatal del estado de São Paulo, Brasil (Beluzo et al., 2020); el dataset de prematuros extremos del Centro Médico Samsung, publicado en PLOS ONE (Park et al., 2017); y el dataset "Development Preterm Infant" de Mendeley Data (Faramarzi et al., 2023), utilizado como fuente complementaria para el análisis del desarrollo motor. A partir de la limpieza, el análisis exploratorio y la validación estadística, se entrenaron y compararon algoritmos de clasificación supervisada para predecir la mortalidad neonatal en esta población. El desempeño de los modelos fue evaluado mediante métricas de clasificación, obteniendo un AUC de 0.959 con Regresión Logística sobre el dataset principal, lo que sugiere que las variables perinatales medidas en los primeros minutos de vida tienen alto poder predictivo sobre el riesgo neonatal.

Palabras clave: *Riesgo neonatal, Prematuridad, Mortalidad neonatal, Supervivencia neonatal, Desarrollo motor, Aprendizaje automático supervisado, Datos de repositorios públicos, Modelos predictivos, Variables perinatales, APGAR.*

Temática de Interés en Ciencia de Datos: *Aprendizaje Automático Supervisado.*

1. Problemática observada

Hoy en día, la detección temprana de riesgos en recién nacidos representa un desafío en el ámbito de la salud. En particular, los recién nacidos prematuros constituyen una población vulnerable, con mayor probabilidad de presentar complicaciones a corto y largo plazo, además de una tasa de mortalidad significativa. La *Organización Mundial de la Salud (OMS, 2023)* define como prematuro a todo aquel nacido vivo antes de las 37 semanas completas de gestación.

En este contexto, si bien existe una creciente disponibilidad de datos abiertos en salud, aún es limitado el desarrollo de herramientas que permitan utilizarlos para la predicción temprana de estos riesgos. Esta limitación dificulta la generación de herramientas predictivas que permitan una intervención oportuna en poblaciones de alto riesgo.

2. Hipótesis de Investigación

“Es posible predecir el riesgo de mortalidad neonatal en recién nacidos prematuros a partir

de variables perinatales como peso al nacer, edad gestacional y puntaje APGAR, mediante modelos de aprendizaje automático supervisado, utilizando datos recolectados por el Sistema de Información sobre Nacidos Vivos de São Paulo, Brasil (SPNeoDeath), el Centro Médico Samsung (Corea del Sur) y el repositorio Mendeley Data (Faramarzi et al., 2023)."

3. Objetivo General

Construir un modelo de aprendizaje automático supervisado para la predicción del riesgo de mortalidad neonatal en recién nacidos prematuros, utilizando datos de repositorios públicos internacionales de acceso abierto (SPNeoDeath, Centro Médico Samsung y Mendeley Data).

4. Objetivos Específicos

- Analizar las variables clínicas y perinatales de los datasets seleccionados mediante análisis exploratorio y pruebas estadísticas (ANOVA) para identificar los predictores más relevantes de la mortalidad neonatal en prematuros.
- Preparar los datos mediante técnicas de limpieza, selección de variables, balanceo de clases y escalado, siguiendo la metodología CRISP-DM.
- Aplicar modelos de clasificación supervisada, incluyendo Regresión Logística, Árbol de Decisión y Random Forest, para predecir el riesgo de mortalidad neonatal en recién nacidos prematuros, midiendo su desempeño mediante métricas de clasificación como AUC, F1 y Accuracy.

5. Marco teórico

5. 1. Prematuridad y clasificaciones OMS

La prematuridad se define como el nacimiento antes de las 37 semanas completas de gestación, según los criterios establecidos por la Organización Mundial de la Salud (OMS, 2023). Es la principal causa de mortalidad neonatal a nivel mundial y una de las causas más frecuentes de morbilidad infantil a corto y largo plazo.

La OMS clasifica la prematuridad en tres categorías según la edad gestacional al momento del nacimiento (OMS, 2023):

- Prematuros extremos: menos de 28 semanas de gestación.
- Muy prematuros: entre 28 y 32 semanas de gestación.
- Prematuros moderados o tardíos: entre 32 y 37 semanas de gestación.

Esta clasificación es clínicamente relevante porque el riesgo de complicaciones y mortalidad aumenta de forma inversamente proporcional a la edad gestacional. Los prematuros extremos presentan la mayor vulnerabilidad, requiriendo internación en Unidades de Cuidados Intensivos Neonatales (UCIN) y seguimiento especializado prolongado.

5. 2. Variables perinatales:

Las **variables perinatales** son aquellas mediciones clínicas registradas en el período inmediatamente anterior, durante y posterior al nacimiento. En el contexto de la

neonatología, tres variables concentran el mayor poder predictivo sobre el riesgo neonatal (WHO, 2022):

El **peso al nacer** es el indicador más utilizado para evaluar el estado de salud del recién nacido. Se considera bajo peso al nacer cuando es inferior a 2.500 gramos, muy bajo peso cuando es inferior a 1.500 gramos, y extremadamente bajo peso cuando es inferior a 1.000 gramos. En recién nacidos prematuros, el peso al nacer tiene una correlación directa con la edad gestacional y es el predictor más fuerte de mortalidad neonatal, como se evidenció en el presente trabajo con una importancia del 38.2% en el modelo de Random Forest.

La **edad gestacional** refiere al tiempo transcurrido desde la última menstruación hasta el nacimiento, expresado en semanas. Es el criterio principal para clasificar la prematuridad y determinar el nivel de madurez orgánica del neonato.

El **puntaje o test de APGAR** es una escala de evaluación clínica propuesta por la Dra. Virginia Apgar en 1952, que valora cinco parámetros del recién nacido al primer y quinto minuto de vida: apariencia (color de piel), pulso, gesticulación (respuesta a estímulos), actividad (tono muscular) y respiración. Cada parámetro se puntúa de 0 a 2, con un máximo de 10 puntos. Un puntaje inferior a 7 al quinto minuto indica riesgo clínico significativo (Apgar, 1953).

5. 3. Datos abiertos en salud

Los datos abiertos son aquellos que pueden ser libremente accedidos, reutilizados y redistribuidos por cualquier persona, sin restricciones técnicas, legales ni económicas (Open Knowledge Foundation, 2015). La Open Knowledge Foundation define un dato como verdaderamente abierto cuando está disponible en formatos accesibles, sin barreras de acceso y bajo licencias que permitan su reutilización.

En el ámbito de la salud pública, existe una distinción importante entre datos abiertos y datos de acceso controlado. Los primeros permiten la descarga directa sin ningún requisito previo, como es el caso del dataset del Centro Médico Samsung disponible en Figshare o el dataset de Mendeley Data utilizado en este trabajo. Los segundos, si bien son públicos y gratuitos, requieren registro, acreditación o la aprobación de un comité de acceso, como ocurre con repositorios como PhysioNet o Synapse, donde se aloja el dataset SPNeoDeath.

Esta distinción es relevante desde el punto de vista metodológico, ya que condiciona la reproducibilidad del trabajo y el tiempo necesario para acceder a los datos. En el presente trabajo se utilizaron exclusivamente fuentes de acceso público y gratuito, priorizando la transparencia y permitiendo que otros investigadores puedan verificar y extender los resultados obtenidos, respetando los términos de licencia de cada repositorio.

5. 4. Aprendizaje automático supervisado

El aprendizaje automático (del inglés *machine learning*) es una rama de la inteligencia artificial que desarrolla algoritmos capaces de aprender patrones a partir de datos y realizar predicciones o clasificaciones sobre nuevos casos sin ser explícitamente programados para ello (Géron, 2022).

El aprendizaje supervisado es la modalidad en la que el algoritmo aprende a partir de un conjunto de datos etiquetados, es decir, datos donde cada registro tiene asociada una

variable objetivo conocida. En el caso de este trabajo, la variable objetivo es la mortalidad neonatal (0 = sobrevive, 1 = fallece), y el modelo aprende la relación entre las variables perinatales y ese resultado para predecirlo en casos nuevos.

