Identificación de Patrones de Riesgo en Salud Neonatal mediante Clustering Basado en Densidad

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

Octubre 2025

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

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 poder explorar si es posible agrupar a los bebés según su condición de salud y evaluar si los patrones hallados corresponden con distintos niveles de riesgo.

2 Descripción de los datos

El conjunto de datos proviene de la base publica *Infant Wellness and Risk Evaluation Dataset* disponible en Kaggle. Contiene atributos tanto numéricos como categóricos, incluyendo:

- Edad gestacional (semanas)

- Peso y talla al nacer
- Temperatura corporal
- Frecuencia cardiaca y respiratoria
- Saturación de oxígeno
- Tipo de alimentación y frecuencia
- Nivel de ictericia (mg/dL)
- Puntuación Apgar (Puntuación de salud del recién nacido según el peso al nacer)
- Nivel de riesgo (bajo, medio, alto)

Se aplicaron transformaciones previas como relleno de valores faltantes mediante *forward* y *backward filling*, 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 de 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 supervisados para identificar subgrupos de riesgo en poblaciones pediátricas con defectos del corazón. Otros estudios, como los publicados en *The Lancet Child & Adolescent Health* [2], destacan la importancia de la detección temprana de patrones para el diágnostico oprtuno de epilepsia. Investigaciones como MacBean et al. [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.

4 Metodología

4.1 Algortimo No Supervisado y métricas

Se aplicó el algoritmo **DBSCAN**, que define clústeres como regiones de alta densidad separadas por áreas de baja densidad. Formalmente, para un punto p_i, el conjunto de vecinos $\mathcal{N}_\epsilon(p_i)$ se define como:

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

donde ϵ es el radio máximo de vecindad. Si $|\mathcal{N}_\epsilon(p_i)| \geq minPts$, el punto se considera un *núcleo*. Los clústeres se expanden conectando puntos núcleo adyacentes, mientras que los puntos que no pertenecen a ningún clúster se clasifican como *ruido*.

Los parámetros ϵ y *minPts* se determinaron mediante el método del *k-dist graph* usando *Nearest Neighbors*, observando el punto de inflexión en la curva de distancias promedio.

Posteriormente se evaluó la calidad del agrupamiento con tres métricas:

- **Coeficiente de Silhouette** (S): mide qué tan separados están los grupos.
- **Índice de Davies-Bouldin** (DBI): valores más bajos indican clústeres bien definidos.
- **Índice de Calinski-Harabasz** (CHI): valores altos reflejan mayor dispersión intergrupo.

Para visualización, se utilizó *PCA* (Análisis de Componentes Principales) reduciendo la dimensionalidad a dos componentes principales.

4.2 Preprocesamiento y selección de variables

Dado el número de variables y la necesidad de visualización, se aplicó **Análisis de Componentes Principales (PCA)**. Este método transforma las variables originales en un nuevo conjunto de variables no correlacionadas (componentes principales) que capturan la mayor varianza posible de los datos.

El análisis de varianza explicada reveló que:

- Las primeras 2 componentes explican aproximadamente el 26% de la varianza total (PC1: 13.05%, PC2: 12.68%)

- Se requieren 12 componentes para explicar el 90% de la varianza total

Del conjunto inicial de 25 variables, se seleccionaron solo 12 variables numericas relevantes para el análisis:

- Variables perinatales: edad gestacional, peso al nacer, talla al nacer, circunferencia cefálica al nacer
- Variables de seguimiento: edad en días, peso actual, talla actual, circunferencia cefálica actual
- Variables fisiológicas: frecuencia cardíaca, producción de orina
- Indicadores clínicos: nivel de ictericia, puntuación Apgar

Todas las variables fueron estandarizadas mediante *StandardScaler* para asegurar que ninguna variable dominara el análisis debido a su escala.

