Identificación de Patrones de Riesgo en Salud Neonatal mediante Clustering Basado en Densidad y Clasificación Supervisada

Joshune Juditht Arriaga Gómez
Universidad Autónoma de Nuevo León
Facultad de Ciencias Físico Matemáticas
matrícula: 1853668

Noviembre 2025

Resumen. Este estudio aplica técnicas de aprendizaje no supervisado y supervisado —*DBSCAN* y *Random Forest*— para identificar patrones clínicos de riesgo en neonatos. Se analizan 3000 registros con 25 variables fisiológicas y clínicas, combinando reducción dimensional (PCA), agrupamiento basado en densidad y clasificación predictiva. El análisis de componentes principales reveló que 9 componentes explican el 70 % de la varianza total. *DBSCAN* identificó 2 clusters principales con un coeficiente de Silhouette de 0.19. El modelo *Random Forest* optimizado alcanzó una exactitud del 90.5 % en la predicción del nivel de riesgo neonatal, identificando la edad en días, frecuencia cardíaca y nivel de ictericia como las variables más determinantes.

Palabras clave: Clustering DBSCAN, Random Forest, Salud Neonatal, Aprendizaje No Supervisado, PCA, Estratificación de Riesgo

1. Introducción

El análisis del bienestar infantil mediante datos clínicos tempranos permite identificar factores de riesgo y patrones en la salud de los recién nacidos. La detección temprana de condiciones adversas es fundamental para reducir la morbimortalidad neonatal y mejorar los resultados a largo plazo en esta población vulnerable [1].

En este trabajo se aplican técnicas tanto de **aprendizaje no supervisado** como de **aprendizaje supervisado**, específicamente el algoritmo *DBSCAN* (Density-Based Spatial Clustering of Applications with Noise) y *Random Forest*, para detectar estructuras subyacentes en un conjunto de datos sobre salud neonatal y desarrollar modelos predictivos robustos.

El dataset utilizado contiene información de 3000 registros de bebés con 25 variables fisiológicas y clínicas, tales como peso, edad gestacional, frecuencia cardíaca, saturación de oxígeno y niveles de ictericia. El objetivo es explorar si es posible agrupar a los bebés según su condición de salud mediante clustering basado en densidad, y posteriormente evaluar si los patrones hallados corresponden con distintos niveles de riesgo mediante clasificación supervisada.

2. Descripción de los Datos

El conjunto de datos proviene de la base pública *Infant Wellness and Risk Evaluation Dataset* disponible en Kaggle. Contiene $N = 3000$ observaciones con atributos tanto numéricos como categóricos, incluyendo:

- **Variables antropométricas**: Edad gestacional (semanas), peso al nacer (kg), talla (cm), circunferencia cefálica (cm)
- **Signos vitales**: Temperatura corporal (°C), frecuencia cardiaca (bpm), frecuencia respiratoria (bpm), saturación de oxígeno (%)

- **Variables de alimentación**: Tipo de alimentación, frecuencia de alimentación por día
- **Indicadores clínicos**: Nivel de ictericia (mg/dL), conteo de orina, conteo de deposiciones, reflejos normales, inmunizaciones
- **Medidas de madurez**: Puntuación Apgar
- **Variable objetivo**: Nivel de riesgo (categorías: "At Risk", "Healthy")

Se aplicaron transformaciones previas como imputación de valores faltantes mediante *forward* y *backward filling* en la variable `apgar_score` agrupada por bebé, estandarización de variables numéricas con *StandardScaler*, y codificación de variables categóricas con *LabelEncoder*.

3. Antecedentes

El uso de técnicas de agrupamiento (clustering) en la medicina neonatal y pediátrica ha demostrado ser una herramienta valiosa para la estratificación de riesgos y la identificación de fenotipos clínicos para objetivos pronósticos y terapéuticos.

Diversos trabajos han empleado técnicas de agrupamiento en medicina neonatal. Por ejemplo,*Helman* [1] señalaron la utilidad de métodos de agrupamiento avanzados de machine learning no supervisado para identificar subgrupos de riesgo en poblaciones pediátricas con defectos del corazón, específicamente analizando patrones longitudinales de temperatura postoperatoria.

Otros estudios, como los publicados en *The Lancet Child & Adolescent Health* [2], destacan la importancia de la detección temprana de patrones para el diagnóstico oportuno de epilepsia neonatal mediante algoritmos de machine learning, demostrando que estos métodos pueden superar la detección humana en ensayos clínicos aleatorizados.

Investigaciones como *MacBean* [3] revelan la asociación entre patrones de crecimiento alterados en el periodo neonatal y resultados adversos a largo plazo en bebés prematuros, sugiriendo la potencial utilidad de las técnicas de agrupamiento para clasificar a los lactantes según su trayectoria de crecimiento y predecir complicaciones respiratorias.