Los tres algoritmos utilizados en este trabajo son:

La **Regresión Logística** es un modelo lineal que estima la probabilidad de pertenencia a una clase mediante una función sigmoide. Es interpretable, computacionalmente eficiente y especialmente efectivo cuando la relación entre las variables predictoras y la variable objetivo es predominantemente lineal.

El **Árbol de Decisión** es un modelo que divide el espacio de características mediante reglas de decisión jerárquicas, generando una estructura en forma de árbol. Es altamente interpretable visualmente pero propenso al sobreajuste cuando no se limita su profundidad.

El **Random Forest** es un modelo de conjunto que construye múltiples árboles de decisión sobre muestras aleatorias del conjunto de entrenamiento y combina sus predicciones. Es más robusto que un árbol individual y permite calcular la importancia relativa de cada variable predictora (Géron, 2022).

5. 5. Métricas de evaluación

Para evaluar el desempeño de los modelos de clasificación se utilizaron las siguientes métricas:

La **Accuracy** o exactitud mide el porcentaje de predicciones correctas sobre el total de casos. Si bien es la métrica más intuitiva, resulta engañosa en datasets desbalanceados: un modelo que prediga siempre "sin riesgo" en un dataset con 4.2% de mortalidad alcanzaría un 95.8% de accuracy sin detectar ningún caso de riesgo real.

El **F1-Score** es la media armónica entre precisión y recall. Es especialmente útil en contextos desbalanceados porque penaliza tanto los falsos positivos como los falsos negativos. En este trabajo se reportó el F1 para cada clase por separado: clase 0 (sobrevive) y clase 1 (fallece), siendo esta última la más crítica clínicamente.

El **AUC** (Área Bajo la Curva ROC) mide la capacidad del modelo para distinguir entre las dos clases independientemente del umbral de clasificación. Un valor de 0.5 equivale a una predicción aleatoria, mientras que un valor de 1.0 representa una clasificación perfecta. Es la métrica principal de este trabajo por su robustez frente al desbalance de clases.

Las **Curvas ROC** grafican la tasa de verdaderos positivos versus la tasa de falsos positivos para distintos umbrales de clasificación, permitiendo comparar visualmente el desempeño de múltiples modelos.

5. 6. Metodología CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) es una metodología estándar para el desarrollo de proyectos de minería de datos y aprendizaje automático, desarrollada en 1996 y documentada por Chapman et al. (2000). Se convirtió en el proceso más utilizado en la industria y la academia por su estructura iterativa y su independencia del dominio de aplicación.

El proceso se organiza en siete etapas, aplicadas en el presente trabajo de la siguiente manera:

- 1. Recolección de datos:** Obtención de los datasets SPNeoDeath y Samsung Medical Center desde repositorios públicos de acceso abierto para el modelado principal de mortalidad neonatal. El dataset Development Preterm Infant (Mendeley Data) fue incorporado como análisis complementario para explorar la posibilidad de predecir el desarrollo motor grueso en prematuros sobrevivientes, reconociendo que el tamaño muestral reducido ($n=89$) limita la generalización de los resultados."
- 2. Limpieza y preprocesamiento:** Tratamiento de valores faltantes, estandarización de variables, codificación de categóricas y balanceo de clases mediante oversampling.
- 3. Análisis exploratorio (EDA):** Exploración y caracterización de las variables mediante boxplots, matrices de correlación y pruebas ANOVA para validar su capacidad discriminativa.
- 4. Selección de variables:** Identificación de las variables perinatales con mayor poder predictivo sobre la mortalidad neonatal.
- 5. Entrenamiento del modelo:** Aplicación de tres algoritmos de clasificación supervisada: Regresión Logística, Árbol de Decisión y Random Forest.
- 6. Evaluación:** Comparación de modelos mediante métricas de clasificación (AUC, F1, Accuracy) y curvas ROC. Se contempla también el análisis de Recall como métrica complementaria de relevancia clínica.
- 7. Despliegue:** Implementación del prototipo web PrematurApp, desplegado en Streamlit Community Cloud y accesible desde cualquier dispositivo con conexión a internet.

5. 7. Mortalidad neonatal y principales complicaciones en prematuros

La mortalidad neonatal se define como el fallecimiento del recién nacido durante los primeros 28 días de vida (OMS, 2023). Es la principal consecuencia del nacimiento prematuro y está fuertemente asociada a la inmadurez orgánica de los sistemas respiratorio, inmunológico y neurológico.

Entre las complicaciones más frecuentes en prematuros se encuentran:

- **Síndrome de dificultad respiratoria (SDR):** deficiencia de surfactante pulmonar que dificulta la respiración autónoma, especialmente en menores de 32 semanas.
- **Displasia broncopulmonar (DBP):** daño pulmonar crónico secundario al uso prolongado de oxígeno suplementario o ventilación mecánica.
- **Corioamnionitis:** infección de las membranas fetales que aumenta el riesgo de parto prematuro y sepsis neonatal.
- **Hemorragia intraventricular:** sangrado en el cerebro del neonato, más frecuente en prematuros extremos, con potencial impacto en el desarrollo neuromotor.

Estas complicaciones justifican la necesidad de herramientas de detección temprana que permitan anticipar el riesgo clínico desde los primeros minutos de vida (OMS, 2023).

6. Metodología

El presente trabajo siguió la metodología CRISP-DM (Chapman et al., 2000), estructurada en siete etapas que guiaron el desarrollo del proyecto desde la comprensión del problema hasta el despliegue de la herramienta:

En la **etapa de recolección de datos** se identificaron y adquirieron tres datasets de repositorios públicos internacionales: SPNeoDeath (Synapse), con 1.435.834 registros de nacimientos y mortalidad neonatal del estado de São Paulo, Brasil, del cual se extrajeron 141.977 registros correspondientes a recién nacidos prematuros; el dataset del Centro Médico Samsung (Figshare), con 382 registros de prematuros extremos (23-26 semanas); y el dataset "Development Preterm Infant" de Mendeley Data, con 89 registros clínicos de UCIN, utilizado como fuente complementaria para el análisis del desarrollo motor. Se evaluaron adicionalmente registros estadísticos del DEIS de Argentina y del dataset de Kaggle "Fetal Health Classification", sin embargo no fueron incorporados al modelado: el primero porque presenta datos agregados y no individuales, incompatibles con el aprendizaje supervisado; el segundo porque corresponde a una población y un momento clínico diferente al objeto de estudio.

En la **etapa de limpieza y preprocesamiento** se realizó el tratamiento de valores nulos, la estandarización de variables numéricas mediante StandardScaler, y el balanceo de clases a través de oversampling para compensar el desbalance natural de la variable objetivo (4.2% de mortalidad neonatal en prematuros).

En la **etapa de análisis exploratorio (EDA)** se caracterizaron las variables perinatales mediante boxplots y matrices de correlación, y se aplicaron pruebas ANOVA para validar estadísticamente su capacidad discriminativa entre los grupos de supervivencia y mortalidad.

En la **etapa de selección de variables** se identificaron las ocho variables con mayor poder predictivo: semanas gestacionales, peso al nacer, puntaje APGAR al primer y quinto minuto, tipo de parto, presencia de malformación congénita, consultas prenatales y edad materna.

En la **etapa de entrenamiento** se aplicaron tres algoritmos de clasificación supervisada: Regresión Logística, Árbol de Decisión y Random Forest, sobre los datasets SPNeoDeath y Samsung Medical Center. Como análisis complementario, se entrenó un modelo Random Forest sobre el dataset Mendeley (n=89) para explorar la predicción del desarrollo motor grueso en prematuros sobrevivientes, y se generó una curva de aprendizaje para analizar el comportamiento del modelo en función del tamaño muestral.

En la **etapa de evaluación** se compararon los modelos mediante métricas de clasificación (AUC, F1 y Accuracy) y curvas ROC. Se contempla el Recall como métrica complementaria de relevancia clínica, dado que minimizar los falsos negativos es prioritario en un contexto de predicción de mortalidad neonatal.

