Identificación de Patrones de Riesgo en Salud Neonatal mediante Clustering

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

Noviembre 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 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.

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

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

4.1 Preprocesamiento y Selección de Variables

4.1.1 Detección y Tratamiento de Valores Atípicos

Se implementó un enfoque dual para la identificación de outliers:

1. **Método Z-score**: Se calcularon puntuaciones Z estandarizadas para identificar observaciones extremas ($|z| > 3$). Este método permite detectar valores atípicos que se desvían significativamente de la media poblacional en la distribución multivariada.

2. **Rango Intercuartílico (IQR)**: Se aplicó el criterio de Tukey, donde los outliers se definieron como observaciones fuera del rango $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$. Este método es robusto ante distribuciones asimétricas comunes en variables fisiológicas.

El análisis reveló porcentajes variables de outliers por variable, que fueron visualizados mediante boxplots individuales. Estos gráficos permitieron evaluar la distribución y dispersión de cada característica clínica antes del filtrado, identificando la presencia de valores extremos que podrían afectar el desempeño de los modelos.

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.

4.1.3 Estandarización y Reducción Dimensional

Se aplicó **StandardScaler** a las 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.

Para la visualización y análisis exploratorio se implementó **Análisis de Componentes Principales (PCA)**, una técnica de reducción dimensional que transforma las variables originales correlacionadas en un conjunto de componentes principales ortogonales que capturan la máxima varianza. Las primeras dos componentes explicaron el 26% de la varianza total (PC1: 13.05%, PC2: 12.68%), siendo necesarias 12 componentes para capturar el 90% de la variabilidad. Esta distribución relativamente uniforme indica que los datos de salud neonatal son inherentemente multidimensionales, sin que exista un único factor dominante.

4.2 Fase I: Aprendizaje No Supervisado con DBSCAN

4.2.1 Fundamentos del Algoritmo DBSCAN

Se aplicó el algoritmo **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) para identificar estructuras 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 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.

4.2.2 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:

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

donde ϵ es el radio máximo de vecindad y $dist(p_i, p_j)$ 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 $|\mathcal{N}_\epsilon(p_i)| \geq minPts$, 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 $|\mathcal{N}_\epsilon(q)| < minPts$ 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.

4.2.3 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 utilizando el algoritmo de Nearest Neighbors
2. Ordenar estas distancias en orden descendente
3. Graficar las distancias ordenadas e identificar el "codo" o punto de inflexión en la curva resultante

El punto de inflexión indica el valor óptimo de ϵ, ya que representa la transición entre puntos en regiones densas (donde las distancias al k-ésimo vecino son pequeñas) y puntos en áreas dispersas (donde estas distancias aumentan abruptamente). Los parámetros optimizados fueron $\epsilon = 2.4491$ y $minPts = 10$.

4.2.4 Métricas de Evaluación del Clustering

La calidad del agrupamiento se evaluó mediante tres métricas complementarias:

Coeficiente de Silhouette (S): Cuantifica la relación entre la separación de diferentes clústeres y la similitud entre puntos de un mismo clúster. Para cada punto i, se calcula como:

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

donde $a(i)$ es la distancia promedio entre el punto i y todos los demás puntos dentro del mismo clúster, y $b(i)$ es la distancia promedio entre el punto i y todos los puntos en el clúster más cercano. El coeficiente de Silhouette toma valores en el rango $[-1, 1]$.

Índice de Davies-Bouldin (DBI): Evalúa la similitud promedio entre cada clúster y su clúster más similar. Se calcula en varios pasos:

Índice de Calinski-Harabasz (CHI): También conocido como Criterio de Relación de Varianza, mide el cociente entre la dispersión entre clústeres y la dispersión dentro de los clústeres.

4.3 Fase II: Aprendizaje Supervisado con Random Forest

4.3.1 Fundamentos de Random Forest

Random Forest es un algoritmo de aprendizaje automático que construye múltiples árboles de decisión durante el entrenamiento y combina sus predicciones mediante votación mayoritaria. Formalmente, Random Forest se define como un clasificador consistente en una colección de clasificadores estructurados en árbol $\{h(\mathbf{x}, \Theta_k), k = 1, 2, \ldots\}$, donde $\{\Theta_k\}$ son vectores aleatorios independientes e idénticamente distribuidos, y cada árbol emite un voto unitario para la clase más popular en la entrada $\mathbf{x}$.

El algoritmo introduce aleatoriedad de dos maneras:

1. **Muestreo aleatorio de datos**: Para cada árbol, se genera una muestra bootstrap de tamaño N mediante muestreo con reemplazo del conjunto de entrenamiento original (similar a bagging). Aproximadamente 1/3 de las observaciones originales quedan fuera de cada muestra (datos Out-of-Bag, OOB).