5 Resultados

5.1 Análisis de componentes principales

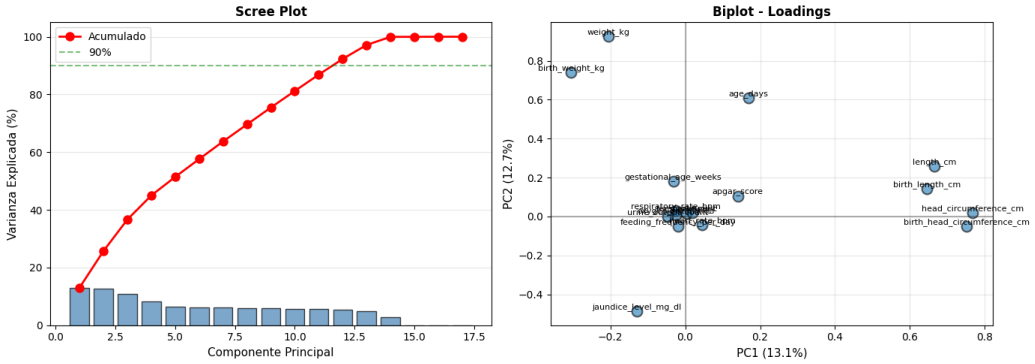

Figure 1: Análisis de componentes principales. Izquierda: *Scree plot* mostrando la varianza explicada por cada componente. Derecha: *Biplot* de loadings de las variables en PC1 y PC2.

La Figura 1 muestra el *scree plot* y el *biplot* de loadings. El *scree plot* revela que la primera componente principal (PC1) explica el 13.05% de la varianza,

mientras que PC2 explica el 12.68%. La varianza explicada decrece gradualmente, requiriendo 12 componentes para alcanzar el 90% de la varianza acumulada. Esta distribución relativamente uniforme de la varianza entre múltiples componentes indica que los datos de salud neonatal son inherentemente multidimensionales, sin que exista un único factor dominante.

En el *biplot* de loadings se observa que:

- Las variables como peso al nacer (`birth_weight_kg`), talla al nacer (`birth_length_cm`) y circunferencia cefálica al nacer (`birth_head_circumference_cm`) se agrupan en el cuadrante superior derecho, indicando alta correlación positiva con ambas componentes principales
- La edad gestacional (`gestational_age_weeks`) muestra una contribución importante hacia PC2
- El nivel de ictericia (`jaundice_level_mg_dl`) aparece con carga negativa en PC2, sugiriendo una relación inversa con la edad gestacional

5.2 Agrupamiento con DBSCAN

El modelo DBSCAN logró detectar grupos con distinta densidad en los datos. Con los parámetros optimizados ($\epsilon = 2.4491$ y $minPts = 10$), se identificaron dos clústeres principales y un pequeño conjunto de puntos clasificados como ruido, aproximadamente 0.53% de los datos.

Las métricas obtenidas fueron:

Table 1: Evaluación del modelo DBSCAN.

Métrica	**Valor**	**Interpretación**
Silhouette Score	0.1860	Separación moderada entre clústeres
Davies-Bouldin Index	1.1460	Grupos definidos razonablemente
Calinski-Harabasz Index	46.36	Dispersión entre grupos aceptable

El coeficiente de Silhouette de 0.186 sugiere una separación moderada entre los clústeres, algo esperable en datos médicos donde las diferencias entre grupos de riesgo suelen ser sutiles. Por su parte, el índice Davies-Bouldin de 1.146 respalda esta interpretación, indicando que los clústeres presentan una definición razonablemente adecuada.