En la etapa de **despliegue** se desarrolló e implementó PrematurApp, una herramienta web de apoyo a la decisión clínica desplegada en Streamlit Community Cloud, accesible de forma gratuita desde cualquier dispositivo con conexión a internet.

7. Descripción de Datasets

7. 1. Dataset 1: SPNeoDeath - Dataset Principal

El dataset SPNeoDeath fue desarrollado por Beluzo et al. (2020) a partir de la integración del Sistema de Información sobre Nacidos Vivos (SINASC) y el Sistema de Información de Mortalidad (SIM) del estado de São Paulo, Brasil, abarcando el período 2012-2018. Contiene 1.435.834 registros individuales de nacimientos, de los cuales se extrajeron 141.977 correspondientes a recién nacidos prematuros (menos de 37 semanas de gestación).

El dataset incluye 24 variables organizadas en tres categorías: variables sociodemográficas de la madre (edad, nivel educativo, estado civil), variables obstétricas (tipo de parto, número de consultas prenatales, semanas de gestación) y variables perinatales del neonato (peso al nacer, puntaje APGAR al primer y quinto minuto, presencia de malformaciones congénitas). La variable objetivo es `neonatal_death`, una variable binaria que indica si el recién nacido falleció durante los primeros 28 días de vida (Beluzo et al., 2020).

7. 2. Dataset 2: Centro Médico Samsung - Dataset Complementario

El dataset del Centro Médico Samsung fue publicado como material suplementario del estudio de Park et al., 2017 en la revista PLOS ONE, disponible en el repositorio Figshare. Contiene 382 registros de recién nacidos prematuros extremos, nacidos entre las semanas 23 y 26 de gestación, en el Centro Médico Samsung de Corea del Sur.

El dataset incluye 44 variables clínicas que abarcan datos obstétricos (tipo de parto, uso de corticoide antenatal, presencia de corioamnionitis), variables perinatales (peso al nacer, puntaje APGAR al primer minuto, sexo) y variables de seguimiento temporal (mortalidad por períodos: día 1, días 2-7, días 8-28 y posterior a los 28 días). La variable objetivo utilizada en este trabajo es `outcome`, una variable binaria que indica supervivencia o fallecimiento del neonato.

A diferencia del SPNeoDeath, este dataset se focaliza exclusivamente en prematuros extremos, lo que lo convierte en un complemento clínicamente valioso para analizar el subgrupo de mayor vulnerabilidad. Su tamaño reducido ($n=382$) limita la generalización de los modelos pero permite una caracterización clínica más detallada (Park et al., 2017).

7. 3. Dataset 3: Development Preterm Infant - Sección Complementaria

El dataset "Development Preterm Infant" fue publicado por Faramarzi et al. (2023) en el repositorio Mendeley Data. Contiene 89 registros de recién nacidos prematuros hospitalizados en Unidades de Cuidados Intensivos Neonatales (UCIN), con variables clínicas de seguimiento longitudinal.

El dataset incluye variables perinatales como edad gestacional, peso al nacer, puntaje APGAR, tipo de reanimación y duración de la hospitalización, junto con variables de evaluación del desarrollo neuromotor: motricidad gruesa, motricidad fina, dominio cognitivo,

dominio del lenguaje y dominio perceptual. La variable objetivo utilizada en este trabajo es **coarse.cat**, que clasifica el desarrollo motor grueso del neonato en dos categorías: con riesgo y sin riesgo.

Este dataset fue utilizado como sección complementaria del trabajo para explorar una dimensión diferente del riesgo neonatal: no la mortalidad, sino el desarrollo motor a futuro. Mediante una curva de aprendizaje se analizó el comportamiento del modelo en función del tamaño muestral, evidenciando que con 89 registros el modelo no alcanza su techo predictivo, lo que sugiere que un dataset de mayor volumen con variables de seguimiento motor podría mejorar significativamente los resultados (Faramarzi et al., 2023).

8. Resultados

8. 1. Análisis exploratorio y ANOVA

El análisis bivariado realizado sobre el dataset SPNeoDeath reveló diferencias estadísticamente significativas entre los grupos de mortalidad y supervivencia para las cuatro variables perinatales analizadas. Las pruebas ANOVA arrojaron valores F extremadamente altos debido al tamaño muestral ($n=141.977$), con $p=0$ en todos los casos: semanas gestacionales ($F=38.432$), peso al nacer ($F=21.346$), APGAR al primer minuto ($F=27.492$) y APGAR al quinto minuto ($F=32.355$). Estos resultados confirman que las cuatro variables presentan poder discriminativo significativo entre los grupos, validando su inclusión como predictores en el modelo (Figura 1).

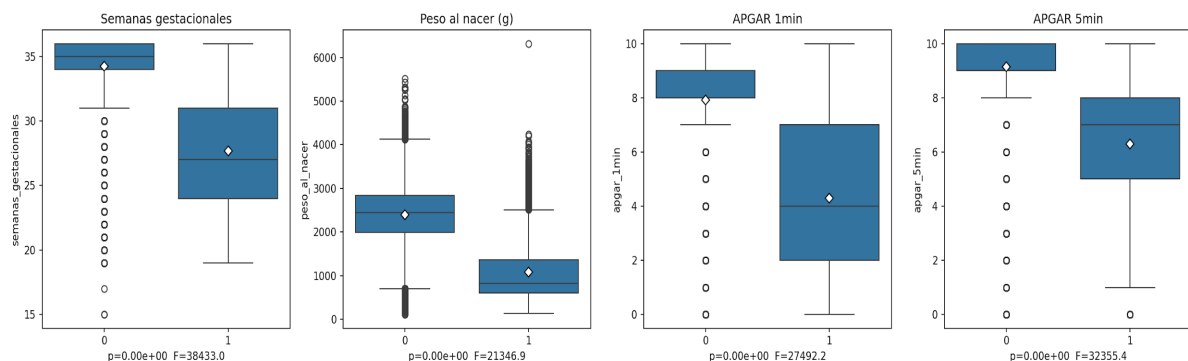


Figura 1: Análisis bivariado - SPNeoDeath. Distribución de semanas gestacionales, peso al nacer, APGAR 1min y APGAR 5min según mortalidad neonatal (0=vivo, 1=fallecido). El diamante indica la media. p-valores y estadísticos F reportados debajo de cada gráfico.

La matriz de correlación entre las variables numéricas del dataset evidenció una correlación positiva alta entre peso al nacer y semanas gestacionales ($r=0.72$), y correlaciones negativas moderadas entre todas las variables perinatales y la mortalidad neonatal, siendo las semanas gestacionales la de mayor correlación negativa con el target ($r=-0.46$). La edad materna no mostró correlación relevante con la mortalidad ($r=-0.03$) (Figura 2).

