Aprendizaje No Supervisado en la Identificación de Patrones de Desnutrición Infantil en Menores de 5 años en Medellín

Javier Ballén | Ever Contreras | Viviana Galindo | David Puerto

Resumen
La desnutrición infantil constituye un desafío global de salud pública para diversos gobiernos alrededor del mundo. Este problema es influenciado por una variedad de factores que incluyen variables de salud y nutrición, así como aspectos demográficos y socioeconómicos, que resultan en una población altamente heterogénea y que, en cualquier caso, representa un riesgo para el desarrollo y calidad de vida de los niños menores de 5 años. Con el objetivo de identificar y diferenciar los diversos perfiles de nutrición entre esta población vulnerable, se propone el uso de técnicas de aprendizaje no supervisado como Clustering combinado con PCA para la reducción de dimensionalidad, centrado en los factores más dominantes. Este enfoque permitirá descubrir y clasificar patrones distintivos de desnutrición en la ciudad de Medellín, facilitando intervenciones más dirigidas y políticas de salud pública eficientes.

1. Introducción

En Medellín, la desnutrición infantil se presenta como un desafío crítico de salud pública que afecta fuertemente a niños menores de cinco años, lo cual tiene consecuencias devastadoras en su desarrollo físico y cognitivo, afectando su calidad de vida y capacidad de aprendizaje. Dada la complejidad y la variabilidad de las condiciones socioeconómicas y demográficas en diferentes comunidades, existe una necesidad urgente de enfoques analíticos que permitan una comprensión más profunda y localizada de este problema.

La pregunta que pretende resolver este proyecto es: ¿Cómo pueden las técnicas de clustering basadas en variables de salud y nutrición, demográficas y socioeconómicas ayudar a identificar y diferenciar perfiles de desnutrición en niños menores de 5 años en Medellín, y cuáles son los factores que más contribuyen al diagnóstico de desnutrición? Este estudio serviría de guía para gestionar políticas de salud pública en general y supervisar programas de nutrición diferenciados de acuerdo con las características de cada grupo.

Por lo tanto, la utilización del clustering como técnica de aprendizaje no supervisado se enmarca en la necesidad de identificar grupos homogéneos dentro de una población heterogénea. El proyecto espera, mediante la aplicación de esta técnica, destacar los factores de intervención urgentes, lo cual podría mejorar la coordinación de esfuerzos del gobierno local.

2. Revisión preliminar de la literatura

La malnutrición, que incluye tanto la falta como el exceso de nutrientes, es un problema global significativo, especialmente la desnutrición en menores de 5 años. Recientes estudios utilizan analítica de datos y aprendizaje automático para identificar patrones y factores de desnutrición, destacando la necesidad de enfoques avanzados para abordar este desafío de salud pública.

En el contexto internacional, Striessnig y Bora (2019) realizaron un estudio en India utilizando datos de la Encuesta Nacional de Salud Familiar de India (NFHS-4) y empleando técnicas de PCA y Clustering jerárquico con el criterio de Ward clasificaron los distritos según el estado nutricional de niños menores de 5 años, identificando patrones de malnutrición basados en indicadores como el retraso en el crecimiento, emaciación y bajo peso. Factores como la educación femenina y el acceso a recursos básicos se asociaron con mejores resultados, mientras que la pobreza y el bajo IMC materno se relacionaron con peores resultados nutricionales.

Por otra parte, Leyso y Palatino (2020) enfocaron su estudio en la malnutrición en niños menores de 5 años en Marinduque, Filipinas, identificando clusters de bajo peso y sobrepeso mediante un censo provincial de 2014 a 2016. Utilizando el modelo de escaneo espacial elíptico de Kulldorff, ajustado por edad y estatus socioeconómico, descubrieron cuatro clusters significativos en Boac, Buenavista, Gasan y Torrijos. El estudio enfatiza la importancia de políticas públicas localizadas y efectivas basadas en la detección de clusters.

En un enfoque similar, Bitew et al. (2021) centraron su estudio en predecir la desnutrición en niños menores de cinco años en Etiopía destacando la eficacia de técnicas como XGBTree, k-nearest neighbours, random forest, y redes neuronales. Los modelos identificaron factores críticos de riesgo como acceso al agua, anemia, edad del niño, tamaño al nacer y peso materno, demostrando cómo el aprendizaje automático puede mejorar la predicción y manejo de la desnutrición infantil.