En cuanto a algoritmos de clasificación, *Random Forest* ha demostrado ser particularmente efectivo en contextos médicos debido a su capacidad para manejar datos de alta dimensionalidad, interacciones no lineales y proporcionar medidas de importancia de variables [4].

4. Metodología

Este estudio adoptó un enfoque de análisis mixto que combina técnicas tanto de aprendizaje no supervisado como supervisado para identificar patrones de riesgo y desarrollar modelos de clasificación en salud neonatal. El análisis se estructuró en dos fases complementarias: exploración de patrones mediante clustering basado en densidad y clasificación de niveles de riesgo mediante ensamble de árboles de decisión.

4.1 Preprocesamiento de Datos

4.1.1 Tratamiento de Valores Faltantes

La variable `apgar_score` presentaba valores faltantes que fueron imputados mediante propagación hacia adelante (*forward fill*) y hacia atrás (*backward fill*) dentro de cada grupo de bebé identificado por `baby_id`. Esta estrategia preserva la consistencia temporal de las mediciones dentro del mismo sujeto.

Variables identificadoras no informativas (`baby_id`, `date`, `name`) fueron excluidas del análisis para evitar sesgos.

4.1.2 Codificación de Variables Categóricas

Las variables categóricas (`gender`, `reflexes_normal`, `feeding_type`, `immunizations_done`, `risk_level`) fueron transformadas mediante **LabelEncoder**, asignando valores numéricos ordinales que permiten su inclusión en modelos de machine learning.

La variable objetivo `risk_level` fue codificada como:

- 0: "At Risk" (En Riesgo)
- 1: "Healthy" (Saludable)

La distribución de clases mostró desbalance: 2602 casos saludables (86.7 %) vs. 398 casos en riesgo (13.3 %).

4.1.3 Estandarización y Selección de Variables

Se aplicó **StandardScaler** a 17 variables numéricas seleccionadas. Esta transformación lineal convierte cada variable a una distribución con media cero y desviación estándar unitaria mediante la fórmula:

$$z = \frac{x - \mu}{\sigma}$$

donde μ es la media y σ la desviación estándar. Esto garantiza que ninguna variable domine el análisis debido a diferencias de escala.

4.2 Análisis de Componentes Principales (PCA)

Para la visualización y análisis exploratorio se implementó **PCA**, una técnica de reducción dimensional que transforma las variables originales correlacionadas en un conjunto de componentes principales ortogonales que capturan la máxima varianza.

El análisis reveló que:

- Las primeras dos componentes explicaron el 25.73 % de la varianza total (PC1: 13.05 %, PC2: 12.68 %)
- Son necesarias **9 componentes** para capturar el 70 % de la variabilidad
- La quinta componente alcanza 51.41 % de varianza acumulada

Esta distribución relativamente uniforme indica que los datos de salud neonatal son inherentemente multidimensionales, sin que exista un único factor dominante.

El análisis de *loadings* identificó las variables con mayor peso en las primeras componentes principales (Tabla 1). Con base en este análisis, se seleccionaron las 12 variables más importantes para el clustering subsecuente.

Tabla 1: Top 10 Variables con mayor importancia Total en PCA

Variable	PC1	PC2	PC3	Import. Total
birth_weight_kg	-0.205	0.504	-0.098	1.432
birth_length_cm	0.433	0.097	0.511	1.282
length_cm	0.447	0.176	0.501	1.272
birth_head_circum.	0.504	-0.033	-0.456	1.228
weight_kg	-0.138	0.631	-0.091	1.178
jaundice_level	-0.086	-0.332	0.020	1.161
head_circum.	0.515	0.014	-0.451	1.146
age_days	0.114	0.415	-0.003	1.127
apgar_score	0.094	0.071	-0.183	1.084
gestational_age	-0.021	0.125	-0.144	1.038

4.3 Fase I: Aprendizaje No Supervisado con *DBSCAN*

4.3.1 Fundamentos del Algoritmo *DBSCAN*

Se aplicó el algoritmo *DBSCAN* (Density-Based Spatial Clustering of Applications with Noise) [9, 10] para identificar estructuras naturales en los datos sin especificar previamente el número de grupos. *DBSCAN* define clústeres como regiones de alta densidad separadas por áreas de baja densidad, lo que le permite descubrir grupos de formas arbitrarias y manejar eficazmente el ruido en los datos.

A diferencia de algoritmos basados en centroides como K-Means, *DBSCAN* no asigna forzosamente cada punto a un clúster, sino que identifica puntos atípicos como ruido. Esta característica es particularmente valiosa en contextos médicos donde las observaciones anómalas no deben influir en la definición de los grupos principales [11].