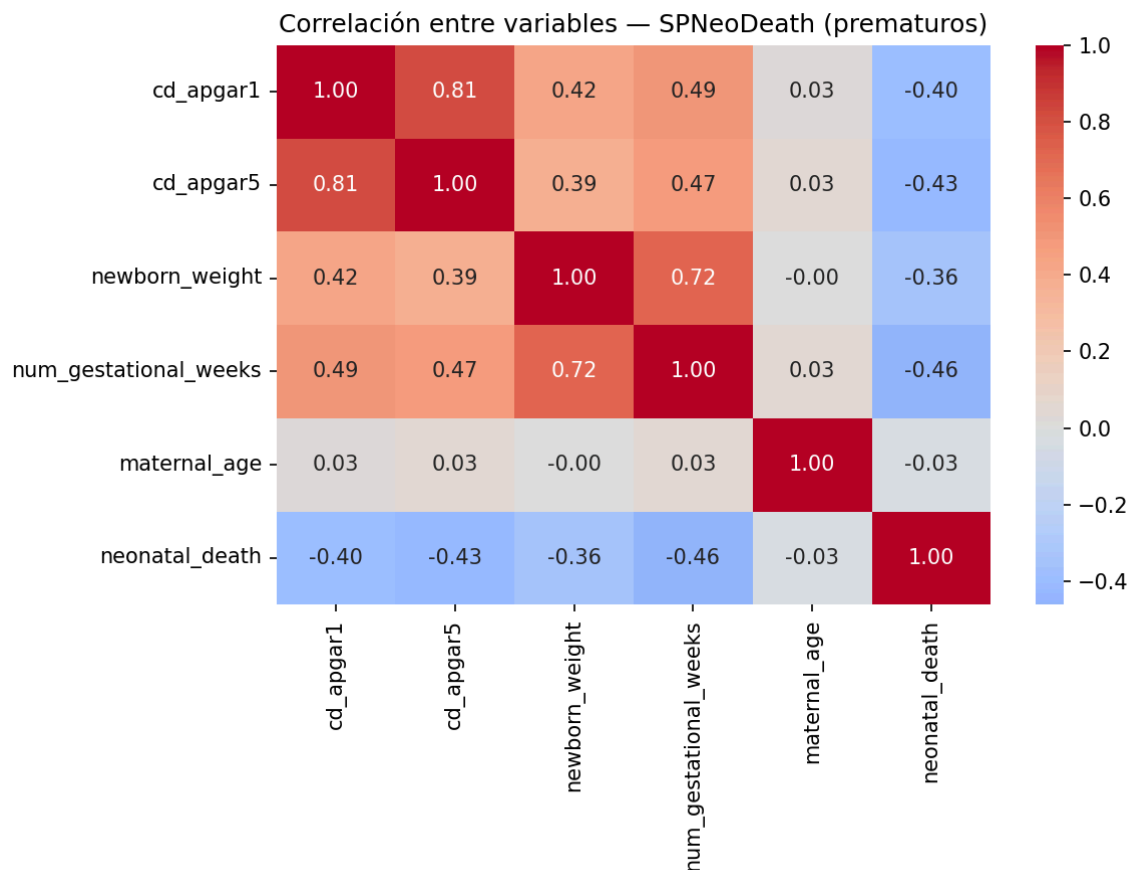


Figura 2: Matriz de correlación entre variables numéricas - SPNeoDeath (prematurados).

El mismo análisis aplicado sobre el dataset Samsung confirmó el patrón en una población de prematuros extremos (23-26 semanas), con diferencias significativas en las tres variables analizadas: semanas gestacionales ($F=18.7$, $p<0.001$), peso al nacer ($F=52.3$, $p<0.001$) y APGAR al primer minuto ($F=26.8$, $p<0.001$) (Figura 3).

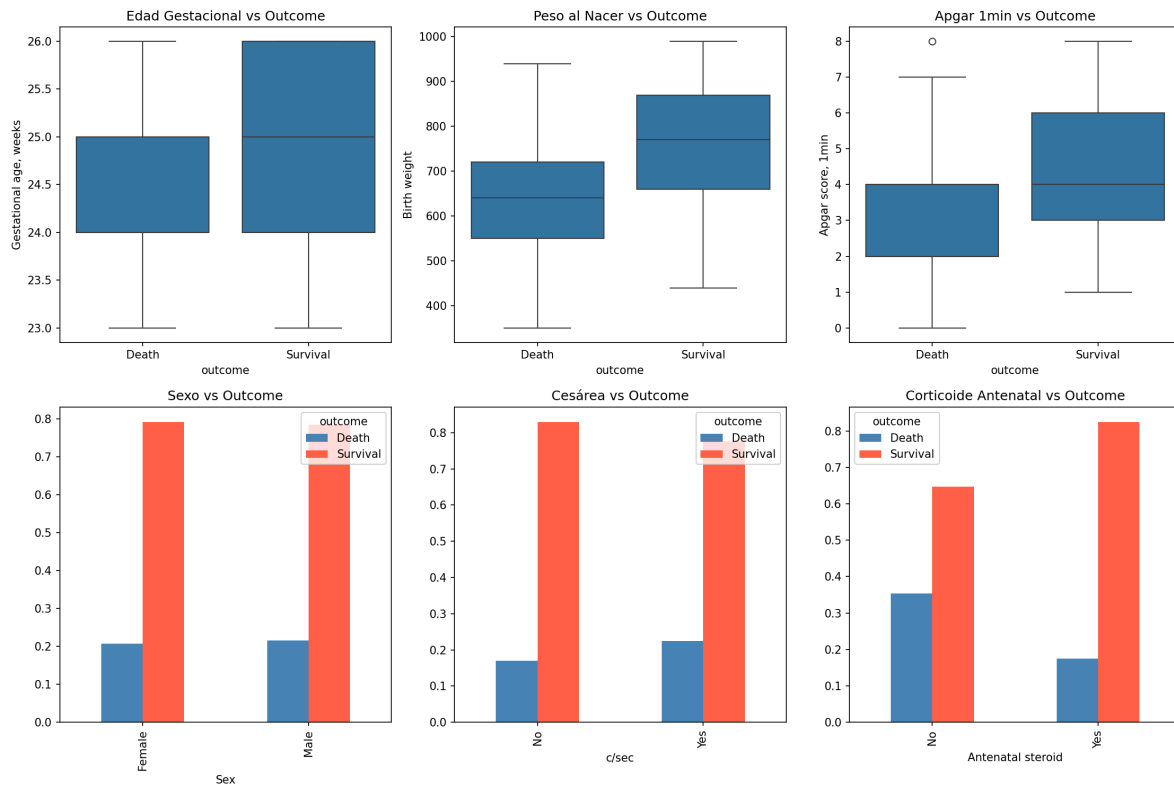


Figura 3: Análisis exploratorio - Samsung. Distribución de variables continuas (semanas gestacionales, peso al nacer y APGAR 1min) y categóricas (sexo, cesárea y corticoide antenatal) según outcome (Death/Survival).

8. 2. Modelado: Dataset SPNeoDeath

Se entrenaron tres modelos de clasificación supervisada sobre el dataset SPNeoDeath, utilizando un conjunto de entrenamiento balanceado mediante oversampling (108.732 registros por clase) y un conjunto de prueba con distribución real (27.184 registros sin mortalidad y 1.206 con mortalidad).

La Regresión Logística obtuvo el mayor AUC (0.959), lo que sugiere que la relación entre las variables perinatales y la mortalidad neonatal tiene una componente predominantemente lineal. El Random Forest alcanzó un AUC de 0.941 y el mejor F1 para la clase de riesgo (0.536). El Árbol de Decisión presentó el menor AUC (0.682), evidenciando sobreajuste al conjunto de entrenamiento. El F1 bajo para la clase 1 en todos los modelos es esperable dado el desbalance real del conjunto de prueba, donde los casos de mortalidad representan el 4.2% de los registros.

Modelo	Accuracy	F1 clase 0	F1 clase 1	AUC
Regresión Logística	0.897	0.943	0.427	0.959
Árbol de Decisión	0.953	0.975	0.410	0.682

Random Forest	0.962	0.980	0.536	0.941
---------------	-------	-------	-------	-------

Tabla 1: Resultados de clasificación - Dataset SPNeoDeath.

Cabe destacar que el Árbol de Decisión fue entrenado sin restricción de profundidad ni ajuste de hiperparámetros, lo cual favorece el sobreajuste observado. Este resultado es consistente con el comportamiento esperado de este tipo de modelo en ausencia de poda, más que un hallazgo específico sobre los datos analizados.

8. 3. Modelado: Dataset Samsung

Se entrenaron los mismos tres modelos sobre el dataset Samsung, utilizando un conjunto de entrenamiento balanceado (236 registros por clase) y un conjunto de prueba con distribución real (60 registros sin mortalidad y 16 con mortalidad). Los resultados fueron considerablemente más bajos que los obtenidos sobre SPNeoDeath. La Regresión Logística nuevamente fue el modelo con mejor AUC (0.756), seguida por Random Forest (0.683). El Árbol de Decisión obtuvo un AUC de 0.527, prácticamente equivalente a una clasificación aleatoria. La menor performance general se explica por el tamaño reducido del dataset ($n=377$), que limita la capacidad de generalización de los modelos y genera alta varianza en las métricas de evaluación.