2. **Selección aleatoria de características**: En cada nodo del árbol, se seleccionan aleatoriamente m variables de las M totales ($m \ll M$, típicamente $m = \sqrt{M}$), y la mejor división se determina solo sobre estas m variables.

4.3.2 Configuración del Modelo

Se configuró el modelo con los siguientes hiperparámetros:

- `n_estimators=100`: Número de árboles en el bosque
- `random_state=42`: Semilla aleatoria fija para garantizar reproducibilidad
- Criterio de división: Índice de Gini para medir la impureza en cada nodo
- Profundidad máxima: Sin restricción (cada árbol crece hasta su máxima extensión)
- Tamaño mínimo del nodo hoja: 1 observación

4.3.3 Preparación de Datos para Clasificación

La variable objetivo `risk_level` (nivel de riesgo) fue recodificada para agrupar categorías de bajo riesgo en una sola clase (`risk_level` $< 1 \rightarrow 0$), simplificando el problema de clasificación multiclase. El conjunto de datos filtrado se dividió mediante la función `train_test_split` de scikit-learn:

- **Conjunto de entrenamiento**: 80% de los datos ($n = 2400$ observaciones aproximadamente)
- **Conjunto de prueba**: 20% de los datos ($n = 600$ observaciones aproximadamente)
- Muestreo aleatorio estratificado para mantener la proporción de clases en ambos conjuntos

4.3.4 Métricas de Evaluación del Clasificador

Para evaluar el desempeño del modelo Random Forest en el conjunto de prueba, se calcularon las siguientes 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:

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

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

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

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

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:

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

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:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

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

Importancia de Variables: Random Forest calcula automáticamente la importancia de cada variable mediante la *reducción promedio de impureza de Gini* que produce cada variable en todos los árboles del bosque. Esta métrica, también conocida como "Gini importance" o "Mean Decrease Impurity", se calcula como:

$$Importancia(X_j) = \frac{1}{N_{trees}} \sum_{t=1}^{N_{trees}} \sum_{n \in nodos} I(ndividepor X_j) \times \Delta i(n)$$

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, por tanto, son más relevantes para la clasificación. Esta información permite identificar qué características clínicas (edad gestacional, peso, frecuencia cardíaca, etc.) son más determinantes para predecir el nivel de riesgo neonatal.

4.4 Integración de Enfoques No Supervisado y Supervisado

La combinación de métodos no supervisados (DBSCAN) y supervisados (Random Forest) proporciona una visión más amplia, aprovechando las fortalezas de cada enfoque para comprender los patrones y validar los resultados desde diferentes perspectivas.

El algoritmo **DBSCAN** permite descubrir la estructura de los datos, identificando grupos de riesgo que emergen sin información previa sobre las etiquetas. Su análisis se basa en la densidad y proximidad de las observaciones, lo que facilita detectar comportamientos atípicos o subpoblaciones que no serían evidentes mediante técnicas convencionales.

Por otro lado, **Random Forest** utiliza las etiquetas de nivel de riesgo conocidas para construir un modelo predictivo robusto, capaz de estimar la probabilidad de pertenencia a cada categoría y de cuantificar la importancia de las variables clínicas en la clasificación.

Al contrastar los clústeres obtenidos por DBSCAN con las predicciones del modelo supervisado, es posible evaluar el grado de correspondencia entre los patrones naturales presentes en los datos y las categorías de riesgo definidas clínicamente. Cuando ambas perspectivas coinciden, se refuerza la confianza en que las definiciones clínicas reflejan estructuras reales y consistentes en el conjunto de datos.

Así, este enfoque dual representa mejor el análisis de datos médicos, ya que la exploración no supervisada contribuye a poder fundamentar y enriquecer la construcción de modelos predictivos supervisados.

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).

[7] Analytics Lane. *Número óptimo de clústeres con Silhouette e implementación en Python*. Recuperado de: https://www.analyticslane.com/2023/06/23/numero-optimo-de-clusteres-con-silhou
(consultado en 2025).

[8] Towards Data Science. *Davies-Bouldin Index for K-Means Clustering Evaluation in Python*. Disponible en: https://towardsdatascience.com/davies-bouldin-index-for-k-means-clustering-eva
(consultado en 2025).

[9] Analytics Lane. *Identificar el número de clústeres con Calinski-Harabasz en K-Means e implementación en Python*. Recuperado de: https://www.analyticslane.com/2023/06/16/identificar-el-numero-de-clusteres-co
(consultado en 2025).

[10] 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).

[11] Vrushali Y Kulkarni. *Random Forest Classifiers: A Survey and Future*. Telkom University (2013). Disponible en: https://adiwijaya.staff.telkomuniversity.ac.id/files/2014/02/Random-Forest-Cla
$Survey - and - Future.pdf (consultado en 2025)$.

[12] The Machine Learners. *Métricas de Clasificación*. Disponible en: https://www.themachinelearners.com/metricas-de-clasificacion/#Accuracy
(consultado en 2025).