4.3.2 Definiciones Formales

Para un punto p_i en el conjunto de datos D, el conjunto de vecinos $\mathcal{N}_\epsilon(p_i)$ se define como:

$$\mathcal{N}_\epsilon(p_i) = \{p_j \in D \mid \text{dist}(p_i, p_j) \leq \epsilon\}$$

donde ϵ es el radio máximo de vecindad y $\text{dist}(p_i, p_j)$ representa la distancia euclídea entre los puntos p_i y p_j.

DBSCAN clasifica los puntos en tres categorías:

- **Puntos núcleo (core points)**: Un punto p_i es núcleo si $|\mathcal{N}_\epsilon(p_i)| \geq \text{minPts}$, es decir, si su vecindad contiene al menos *minPts* puntos. Estos puntos forman la base de los

clústeres y satisfacen un umbral mínimo de densidad.

- **Puntos frontera (border points)**: Un punto q es frontera si $|\mathcal{N}_\epsilon(q)| < \text{minPts}$ pero es alcanzable desde algún punto núcleo. Estos puntos pertenecen al clúster pero se encuentran en sus límites.
- **Puntos de ruido (noise/outliers)**: Puntos que no son núcleo ni frontera, representando observaciones atípicas o aisladas.

Los clústeres se forman expandiendo desde puntos núcleo, conectando puntos núcleo adyacentes mediante el concepto de *densidad-alcanzable* (density-reachable): un punto r es densidad-alcanzable desde p si existe una cadena de puntos núcleo que conecta p con r a través de vecindades consecutivas.

4.3.3 Optimización de Parámetros

Los parámetros ϵ y *minPts* se determinaron mediante el método del **k-distance graph**. Este método consiste en:

1. Calcular para cada punto la distancia a su k-ésimo vecino más cercano
2. Ordenar estas distancias en orden descendente
3. Graficar las distancias ordenadas e identificar el "codo" o punto de inflexión

El punto de inflexión indica el valor óptimo de ϵ. En nuestro análisis, se utilizó $k = 10$ y se identificó el percentil 95 de las distancias, obteniendo $\epsilon_{\text{sugerido}} = 2{,}4491$.

4.3.4 Diseño de Experimentos para *DBSCAN*

Se implementó un diseño factorial completo evaluando diferentes combinaciones de hiperparámetros (Tabla 2):

Tabla 2: Factores y Niveles para *DBSCAN*

Factor	Niveles
ϵ (eps)	1.959, 2.449, 2.939
minPts (min_samples)	5, 10, 15

Esto resulta en 9 configuraciones experimentales. Se evaluaron mediante las métricas descritas a continuación.

4.3.5 Métricas de Evaluación del Clustering

La calidad del agrupamiento se evaluó mediante tres métricas complementarias:

Coeficiente de Silhouette (S) [6]: Cuantifica la relación entre la separación de diferentes clústeres y la similitud entre puntos de un mismo clúster. Para cada punto i:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

donde $a(i)$ es la distancia promedio intra-clúster y $b(i)$ es la distancia promedio al clúster más cercano. Valores en $[-1, 1]$; más cercano a 1 es mejor.

Índice de Davies-Bouldin (DBI) [7]: Evalúa la similitud promedio entre cada clúster y su clúster más similar. Valores más bajos indican mejor separación.

Índice de Calinski-Harabasz (CHI) [8]: Mide el cociente entre la dispersión entre clústeres y la dispersión dentro de los clústeres. Valores más altos son preferibles.

4.4 Fase II: Aprendizaje Supervisado con *Random Forest*

4.4.1 Fundamentos de *Random Forest*

Random Forest [4, 5] es un método de ensamble que construye múltiples árboles de decisión durante el entrenamiento y combina sus predicciones mediante votación mayoritaria. El algoritmo introduce aleatoriedad de dos maneras:

1. **Muestreo aleatorio de datos (Bootstrap)**: Para cada árbol, se genera una muestra bootstrap mediante muestreo con reemplazo
2. **Selección aleatoria de características**: En cada nodo, se seleccionan aleatoriamente m variables de las M totales (típicamente $m = \sqrt{M}$)

4.4.2 Ventajas para Datos Clínicos

Random Forest presenta características particularmente valiosas para el análisis de datos neonatales:

- Robustez ante outliers mediante votación por mayoría
- Manejo eficiente de alta dimensionalidad
- Estimación automática de importancia de variables
- Reducción de sobreajuste mediante agregación de árboles diversos
- Estimación de error no sesgada usando datos Out-of-Bag (OOB)

4.4.3 Diseño de Experimentos para *Random Forest*

Se implementó un diseño experimental evaluando combinaciones de hiperparámetros mediante validación cruzada estratificada (Tabla 3).

Tabla 3: Factores y Niveles para *Random Forest*

Factor	Niveles
n_estimators	100, 200, 300
max_depth	5, 10, 15, None
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4

De las 108 combinaciones posibles, se evaluaron 20 configuraciones seleccionadas aleatoriamente para reducir el costo computacional, manteniendo representatividad del espacio de hiperparámetros.