Modelo	Accuracy	F1 clase 0	F1 clase 1	AUC
Regresión Logística	0.724	0.804	0.533	0.756
Árbol de Decisión	0.724	0.832	0.222	0.527
Random Forest	0.711	0.825	0.154	0.683

Tabla 2: Resultados de clasificación - Dataset Samsung.

8. 4. Curvas ROC

Las curvas ROC permiten visualizar el desempeño de los modelos para distintos umbrales de clasificación, complementando los valores de AUC reportados en las tablas anteriores (Figura 4).

En el dataset SPNeoDeath, las curvas de Regresión Logística y Random Forest se ubican próximas al ángulo superior izquierdo, confirmando su alta capacidad discriminativa. El Árbol de Decisión se aproxima a la diagonal, lo que es consistente con su bajo AUC (0.682) y

refleja sobreajuste al conjunto de entrenamiento.

En el dataset Samsung, las curvas presentan un patrón escalonado característico de conjuntos de prueba pequeños ($n=76$), lo que genera inestabilidad en las métricas pero no invalida los resultados. La Regresión Logística mantiene la curva más alejada de la diagonal en ambos datasets, consolidándose como el modelo con mejor capacidad discriminativa general.

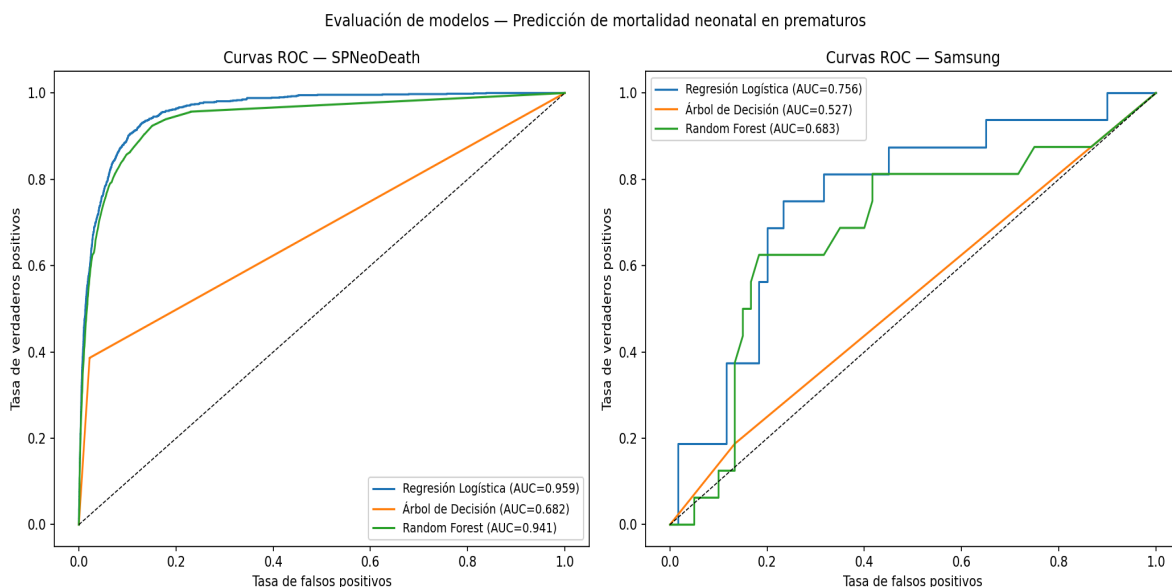


Figura 4: Curvas ROC para los tres modelos evaluados sobre SPNeoDeath (izquierda) y Samsung (derecha).

8. 5. Importancia de variables

El análisis de importancia de variables del modelo Random Forest entrenado sobre SPNeoDeath permitió identificar la contribución relativa de cada predictor en la clasificación del riesgo de mortalidad neonatal (Figura 5).

El peso al nacer resultó la variable más importante con una contribución del 38.2%, casi el doble que la segunda variable en importancia. Las semanas gestacionales aportaron el 16.1%, seguidas por el APGAR al primer minuto (15.2%) y el APGAR al quinto minuto (10.9%). En conjunto, estas cuatro variables perinatales concentran el 80.4% del poder predictivo del modelo, confirmando que la información disponible en los primeros minutos de vida es suficiente para estimar el riesgo de mortalidad neonatal con alta precisión.

Las variables restantes como edad materna (8.7%), presencia de malformación congénita (6.2%), consultas prenatales (3.7%) y tipo de parto (0.9%) mostraron una contribución marginal, siendo el tipo de parto prácticamente irrelevante para el modelo.

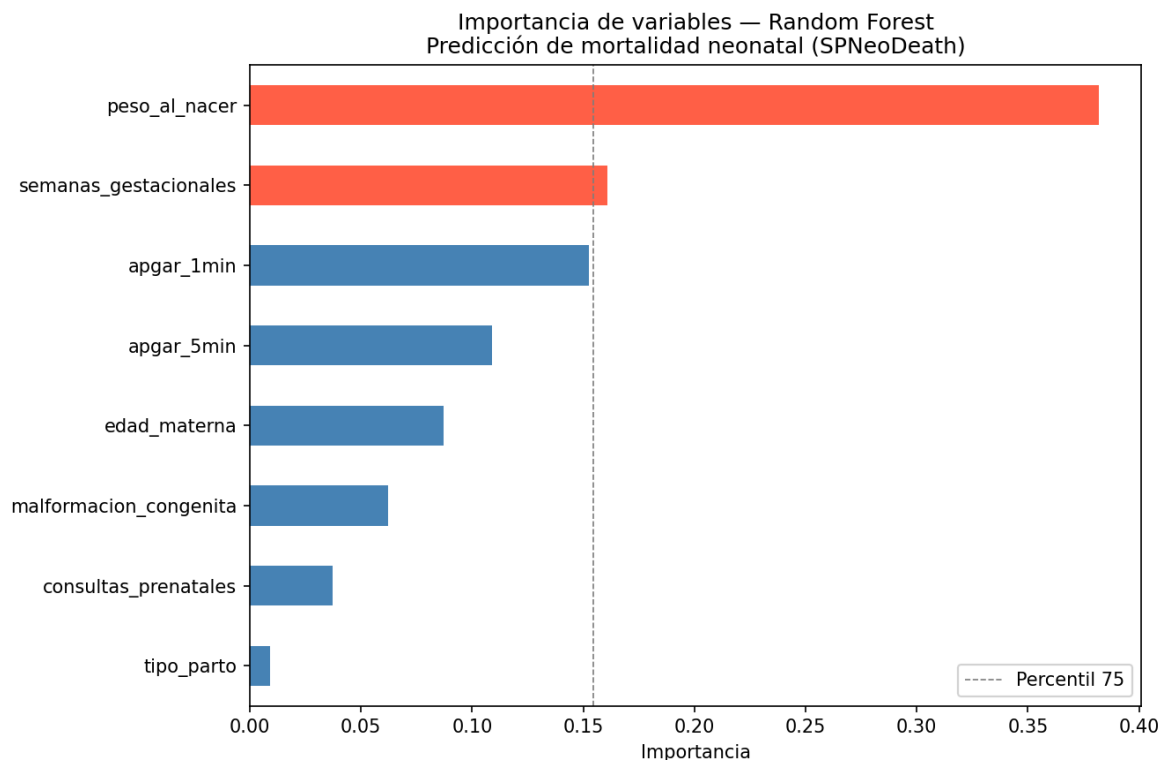


Figura 5: Importancia de variables: Random Forest sobre SPNeoDeath. Las barras en rojo corresponden a las variables por encima del percentil 75 de importancia.

8. 6. Curva de aprendizaje: Dataset Mendeley

Como análisis complementario, se generó una curva de aprendizaje sobre el dataset Mendeley (n=89) para explorar el comportamiento del modelo Random Forest en función del tamaño muestral aplicado a la predicción del desarrollo motor en prematuros (Figura 6).