Adicionalmente, Pavitra *et al.* (2021) analizaron los determinantes del retraso en el crecimiento en niños en Armenia, utilizando datos de la Encuesta Demográfica y de Salud desde 2000 hasta 2015. Emplearon regresión y descomposición de Oaxaca para examinar la influencia de factores como la educación materna, el peso al nacer y el tipo de residencia en la malnutrición infantil. El estudio reveló disparidades socioeconómicas y regionales significativas, enfatizando la necesidad de considerar la ubicación geográfica y condiciones locales en las estrategias para combatir la malnutrición.

Finalmente, en estudios más locales en Colombia, Castillo y Suarez-Ortegón (2023) examinaron la doble carga de malnutrición en niños colombianos menores de cinco años usando datos de ENSIN 2015. A través de regresión logística, encontraron que factores como el sexo, la etnicidad indígena y la región de residencia están significativamente asociados con la malnutrición. El estudio subraya la necesidad de políticas públicas que consideren las variaciones regionales y demográficas para abordar eficazmente la malnutrición infantil en Colombia.

De forma complementaria, Loaiza *et al.* (2023) abordaron la analítica de datos para explorar las causas de la desnutrición infantil en Medellín, aplicando metodologías como CRISP-DM y SEMMA. Mediante técnicas como minería de datos y Machine Learning, identificaron factores sociales, económicos y de salud relacionados con la desnutrición en niños de 0 a 5 años. Sus hallazgos subrayan la importancia de datos de alta calidad para mejorar la comprensión de la desnutrición y formular políticas públicas más efectivas.

Aun cuando varios de los estudios anteriores utilizaron desde modelos de regresión hasta Machine Learning para identificar factores de riesgo y patrones de malnutrición, nuestro enfoque se centra específicamente en la aplicación de técnicas de Clustering para agrupar poblaciones según características similares de desnutrición, permitiendo una intervención más dirigida y personalizada.

3. Descripción de los datos

La base *sivigila_desnutricion* se descargó de la Plataforma Nacional de Datos Abiertos de Colombia y contiene datos suministrados por la Secretaría de Salud de Colombia. La base tiene 2.802 registros de niños menores de 5 años con diagnóstico confirmado de desnutrición aguda entre los años 2016 y 2021 en Medellín.

En la Tabla 1 se relaciona el diccionario con las 26 variables de la base. Se observa que 14 variables son de tipo numérico (int64), 8 son de tipo carácter (string) y 4 son de tipo decimal (float64).

Tabla 1

Diccionario de variables

Nombre	Tipo	Descripción Variable
ID	int64	Número consecutivo
semana	int64	semanas del año de 1 a 53
edad	int64	Edad
uni_med_	int64	Unidad de medida:\n0= No aplica, 1=Años, 2=Meses, 3=Días, 4=Horas, 5=Minutos SD=Sin informacion
sexo_	string	M=Masculino, F=Femenino, SD=Sin informacion
nombre_barrio	string	Texto asociado a la tabla de barrios definidos por la entidad territorial, Vacios se diligencian con "Sin iNformacion", Sin ubicación en zona urbana.
comuna	string	Texto asociado a la tabla de barrios definidos por la entidad territorial, Vacios se diligencian con "Sin iNformacion", Sin ubicación en zona urbana.
tipo_ss_	string	Tipo de Régimen de seguridad social:\nC= Contributivo, S=Subsidiado, P=Excepción, E=Especial,\nN= No asegurado, I= Indeterminado/Pendiente, SD=Sin informacion.
cod_ase_	string	Codigo de la aseguradora
fec_con_	string	Fecha de Consulta
ini_sin_	string	Fecha de inicio de síntomas
tip_cas_	int64	Tipo de caso:\n1=Sospechoso, 2= Probable , 3=Confirmado por laboratorio , 4=Confirmado por clinica , 5= Confirmado por nexo epidemiológico.
pac_hos_	int64	Paciente hospitalizado.\n1= Si, 2=No.
peso_nac	int64	Peso al nacer.\n>=900 y <=5000
talla_nac	float64	Talla al nacer.\n>=30,0 y <=55,0
edad_ges	int64	>=0 y <=45
t_lechem	int64	Tiempo de leche materna.\nnumero
e_complem	int64	Tiempo de alimentación complementaria.\nnumero
crec_dllo	int64	1= Si\n2= No
esq_vac	int64	esquema de vacunacion.\n1= Si\n2= No\n3= Desconocido
carne_vac	int64	Carnet de vacunación.\n1= Si\n2= No
peso_act	float64	peso actual
talla_act	float64	talla actual
per_braqu	float64	Perimetro braqueal.\nnumero
evento	string	Descripción del evento notificado
year_	int64	Año de notificación