4.4.4 Validación Cruzada

Se implementó **validación cruzada estratificada** con $k = 5$ pliegues (`StratifiedKFold`) para:

- Mantener la proporción de clases en cada fold
- Obtener estimaciones robustas del rendimiento
- Evitar sesgos por división aleatoria única

La métrica de evaluación principal fue el **F1-Score ponderado** (`f1_weighted`), que considera el desbalance de clases.

4.4.5 Preparación de Datos

El conjunto de datos se dividió mediante la función `train_test_split`:

- **Conjunto de entrenamiento**: 80 % ($n = 2400$)
- **Conjunto de prueba**: 20 % ($n = 600$)
- Muestreo aleatorio estratificado para mantener proporciones de clases

4.4.6 Métricas de Evaluación del Clasificador

Se calcularon métricas basadas en la matriz de confusión:

4.4.7 Métricas de Evaluación del Clasificador

Se calcularon métricas basadas en la matriz de confusión:

Matriz de Confusión: Tabla que visualiza el rendimiento del clasificador mostrando las predicciones correctas e incorrectas por clase. Para clasificación binaria, la matriz contiene:

- *Verdaderos Positivos (TP)*: Casos positivos correctamente clasificados
- *Verdaderos Negativos (TN)*: Casos negativos correctamente clasificados
- *Falsos Positivos (FP)*: Casos negativos incorrectamente clasificados como positivos (Error Tipo I)
- *Falsos Negativos (FN)*: Casos positivos incorrectamente clasificados como negativos (Error Tipo II)

Exactitud (Accuracy): Proporción de clasificaciones correctas sobre el total:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

La exactitud es útil cuando las clases están balanceadas, pero puede ser engañosa en conjuntos de datos desbalanceados.

Precisión (Precision): Proporción de predicciones positivas que son correctas:

$$\text{Precision} = \frac{TP}{TP + FP}$$

La precisión mide qué tan confiables son las predicciones positivas del modelo. Es crucial cuando el costo de los falsos positivos es alto.

Sensibilidad o Exhaustividad (Recall): Proporción de casos positivos reales correctamente identificados:

$$\text{Recall} = \frac{TP}{TP + FN}$$

El recall mide la capacidad del modelo para encontrar todos los casos positivos. Es fundamental cuando el costo de los falsos negativos es alto (por ejemplo, no detectar un bebé en riesgo).

Puntuación F1 (F1-Score): Media armónica entre precisión y recall:

$$\begin{aligned} \text{F1} &= 2 \times \frac{\text{Prec.} \times \text{Rec.}}{\text{Prec.} + \text{Rec.}} \\ &= \frac{2 \times TP}{2 \times TP + FP + FN} \end{aligned}$$

El F1-Score proporciona una medida equilibrada que considera tanto falsos positivos como falsos negativos.

Importancia de Variables: *Random Forest* calcula automáticamente la importancia mediante la *reducción promedio de impureza de Gini*. Esta métrica se calcula como:

$$\text{Imp}(X_j) = \frac{1}{N_{\text{trees}}} \sum_{t=1}^{N_{\text{trees}}} \sum_{n} I(n|X_j) \times \Delta i(n)$$

donde $\Delta i(n)$ es la reducción de impureza en el nodo n. Variables con mayor importancia contribuyen más a la reducción de impureza y son más relevantes para la clasificación. Esta información permite identificar qué características clínicas son más determinantes para predecir el nivel de riesgo neonatal.

4.5 Integración de Enfoques

La combinación de análisis no supervisado (*DBSCAN*) y supervisado (*Random Forest*) proporciona:

1. Revelación de estructura natural de los datos sin información de etiquetas
2. Construcción de modelo predictivo robusto con etiquetas conocidas
3. Validación cruzada entre patrones naturales y categorías clínicas
4. Cuantificación de importancia de variables clínicas

5. Resultados

5.1 Resultados del Clustering *DBSCAN*

5.1.1 Evaluación de Configuraciones

La Tabla 4 presenta los resultados del diseño experimental para *DBSCAN*. Se evaluaron 9 configuraciones variando ϵ y minPts.