La curva de entrenamiento se mantiene en AUC=1.0 a lo largo de todos los tamaños evaluados, lo que indica que el modelo memoriza perfectamente los datos de entrenamiento. La curva de validación cruzada alcanza un AUC aproximado de 0.87 con el máximo de datos disponibles, pero sin llegar a estabilizarse: la tendencia ascendente en el extremo derecho de la curva sugiere que el modelo aún podría mejorar con más datos.

La brecha persistente entre ambas curvas refleja sobreajuste característico de muestras pequeñas, donde el modelo no cuenta con suficientes registros para generalizar correctamente. Este resultado no invalida el análisis sino que evidencia una limitación inherente al tamaño del dataset y abre una línea de investigación futura orientada a la construcción de datasets clínicos de mayor escala con variables de seguimiento motor en prematuros.

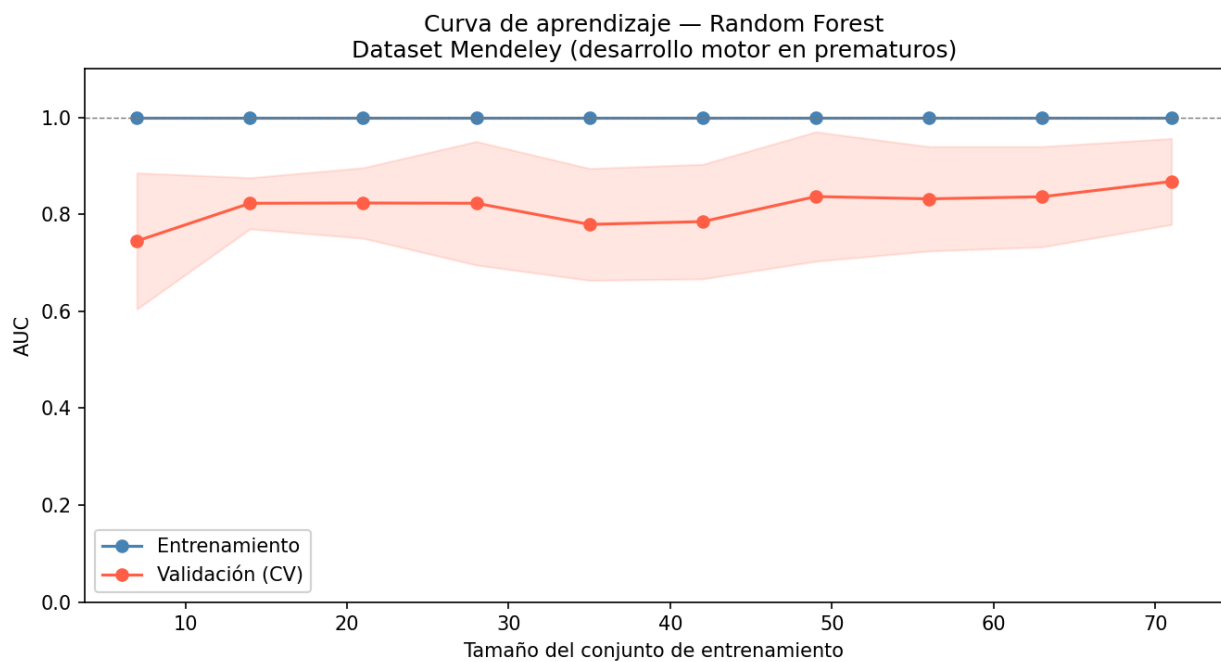


Figura 6: Curva de aprendizaje: Random Forest sobre dataset Mendeley. La línea azul representa el AUC de entrenamiento y la línea roja el AUC de validación cruzada (CV). El área sombreada indica la desviación estándar entre folds.

9. Despliegue: PrematurApp

Como resultado de la fase de despliegue de la metodología CRISP-DM, se desarrolló PrematurApp, una aplicación web de apoyo a la decisión clínica disponible en: <https://prematurapp.streamlit.app/>

La herramienta permite al personal de salud ingresar las variables perinatales del recién nacido al momento del nacimiento y obtener una estimación del riesgo de mortalidad neonatal en tiempo real. El formulario solicita: semanas gestacionales, peso al nacer, puntaje APGAR al primer y quinto minuto, tipo de parto, presencia de malformación congénita, cantidad de consultas prenatales y edad materna.

El modelo devuelve una probabilidad de riesgo clasificada en tres niveles clínicos: riesgo bajo (menor al 10%), riesgo moderado (entre 10% y 50%) y riesgo alto (mayor al 50%), acompañada de una recomendación de acción clínica correspondiente a cada nivel.

PrematurApp está diseñada específicamente para su uso en hospitales públicos de baja complejidad, donde el acceso a herramientas diagnósticas especializadas suele ser limitado. Al requerir únicamente variables disponibles en los primeros minutos de vida y funcionar en cualquier dispositivo con conexión a internet, representa una herramienta accesible y de bajo costo de implementación.

La herramienta tiene fines de apoyo a la decisión clínica y no reemplaza el criterio del profesional de la salud.