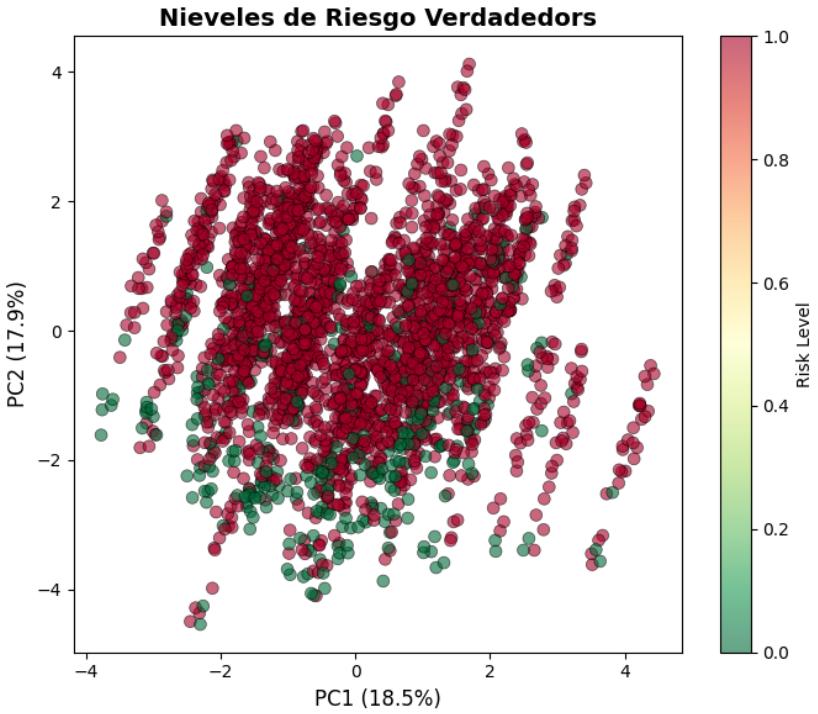

Figure 2: Visualización de clústeres obtenidos con DBSCAN proyectados en el espacio de las dos primeras componentes principales.

En la Figura 2, se observa la distribución de los clústeres obtenidos tras la proyección con PCA. La visualización revela:

- Una concentración de puntos de bajo riesgo (verde) en la región inferior izquierda, correspondiente a valores bajos en ambas componentes principales
- Una dispersión mayor de puntos de alto riesgo (rojo) en las regiones superior y derecha del gráfico
- Presencia de puntos aislados clasificados como ruido, que corresponden a casos atípicos con combinaciones inusuales de características

El análisis muestra que existe una correspondencia moderada entre los clústeres identificados por DBSCAN y los niveles de riesgo reales. El clúster de mayor densidad concentra principalmente casos de bajo riesgo, mientras

que los casos de alto riesgo tienden a estar más dispersos o en regiones de menor densidad.

6 Conclusiones y discusión

Al usar DBSCAN me permitió descubrir subgrupos dentro del conjunto de datos, sin requerir especificar el número de clústeres. A diferencia de algoritmos como K-Means, DBSCAN maneja bien la presencia de ruido y datos no lineales, lo cual es útil para contextos médicos donde las variables fisiológicas pueden presentar alta variabilidad.

La combinación de métricas (Silhouette, Davies-Bouldin y Calinski-Harabasz) brindó una evaluación integral de la coherencia de los grupos. Los resultados sugieren que la estructura encontrada refleja posibles perfiles de riesgo clínico, lo que podría servir para futuros modelos predictivos en neonatología.

References

[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] *A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial* Pavel, Andreea M et al. The Lancet Child Adolescent Health, Volume 4, Issue 10, 740 - 749

[3] MacBean V, Lunt A, Drysdale SB, Yarzi MN, Rafferty GF, Greenough A. Predicting healthcare outcomes in prematurely born infants using cluster analysis. Pediatr Pulmonol. 2018 Aug;53(8):1067-1072. doi: 10.1002/ppul.24050. Epub 2018 May 23. PMID: 29790677.

[4] Domino Data Lab. *Topology and Density-Based Clustering.* Recuperado de: `https://domino.ai/blog/topology-and-density-based-clustering` (consultado en 2025).

[5] Esri. *How Density-Based Clustering Works.* ArcGIS Pro Documentation. Disponible en:

https://pro.arcgis.com/es/pro-app/3.3/tool-reference/spatial-statistics/how-de
(consultado en 2025).

[6] Raghda Altaei. *Understand the Math Behind DBSCAN*. Medium (2023). Disponible en: https://raghda-altaei.medium.com/understand-the-math-behind-dbscan-ae51672c042
(consultado en 2025).