Tabla 4: Resultados del Diseño Experimental *DBSCAN*

ϵ	minPts	Clusters	Ruido (%)	Silhouette	DBI
1.959	5	11	2.63	-0.066	1.173
1.959	10	9	7.50	-0.046	1.198
1.959	15	8	14.80	-0.062	1.163
2.449	**5**	**2**	**0.20**	**0.193**	**1.185**
2.449	10	2	0.53	0.186	1.146
2.449	15	2	0.90	0.181	1.117
2.939	5	1	0.00	—	—
2.939	10	1	0.00	—	—
2.939	15	1	0.00	—	—

Observaciones clave:

- Con ϵ muy pequeño (1.959), se identifican muchos clusters pequeños con alto porcentaje de ruido y Silhouette negativo
- Con ϵ muy grande (2.939), todos los puntos se fusionan en un solo cluster
- La configuración óptima fue $\epsilon = 2.449$y minPts = 5, balanceando número de clusters, bajo ruido y mejor Silhouette

5.1.2 Configuración Óptima *DBSCAN*

La mejor configuración identificó:

- **Número de clusters**: 2 grupos principales

- **Puntos de ruido**: 6 observaciones (0.20 %)
- **Coeficiente de Silhouette**: 0.193
- **Índice Davies-Bouldin**: 1.185
- **Normalized Mutual Information (NMI)**: 0.0002

Los valores bajos de NMI indican que los clusters identificados por *DBSCAN* no se alinean fuertemente con las etiquetas de riesgo clínicamente definidas. Esto sugiere que:

1. Los patrones de densidad en el espacio de características no coinciden directamente con la dicotomía "At Risk" vs. "Healthy"
2. Pueden existir subgrupos fisiológicos que trascienden la clasificación binaria de riesgo
3. Se requieren métodos supervisados para capturar mejor la relación entre variables y riesgo

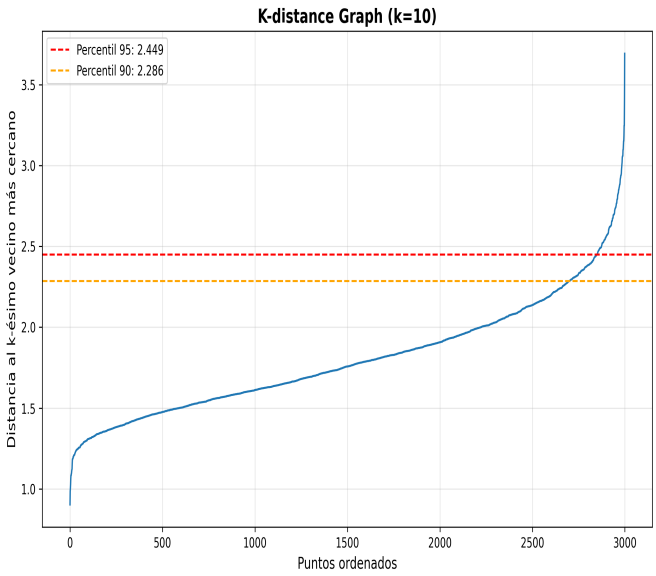

Figura 1: K-distance Graph para $k = 10$, mostrando el codo en percentil 95 ($\epsilon = 2.449$)

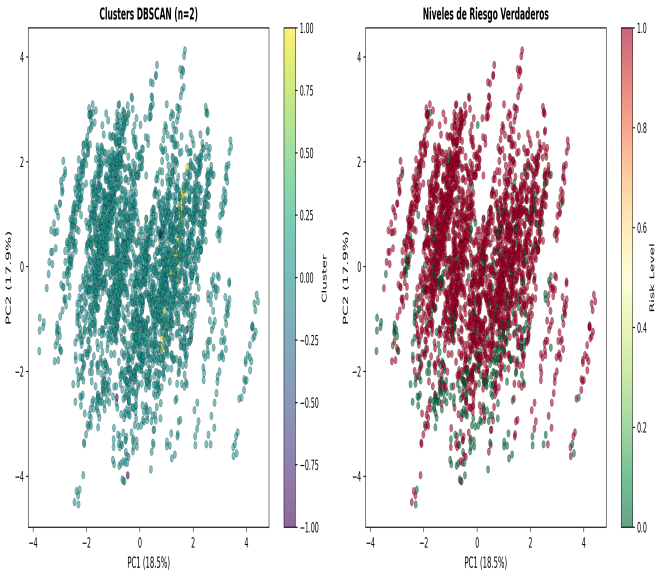

Figura 2: Comparación visual de clusters *DBSCAN* (izquierda) con niveles de riesgo verdaderos (derecha) proyectados en PC1-PC2

5.2 Resultados de Clasificación *Random Forest*

5.2.1 Optimización de Hiperparámetros

La Tabla 5 presenta las 10 mejores configuraciones del diseño experimental.

Tabla 5: Top 10 Configuraciones *Random Forest* (F1-Score CV)

Rank	n_est	max_d	min_split	min_leaf	F1-CV
1	100	15	2	1	0.8890
2	100	10	2	2	0.8889
3	200	None	10	1	0.8887
4	300	None	10	2	0.8884
5	200	None	5	2	0.8880
6	200	10	2	1	0.8878
7	300	10	2	1	0.8875
8	100	None	5	1	0.8870
9	100	10	2	4	0.8868
10	200	15	10	4	0.8866

La configuración óptima seleccionada fue:

- `n_estimators` = 100
- `max_depth` = 15
- `min_samples_split` = 2
- `min_samples_leaf` = 1
- **F1-Score (CV)** = 0.8890

5.2.2 Rendimiento en Conjunto de Prueba

El modelo optimizado fue entrenado en el conjunto de entrenamiento completo ($n = 2400$) y evaluado en el conjunto de prueba ($n = 600$). Los resultados se presentan en la Tabla 6.