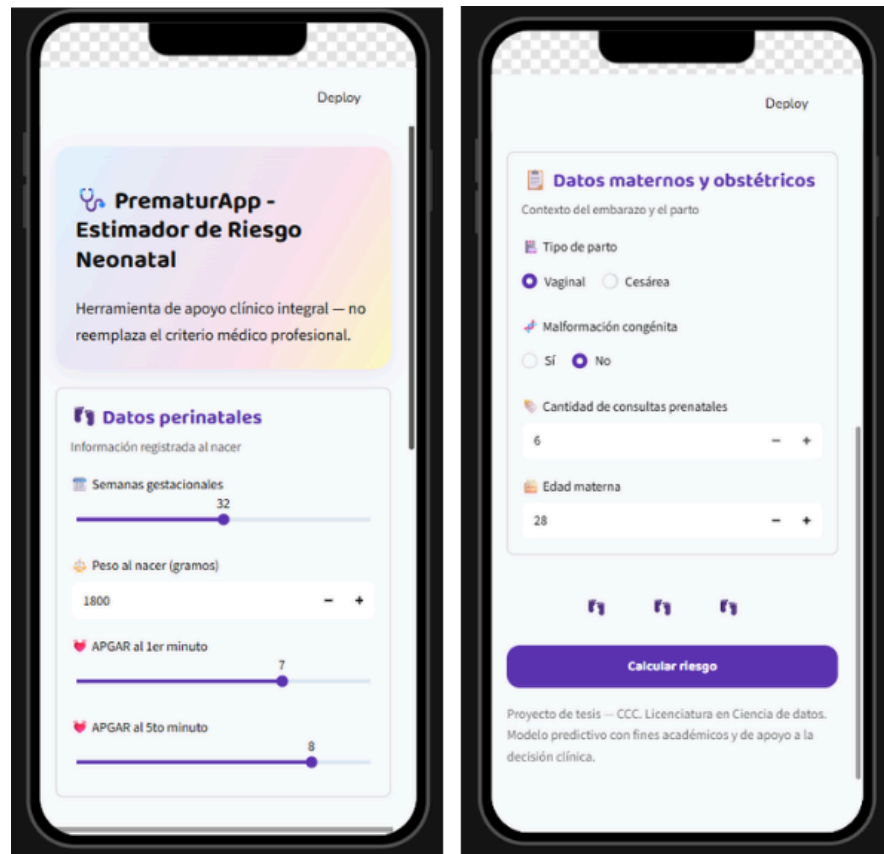


Figura 7: Vista mobile de PrematurApp

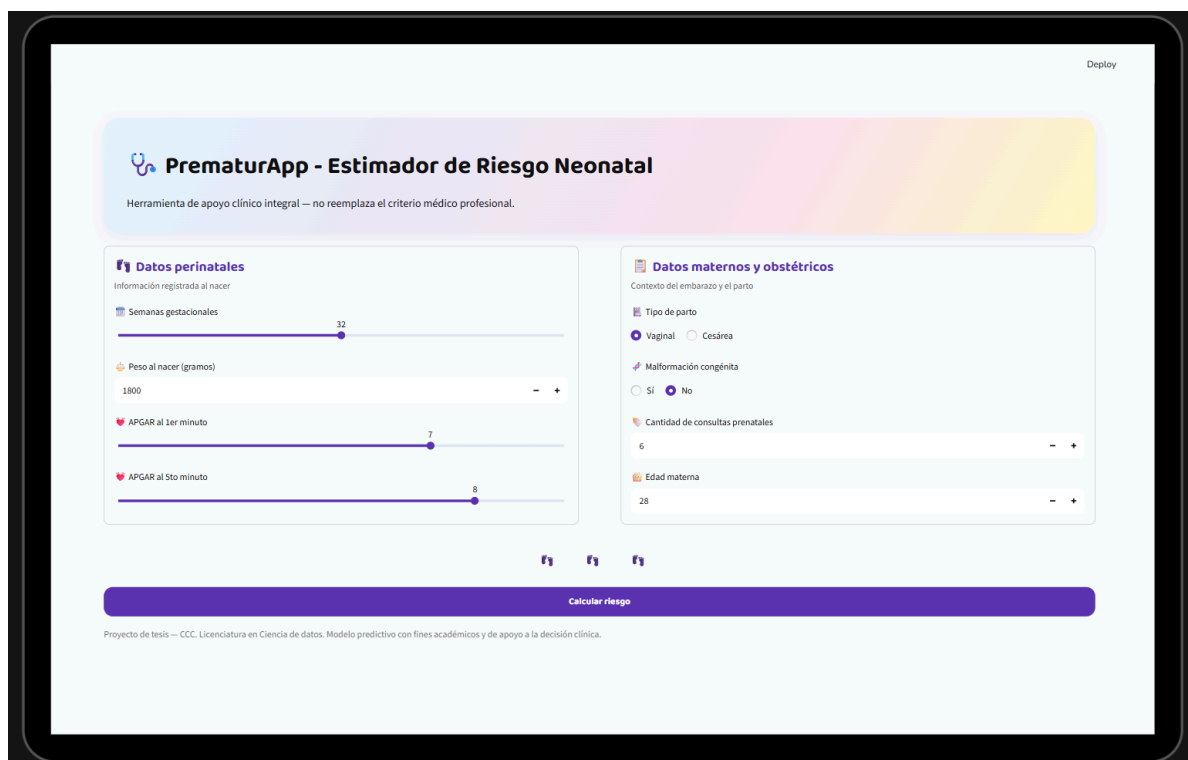


Figura 8: Vista desktop de PrematurApp

10. Discusión

Los resultados obtenidos en este trabajo permiten validar la hipótesis de investigación: es posible predecir el riesgo de mortalidad neonatal en recién nacidos prematuros a partir de variables perinatales mediante modelos de aprendizaje automático supervisado. El modelo de Regresión Logística entrenado sobre el dataset SPNeoDeath alcanzó un AUC de 0.959, resultado consistente con los reportados en la literatura para problemas similares. Chiavegatto Filho et al. (2025) obtuvieron resultados comparables utilizando el mismo dataset con arquitecturas más complejas, lo que sugiere que la relación entre las variables perinatales y la mortalidad neonatal tiene una componente predominantemente lineal que los modelos simples capturan eficientemente.

El hallazgo más relevante del análisis de importancia de variables es que las cuatro variables perinatales medidas en los primeros minutos de vida, peso al nacer, semanas gestacionales, APGAR al primer y quinto minuto, concentran el 80.4% del poder predictivo del modelo. Esto es clínicamente significativo porque implica que la información necesaria para estimar el riesgo está disponible de forma inmediata al nacimiento, sin requerir estudios complementarios ni equipamiento especializado. Este resultado respalda la viabilidad de implementar la herramienta en hospitales públicos de baja complejidad.

La diferencia de desempeño entre los dos datasets principales merece atención. Mientras que sobre SPNeoDeath ($n=141.977$) los modelos alcanzaron AUC superiores a 0.94, sobre Samsung ($n=377$) el máximo fue 0.756. Esta brecha no refleja necesariamente una diferencia en la calidad predictiva de las variables, sino el efecto directo del tamaño muestral sobre la capacidad de generalización de los modelos. La curva de aprendizaje generada sobre el dataset Mendeley refuerza esta interpretación, evidenciando que con muestras pequeñas el modelo sobreajusta y no alcanza su potencial predictivo.

El análisis del dataset Samsung aportó además un hallazgo clínico relevante: la administración de corticoide antenatal mostró una asociación marcada con la supervivencia neonatal, siendo los neonatos que no lo recibieron los que presentaron mayor mortalidad. Este resultado es consistente con la evidencia médica establecida sobre el beneficio de los corticoides en la maduración pulmonar fetal y refuerza la validez clínica del análisis exploratorio realizado.

Finalmente, la comparación entre datasets provenientes de distintas regiones geográficas, São Paulo (Brasil) y Corea del Sur, mostró patrones consistentes en las variables perinatales analizadas, lo que sugiere que los modelos entrenados sobre datos latinoamericanos podrían ser aplicables en contextos similares, incluyendo Argentina.

11. Conclusiones

El presente trabajo demostró que es posible predecir el riesgo de mortalidad neonatal en recién nacidos prematuros a partir de variables perinatales disponibles en los primeros minutos de vida, mediante modelos de aprendizaje automático supervisado entrenados sobre datos de repositorios públicos de acceso abierto. La hipótesis de investigación fue validada satisfactoriamente.

El modelo de Regresión Logística entrenado sobre el dataset SPNeoDeath obtuvo el mejor desempeño general con un AUC de 0.959, evidenciando que la relación entre las variables perinatales y la mortalidad neonatal tiene una naturaleza predominantemente lineal. El Random Forest complementó este resultado con el mejor F1 para la clase de riesgo (0.536), siendo ambos modelos superiores al Árbol de Decisión en todos los datasets evaluados.

El análisis de importancia de variables confirmó que el peso al nacer, las semanas gestacionales y el puntaje APGAR al primer y quinto minuto concentran el 80.4% del poder predictivo del modelo. Estas variables son universalmente registradas al momento del nacimiento, lo que hace viable la implementación de la herramienta en entornos con recursos limitados.

La comparación entre datasets de distinto volumen evidenció que el tamaño muestral es un factor determinante en la capacidad predictiva de los modelos. El análisis complementario sobre el dataset Mendeley mediante curva de aprendizaje sugiere que el desarrollo de datasets clínicos de mayor escala con variables de seguimiento motor podría abrir una línea de investigación futura de alto valor para la neonatología.

Finalmente, la consistencia de los patrones observados en datasets de Brasil y Corea del Sur refuerza la generalización de los resultados y su potencial aplicabilidad en el contexto de la salud pública argentina.

12. Limitaciones y trabajo futuro

12.1. Limitaciones

El principal desafío del trabajo fue la disponibilidad de datos clínicos individuales de prematuros con variables de seguimiento longitudinal. Los datasets de mayor volumen disponibles públicamente, como SPNeoDeath, provienen de sistemas de información de salud que no incluyen seguimiento post-alta, lo que limita el análisis a la mortalidad neonatal temprana y no permite modelar complicaciones a largo plazo como el desarrollo motor o neurológico.

El tamaño reducido del dataset Samsung ($n=377$) limitó la capacidad de generalización de los modelos entrenados sobre esa fuente, generando alta varianza en las métricas de evaluación. Asimismo, la ausencia de datasets clínicos individuales de Argentina con variables perinatales completas impidió entrenar modelos sobre población local, siendo Brasil el proxy geográfico más cercano disponible.

El balanceo de clases mediante oversampling simple, si bien efectivo, introduce el riesgo de sobrerrepresentar patrones existentes en la clase minoritaria sin generar nueva información. Técnicas más avanzadas como SMOTE o redes generativas adversariales (GANs) podrían mejorar este aspecto en trabajos futuros.