En el procesamiento de datos se encuentra que las variables *talla_nac*, *peso_act, talla_act* y *per_braqu* tienen valores decimales separados por "," y ".", por lo que se reemplaza la "," por "." para que los valores puedan ser correctamente interpretados como decimales por Python. Por otro lado, teniendo en cuenta que la variable *edad* puede estar dada en días, meses o años, se crea una nueva variable de tipo decimal *edad_mes*, que relaciona la edad del niño en meses.

Tabla 2

Estadísticas descriptivas

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
id	2802.0	NaN	NaN	NaN	1401.5	809.012052	1.0	701.25	1401.5	2101.75	2802.0
semana	2802.0	NaN	NaN	NaN	26.838687	14.567048	1.0	14.0	27.0	39.0	53.0
sexo_	2802	2	M	1670	NaN	NaN	NaN	NaN	NaN	NaN	NaN
nombre_barrio	2802	365	SIN INFORMACION	62	NaN	NaN	NaN	NaN	NaN	NaN	NaN
comuna	2802	28	Manrique	271	NaN	NaN	NaN	NaN	NaN	NaN	NaN
tipo_ss_	2802	6	C	1620	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cod_ase_	2802	55	EPS010	754	NaN	NaN	NaN	NaN	NaN	NaN	NaN
fec_con_	2802	1286	15/06/2018	12	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ini_sin_	2802	1212	01/01/1900	577	NaN	NaN	NaN	NaN	NaN	NaN	NaN
tip_cas_	2802.0	NaN	NaN	NaN	4.0	0.0	4.0	4.0	4.0	4.0	4.0
pac_hos_	2802.0	NaN	NaN	NaN	1.862241	0.344708	1.0	2.0	2.0	2.0	2.0
peso_nac	2802.0	NaN	NaN	NaN	2503.56424	927.607889	0.0	2291.0	2750.0	3050.0	5000.0
talla_nac	2802.0	NaN	NaN	NaN	41.434297	16.586429	0.0	44.0	48.0	50.0	55.0
edad_ges	2802.0	NaN	NaN	NaN	35.349036	9.535904	0.0	37.0	38.0	39.0	42.0
t_lechem	2802.0	NaN	NaN	NaN	7.02177	7.871824	0.0	1.0	6.0	10.0	99.0
e_complem	2802.0	NaN	NaN	NaN	4.821913	5.205596	0.0	2.0	6.0	6.0	90.0
crec_dllo	2802.0	NaN	NaN	NaN	1.101713	0.302325	1.0	1.0	1.0	1.0	2.0
esq_vac	2802.0	NaN	NaN	NaN	1.171663	0.481144	1.0	1.0	1.0	1.0	3.0
carne_vac	2802.0	NaN	NaN	NaN	1.437901	0.496217	1.0	1.0	1.0	2.0	2.0
peso_act	2802.0	NaN	NaN	NaN	8.28576	3.012254	2.0	6.2	7.9	10.2	50.0
talla_act	2802.0	NaN	NaN	NaN	78.168415	13.771134	45.0	67.0	76.7	89.0	116.0
per_braqu	2802.0	NaN	NaN	NaN	8.974126	6.312378	0.0	0.0	12.0	13.0	30.0
evento	2802	1	DESNUTRICION AGUDA EN MENORES DE 5 AÑOS	2802	NaN	NaN	NaN	NaN	NaN	NaN	NaN
year_	2802.0	NaN	NaN	NaN	2018.576374	1.576804	2016.0	2017.0	2018.0	2020.0	2021.0
edad_mes	2802.0	NaN	NaN	NaN	18.050471	13.677111	0.17	8.0	12.0	24.0	48.0

En la Tabla 2 se relacionan las estadísticas descriptivas de las variables de la base. Entre los resultados, se destaca el valor máximo del peso al nacer de 5.000 gr., que corresponde a un único individuo. Este valor está fuera de los rangos normales, pero según la literatura puede ser posible, aunque son casos muy raros.