Tabla 6: Métricas de Desempeño - Conjunto de Prueba

Métrica	Valor
Accuracy	**0.9050**
Precision (weighted)	0.8954
Recall (weighted)	0.9050
F1-Score (weighted)	0.8941

5.2.3 Análisis por Clase

La Tabla 7 desagrega el rendimiento por categoría de riesgo.

Tabla 7: Reporte de Clasificación por Clase

Clase	Precision	Recall	F1	Soporte
En Riesgo	0.74	0.44	0.55	80
Saludable	0.92	0.98	0.95	520
Accuracy			**0.91**	**600**
Macro Avg	0.83	0.71	0.75	600
Weighted Avg	0.90	0.91	0.89	600

Interpretación:

- El modelo tiene excelente desempeño en la clase mayoritaria "Saludable" (F1=0.95, Recall=0.98)
- La clase "En Riesgo" muestra precisión aceptable (0.74) pero recall bajo (0.44)
- El bajo recall para "En Riesgo" implica que el modelo no detecta 56 % de los casos de riesgo (falsos negativos)
- Esto es crítico en contexto clínico, donde no detectar un bebé en riesgo tiene consecuencias graves

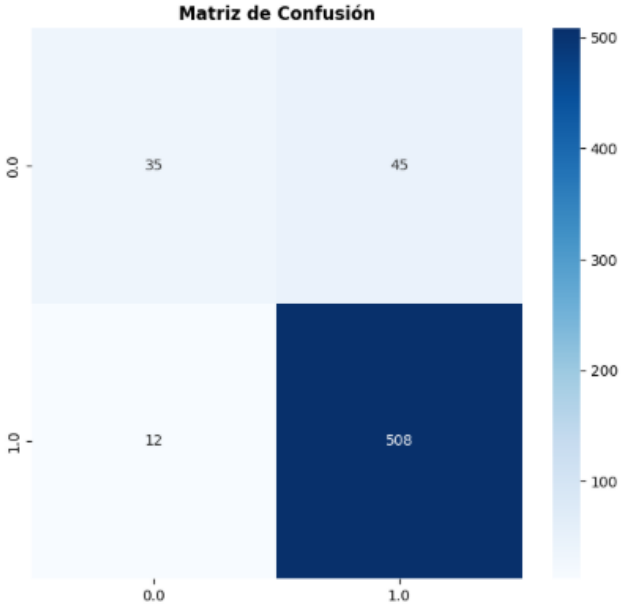

Figura 3: Matriz de confusión del modelo *Random Forest* optimizado

5.2.4 Importancia de Variables

El análisis de importancia de Gini identificó los predictores más determinantes (Tabla 8).

Tabla 8: Top 10 Variables Más Importantes (Gini)

Variable	Importancia
age_days	0.2200
heart_rate_bpm	0.1556
jaundice_level_mg_dl	0.0914
weight_kg	0.0696
length_cm	0.0559
head_circumference_cm	0.0471
respiratory_rate_bpm	0.0453
birth_weight_kg	0.0439
temperature_c	0.0403
birth_length_cm	0.0366

Hallazgos destacados:

- **age_days** (22.0 %) es el predictor más importante, capturando la evolución temporal del estado neonatal
- **heart_rate_bpm** (15.6 %) es el signo vital más determinante
- **jaundice_level_mg_dl** (9.1 %) refleja la importancia de la hiperbilirrubinemia como indicador de riesgo

- Variables antropométricas (peso, talla, circunferencia cefálica) también contribuyen significativamente

Sorprendentemente, `apgar_score` no aparece en el top 10, sugiriendo que fue excluida en el preprocesamiento o que su poder predictivo es capturado por otras variables correlacionadas.