12. 2. Trabajo futuro

Como extensión natural de este trabajo, se desarrolló **PrematurApp**, una aplicación web

accesible disponible en <https://prematuapp.streamlit.app>, donde el personal de salud puede ingresar las variables perinatales al momento del nacimiento y obtener una estimación del riesgo neonatal en tiempo real, diseñada específicamente para su uso en hospitales públicos de baja complejidad. En etapas posteriores, se propone explorar la incorporación de esquemas de aprendizaje continuo (*online learning*), de manera que a medida que la herramienta sea utilizada y se confirmen los outcomes reales de los neonatos evaluados, los nuevos casos puedan retroalimentar periódicamente el modelo, permitiendo que este se ajuste de forma autónoma a las características de la población local sin necesidad de reentrenamiento manual.

Adicionalmente, se propone explorar el uso de redes generativas adversariales (GANs) para la generación de datos sintéticos de prematuros como alternativa al oversampling tradicional, y la construcción de datasets clínicos longitudinales con variables de seguimiento motor que permitan extender el modelo hacia la predicción del desarrollo neuromotor en prematuros sobrevivientes.

Finalmente, la construcción de un dataset clínico individual propio con datos argentinos, mediante convenios con hospitales públicos y respetando la normativa de confidencialidad estadística vigente, podría permitir en el futuro el entrenamiento de modelos sobre población local, aumentando significativamente la pertinencia del trabajo para el sistema de salud argentino.

13. Recursos disponibles

Para el desarrollo del presente trabajo se utilizaron los siguientes recursos:

- Dataset SPNeoDeath - Synapse: <https://doi.org/10.7303/syn22240254>
- Dataset Centro Médico Samsung - Figshare: <https://figshare.com/s/b982cd2d40a3b277e882>
- Dataset "Development Preterm Infant" - Mendeley Data: <https://data.mendeley.com/datasets/h464gsf77t/2>
- Google Colab como entorno de desarrollo y ejecución del código.
- Librerías de Python: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn y SciPy.
- Computadora personal y conexión a internet.

14. Impacto esperado

Dentro del impacto esperado de mi trabajo, se destaca la contribución al conocimiento mediante el uso de aprendizaje automático supervisado para abordar la vulnerabilidad neonatal. Tras una investigación de los modelos existentes, se observa que iniciativas como el proyecto "Metapué" en Salta o estudios avanzados en Brasil han demostrado el éxito de la tecnología en la gestión de pacientes. Sin embargo, la mayoría de estas soluciones suelen estar ligadas a centros de alta complejidad o instituciones privadas. Mi proyecto busca democratizar este acceso, proponiendo una herramienta diseñada específicamente para ser implementada en hospitales públicos de baja complejidad, donde la falta de

infraestructura tecnológica suele ser una barrera para la detección temprana.

El impacto social que persigo trasciende lo técnico: busco que este trabajo sirva para visibilizar la salud neonatal y combatir la desatención que sufren muchos recién nacidos. En muchos contextos, la falta de herramientas precisas lleva a diagnósticos erróneos desde el embarazo o a decisiones clínicas donde se descarta prematuramente la posibilidad de supervivencia de estos "pacientitos". Al proporcionar un modelo de clasificación de riesgo basado en datos objetivos, mi intención es ofrecer un respaldo científico que garantice que cada recién nacido reciba el tratamiento y la oportunidad de vida que merece, sin importar los recursos del hospital donde nazca.

Desde el punto de vista técnico, este desarrollo me permite aplicar mi conocimiento sobre registros clínicos reales provenientes de repositorios públicos internacionales como SPNeoDeath, el Centro Médico Samsung y Mendeley Data. El uso de algoritmos supervisados permitió validar la hipótesis de investigación con rigor estadístico, transformando datos abiertos en alertas tempranas de riesgo neonatal.

La ejecución individual del proyecto va impulsada por un motivo simbólico y personal, permitiéndome además un mayor control sobre la gestión de los datos y el tiempo para garantizar una herramienta que minimice el margen de error diagnóstico en la salud pública.

15. Referencias

Apgar, V. (1953). A proposal for a new method of evaluation of the newborn infant. *Current Researches in Anesthesia & Analgesia*, 32(4), 260–267.

Beluzo, C. E., Silva, E., Alves, L. C., Bresan, R. C., Arruda, N. M., Sovat, R., & Carvalho, T. (2020). SPNeoDeath: A demographic and epidemiological data set having infant, mother, prenatal care and childbirth data related to births and neonatal deaths in São Paulo city Brazil – 2012–2018. *Data in Brief*, 32, 106093. <https://doi.org/10.1016/j.dib.2020.106093>

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc.

Dirección de Estadística e Información de Salud [DEIS]. (s.f.). *Estadísticas vitales*. Ministerio de Salud de Argentina. <https://www.argentina.gob.ar/salud/deis>

Faramarzi, R. (2023). *Development preterm infant* [Dataset]. Mendeley Data, V2. <https://data.mendeley.com/datasets/h464gsf77t/2>

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.

Hospital Público Materno Infantil de Salta. (2023). *El HPMI resultó ganador del programa "Transformar Salud" 2023*. Gobierno de Salta. <https://www.salta.gob.ar/prensa/noticias/el-hpmi-resulto-ganador-del-programa-transformar-salud-2023-91785>

Open Knowledge Foundation. (2015). *Open definition* 2.1. <https://opendefinition.org/od/2.1/en/>

Organización Mundial de la Salud. (2023). *Nacimientos prematuros*. <https://www.who.int/es/news-room/fact-sheets/detail/preterm-birth>

Park, J. H., Chang, Y. S., Sung, S., Ahn, S. Y., & Park, W. S. (2017). Trends in overall mortality, and timing and cause of death among extremely preterm infants near the limit of viability. *PLOS ONE*, 12(1), e0170220. <https://doi.org/10.1371/journal.pone.0170220>

Shu, C. H., Zebda, R., Espinosa, C., Reiss, J., Debuyserie, A., Reber, K., Aghaeepour, N., & Pammi, M. (2024). Early prediction of mortality and morbidities in VLBW preterm neonates using machine learning. *Pediatric Research*, 97(6), 2056–2064. <https://doi.org/10.1038/s41390-024-03604-7>

Silva, G. F. S., Wichmann, R. M., da Silva Junior, F. C., & Chiavegatto Filho, A. D. P. (2025). Development and evaluation of machine learning training strategies for neonatal mortality prediction using multicountry data. *Scientific Reports*, 15, 24278. <https://doi.org/10.1038/s41598-025-04066-5>

Sullivan, B. A., Moreira, A. G., McAdams, R. M., Knake, L. A., Husain, A., Qiu, J., Mudireddy, A., Majeedi, A., Shalish, W., Lake, D. E., & Vesoulis, Z. A. (2025). Comparing machine learning techniques for neonatal mortality prediction: insights from a modeling competition. *Pediatric Research*, 98(2), 405–411. <https://doi.org/10.1038/s41390-024-03773-5>

World Health Organization. (2022). *WHO recommendations on antenatal care for a positive pregnancy experience*. WHO Press.

16. Anexos

Anexo A: Reproducibilidad

A. 1. Entorno y librerías

El análisis fue desarrollado íntegramente en Google Colab. Las librerías utilizadas fueron:

- pandas y numpy para el procesamiento y manipulación de datos
- matplotlib y seaborn para la visualización
- scikit-learn para el modelado, preprocesamiento y métricas de evaluación
- scipy para las pruebas estadísticas (ANOVA)

A. 2. Acceso al código

El notebook completo está disponible en Google Colab:

https://colab.research.google.com/drive/1UblZy3tNY8ghuazlxe8SBsGDspn_ZBhx?usp=sharing

A. 3. Acceso a los datos

Los datasets utilizados están disponibles en los repositorios indicados en la sección de Recursos disponibles. El dataset SPNeoDeath requiere registro previo en Synapse. Los

datasets de Samsung y Mendeley son de descarga directa sin registro.

A. 4. Presentación de defensa

La presentación utilizada en la defensa oral está disponible en:

<https://canva.link/modeladodelriesgoneonatal>

Este documento constituye la entrega final del Seminario de Trabajo Final de la Licenciatura en Ciencia de Datos — Universidad del Gran Rosario, 2026.