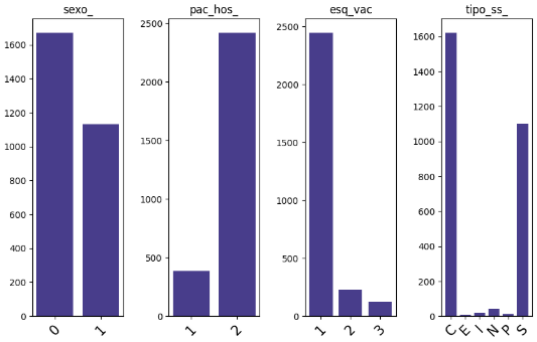
La variable *edad* tiene una dispersión alta, observando que hay niños desde los 5 días de nacido hasta los 4 años. Sin embargo, el 75% de los datos se concentran en niños de hasta

2 años. Las variables *t_lechem* y *e_complem* tiene como valores máximos 99 y 90, los cuales son outliers y serían valores errados, asumiendo que las medidas de estas variables están dadas en meses.

Como se observa en la Figura 1, el 59.6% de los registros pertenecen a niños varones, la mayoría de los niños y las niñas no están hospitalizados y cuentan con esquema de vacunación. Adicionalmente, el 58% de los individuos se encuentran en régimen contributivo, seguido de aproximadamente 39% en régimen subsidiado.

Figura 1

Gráfico de barras de las variables sexo, paciente hospitalizado, esquema de vacunación y tipo de seguridad social

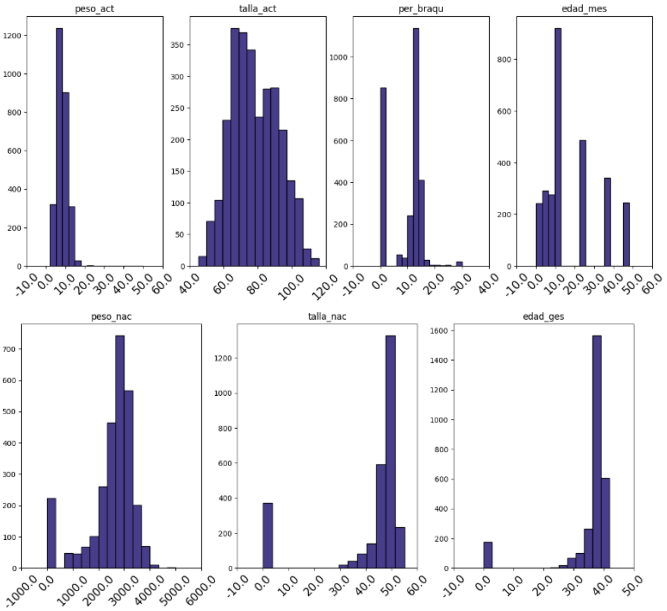


No existen valores nulos en las variables, sin embargo, se observa que las variables *peso_nac*, *per_braqui*, *talla_nac* y *edad_ges* tienen los siguientes valores en cero: 223, 853, 371 y 177, respectivamente, como se observa en la Figura 2. Estos valores no son coherentes con la naturaleza de dichas variables.

En la Figura 2, también se resalta que hay niños con pesos actuales mayores a 20 kg., dichos outliers no son coherentes con los pesos asociados a niños menores de 5 años, por lo que estos registros serán eliminados.

Figura 2

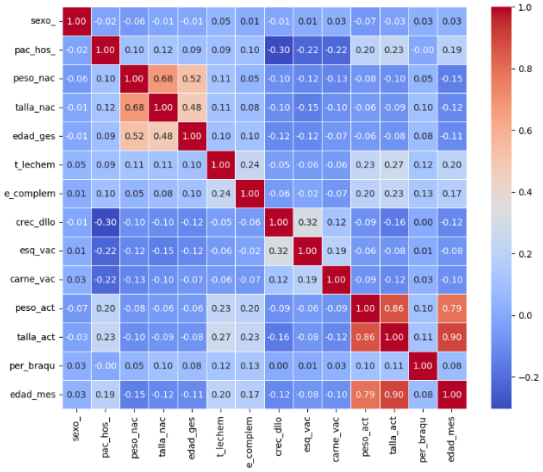
Histogramas de las variables de peso y talla actual, perímetro braquial, edad, peso y talla al nacer y las semanas de gestación



En la Figura 3 se observa que hay una correlación fuerte entre las variables peso y talla actual y la edad, y una correlación moderada entre las variables peso y talla al nacer y semanas de gestación. Ambas correlaciones son consistentes con lo observado en campo. Es importante resaltar, que las correlaciones de las últimas variables puedan estar siendo afectadas por la cantidad de valores en cero que tienen estas, por lo que se esperaría que esta correlación sea más fuerte.