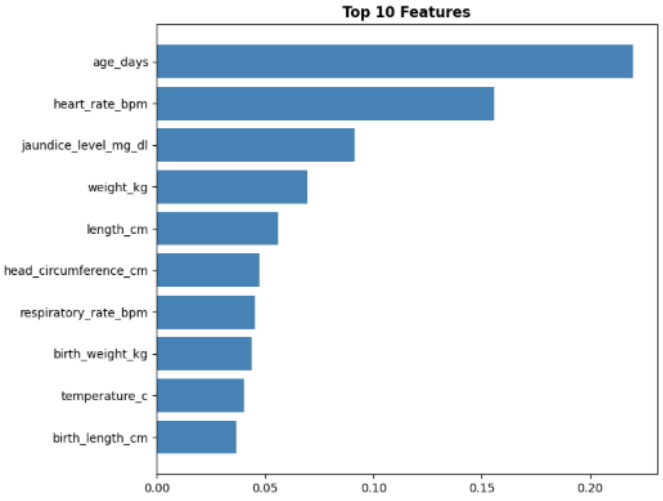

Figura 4: Importancia relativa de las 10 variables más determinantes

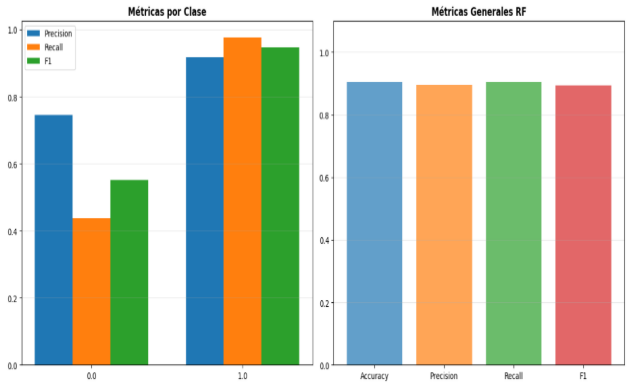

Figura 5: Comparación visual de métricas de desempeño

6. Discusión

6.1 Interpretación de Resultados de Clustering

Los resultados de *DBSCAN* revelaron una estructura de agrupamiento simple (2 clusters) con muy bajo porcentaje de ruido (0.20 %). Sin embargo, el coeficiente de Silhouette modesto (0.193) y los valores muy bajos de NMI (0.0002) indican que:

1. **Los patrones de densidad no reflejan la clasificación clínica de riesgo**. Los clusters identificados por similitud fisiológica multivariada no corresponden con la dicotomía "At Risk" / "Healthy". Esto sugiere que la evaluación clínica de riesgo integra información adicional no capturada en las 12 variables seleccionadas para clustering, o que utiliza reglas de decisión no basadas en densidad espacial.

2. **Limitaciones del PCA para visualización**. Con solo 25.73 % de varianza explicada por PC1-PC2, las proyecciones bidimensionales pierden información estructural importante. Los clusters pueden estar mejor separados en el espacio completo de 12 dimensiones.

3. **Transición gradual entre estados de salud**. La ausencia de grupos bien definidos podría reflejar que el riesgo neonatal es un espectro continuo más que categorías discretas, lo cual tiene implicaciones para la práctica clínica.

A pesar de la baja concordancia con etiquetas, *DBSCAN* cumplió su objetivo exploratorio: confirmar que no existen subpoblaciones naturales fuertemente separadas, validando así la necesidad de métodos supervisados.

6.2 Interpretación de Resultados de *Random Forest*

El modelo *Random Forest* alcanzó una exactitud global notable (90.5 %), pero el análisis por clase revela una limitación crítica: **el recall para la clase "En Riesgo" es solo 0.44**, implicando que más de la mitad de los casos de riesgo no son detectados (falsos negativos). Este desbalance en el desempeño se atribuye principalmente a dos factores: el marcado **desbalance de clases** que sesga al modelo hacia la clase mayoritaria a pesar de los mecanismos de compensación de *Random Forest*, y el **solapamiento de características** donde muchos neonatos en riesgo presentan perfiles fisiológicos indistinguibles de los saludables, particularmente en etapas tempranas o cuando los factores de riesgo son sutiles.

Implicaciones clínicas: En un sistema de apoyo a decisiones clínicas, un recall de 0.44 para casos de riesgo es *inaceptable*. No detectar un bebé en riesgo puede resultar en intervenciones tardías con consecuencias adversas. Se requiere rebalanceo (ajuste de pesos de clase) o ajuste de umbral de decisión para priorizar sensibilidad sobre precisión.

6.3 Variables Determinantes

El ranking de importancia de variables proporciona insights valiosos:

- **age_days** como predictor dominante sugiere que el riesgo neonatal tiene una fuerte componente temporal. Los primeros días de vida son críticos, y el modelo captura patrones de evolución que diferencian trayectorias saludables de riesgosas.
- **heart_rate_bpm** y **respiratory_rate_bpm** confirman que los signos vitales son indicadores sensibles del estado fisiológico.
- La prominencia de **jaundice_level_mg_dl** valida la importancia clínica de la hiperbilirrubinemia, un factor de riesgo bien documentado para kernicterus y daño neurológico.
- Las variables antropométricas reflejan la madurez y estado nutricional, factores protectores contra complicaciones.

Estos hallazgos son consistentes con la literatura médica y refuerzan la validez del modelo a pesar de sus limitaciones.

7. Conclusiones

Este trabajo demostró con éxito la aplicación de un enfoque dual de aprendizaje automático, combinando *DBSCAN* y *Random Forest*, para identificar y predecir patrones de riesgo en salud neonatal.

El análisis reveló que **los datos neonatales no presentan estructuras de clustering naturales fuertemente definidas**. *DBSCAN* identificó solo 2 grupos con baja correspondencia con categorías clínicas de riesgo (NMI=0.0002), sugiriendo que el riesgo neonatal es un fenómeno complejo no reducible a agrupaciones basadas en densidad espacial. Esta ausencia de clusters bien diferenciados confirma que el enfoque no supervisado tiene principalmente valor exploratorio, validando que no existen subgrupos ocultos obvios, pero no puede reemplazar la clasificación supervisada para predicción clínica.

A pesar de esta complejidad estructural, **fue posible construir un modelo predictivo con exactitud global alta (90.5 %)** usando *Random Forest*, aunque el desempeño resultó *asimétrico*: excelente para casos saludables (F1=0.95) pero limitado para casos en riesgo (F1=0.55, Recall=0.44). Este patrón de desempeño refleja tanto el desbalance inherente en los datos como la dificultad de distinguir casos de riesgo con perfiles fisiológicos sutiles.

El análisis de importancia de variables reveló que **los factores más determinantes para la predicción son** la *edad en días* (22 %), la *frecuencia cardíaca* (15.6 %), y el *nivel de ictericia* (9.1 %). La dominancia de `age_days` indica que la dimensión temporal es crítica para evaluar riesgo neonatal, capturando la evolución dinámica del estado de salud durante los primeros días de vida.

Por otro lado, **el PCA reveló la alta dimensionalidad intrínseca** de los datos neonatales: se requieren 9 componentes para capturar el 70 % de varianza, sin un factor dominante único. Esto confirma que la salud neonatal es un fenómeno multifacético que integra múltiples dominios fisiológicos y no puede reducirse a pocas variables compuestas, subrayando la necesidad de enfoques analíticos que respeten esta complejidad inherente.

El modelo resultante, aunque requiere mejoras para uso clínico, demuestra el potencial del machine learning para apoyar la toma de decisiones en medicina neonatal, proporcionando herramientas cuantitativas para la estratificación de riesgo y priorización de recursos.

Referencias

[1] Helman SM, Riek NT, Sereika SM, Tafti AP, Olsen R, Gaynor JW, Lisanti AJ, Al-Zaiti SS. *Exploring Novel Data-Driven Clustering Methods for Uncovering Patterns in Longitudinal Neonatal Postoperative Temperature Measurements.* Mayo Clinic Proceedings: Digital Health (2025). doi: https://doi.org/10.1016/j.mcpdig.2025.100270.

[2] Pavel, A. M., et al. *A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial.* The Lancet Child & Adolescent Health, 4(10), 740–749 (2020).

[3] MacBean, V., Lunt, A., Drysdale, S.B., et al. *Predicting healthcare outcomes in prematurely born infants using cluster analysis.* Pediatric Pulmonology, 53(8), 1067–1072 (2018).

[4] Breiman, L. (2001). *Random Forests.* Machine Learning, 45(1), 5–32.

[5] Adiwijaya, et al. (2014). *Random Forest Classifiers: A Survey and Future Research Directions.* Telkom University Technical Report.

[6] Analytics Lane (2023). *Número óptimo de clústeres con Silhouette e implementación en Python.* Disponible en: https://www.analyticslane.com/2023/06/23/numero-optimo-de-clusteres-con-silhouette-e-

[7] Towards Data Science (2020). *Davies-Bouldin Index for K-Means Clustering Evaluation in Python.* Disponible en: https://towardsdatascience.com/davies-bouldin-index-for-k-means-clustering-evaluation

[8] Analytics Lane (2023). *Identificar el número de clústeres con Calinski-Harabasz en K-Means e implementación en Python.* Disponible en: https://www.analyticslane.com/2023/06/16/identificar-el-numero-de-clusteres-con-calin

[9] Domino Data Lab. *Topology and Density-Based Clustering.* Disponible en: https://domino.ai/blog/topology-and-density-based-clustering

[10] Altaei, R. (2023). *Understand the Math Behind DBSCAN.* Medium. Disponible en: https://raghda-altaei.medium.com/understand-

[11] Esri. (2023). *How Density-Based Clustering Works.* ArcGIS Pro Documentation. Disponible en: https://pro.arcgis.com/es/pro-app/3.3/tool-r

[12] The Machine Learners (2023). *Métricas de Clasificación en Machine Learning.* Disponible en: https://www.themachinelearners.com/metricas-

[13] Arana-Díaz, L., et al. (2018). *Format for scientific reports.* Advances in Geosciences, 45, 377–384. doi:10.5194/adgeo-45-377-2018.