Figura 3

Mapa de calor – correlaciones variables



4. Propuesta metodológica

4.1. Estrategia para el manejo de variables que presentan valores 0

Puesto que se tienen variables importantes para el éxito del presente ejercicio que cuentan con valores 0 en varios registros, a saber: 'per_braqui' (perímetro braquial), 'peso_nac' (peso al nacer), 'talla_nac' (talla al nacer) y 'edad_ges' (tiempo de gestación al momento del nacimiento), para poder aprovechar al máximo posible todos los registros de la base de datos, se plantea la siguiente estrategia de imputación de datos:

Considerando el histograma de la variable 'per_braqui', se aprecia claramente la prevalencia de un rango específico entre 12 y 14, por lo que para completar los datos faltantes en este caso se tomará la Moda.

Respecto a los datos faltantes de ‘peso_nac’, ‘talla_nac’ y ‘edad_ges’, de nuevo considerando los histogramas, se tomará la Moda para la primera y tercera variables, las cuales son las de menores registros en 0 y aprovechando la correlación existente entre las 3 variables, se calcularán con un modelo de regresión de Machine Learning los datos faltantes de ‘talla_nac’.

4.2. Estrategia para la reducción de variables

Se plantea emplear el método de PCA para simplificar el número de variables que se empleará en el modelo de clustering y también para poder determinar los factores clave en la detección de desnutrición en la población infantil menor de 5 años en Medellín.

4.3. Estrategia para el Clustering

Puesto que no se tiene determinado un número esperado de clusters para este ejercicio, se plantea emplear el método de DBSCAN, el cual también permite el uso de métodos de medición de distancias que funcionan bien con combinaciones de variables numéricas y categóricas (Distancia de Gower). Esta sería la primera aproximación y en función de las conclusiones obtenidas, se plantearía emplear un método adicional, para contrastar resultados.

Bibliografía

Ankalaki, S., G. Biradar, V., Naik P., K. K. y S. Hukkeri, G. (2024). A deep learning approach for malnutrition detection. *International Journal of Online and Biomedical Engineering (iJOE)*, *20*(06), pp. 116–138. https://doi.org/10.3991/ijoe.v20i06.46919

Bitew, F.H., Sparks, C.S. y Nyarko, S.H. (2022). Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia. *Public Health Nutrition*, 25(2), pp. 269-280. doi:10.1017/S1368980021004262

Castillo, A.N. y Suarez-Ortegón, M.F. (2023). Dual burden of individual malnutrition in children 1–4 years: Findings from the Colombian nutritional health survey ENSIN 2015. *Pediatric Obesity*, 18(6):e13020. doi:10.1111/ijpo.13020

Leyso, NLC. y Palatino, MC. (2020) Detecting local clusters of under-5 malnutrition in the province of Marinduque, Philippines using spatial scan statistic. *Nutrition and Metabolic Insights,* 13, pp.1-6. doi:10.1177/1178638820940670

Loaiza, A.A., Moreno, W.G. y Ríos, J.D. (2023). Proceso de analítica de datos aplicado a la desnutrición infantil en niños de 0 a 5 años en la ciudad de Medellín. *ITM Institución Universitaria.*

Paul, P., Arra, B., Hakobyan, M., Hovhannisyan, M.G. y Kauhanen, J. (2021). The determinants of under-5 age children malnutrition and the differences in the distribution of stunting—A study from Armenia. *PLoS ONE*, 16(5):e0249776. https://doi.org/10.1371/journal.pone.0249776

Plataforma Nacional de Datos Abiertos. (2024). *Desnutrición aguda en menores de 5 años.* https://www.datos.gov.co/Salud-y-Protecci-n-Social/Desnutrici-n-aguda-en-menores-de-5-a-os/nnww-hpbf/about_data

Striessnig, E. y Bora, J.K. (2019). Under-Five child growth and nutrition status: spatial clustering of Indian districts. *Vienna Institute of Demography Working Papers*, 03. https://doi.org/10.1553/0x003cb432