Potenciando las Intervenciones de CARE: Segmentación Efectiva de Hogares Colombianos a partir del análisis Datos de Pobreza y Desigualdad

Kelly Niño Ramírez Sonia Katherine Olaya O. Guillermo Alberto Ariza

Juan Carlos Acosta. Laura Cristina Martínez

1. Resumen

Este proyecto desarrolló una solución basada en datos para mejorar las intervenciones de CARE, una ONG que combate la pobreza y la injusticia social en Colombia. A través del estudio "Medición de Pobreza Monetaria y Desigualdad 2023" del DANE, se aplicaron técnicas de *clustering* para agrupar hogares según sus características económicas y demográficas, con el fin de identificar patrones en pobreza y desigualdad que permitan a CARE diseñar estrategias de intervención más efectivas.

El objetivo fue segmentar los hogares de manera precisa para que CARE pueda implementar programas específicos y maximizar su impacto en la reducción de la pobreza en Colombia. Para esto, se realizó una selección de características clave, seguida de un análisis mediante cuatro algoritmos de *clustering*: *K-means*, *K-medoides*, *Clustering Jerárquico* y *DBSCAN*. Tras la evaluación de los resultados, *K-medoides* con 7 *clusters* mostró el mejor rendimiento, ofreciendo el equilibrio ideal entre simplicidad, cohesión interna y facilidad de interpretación, superando métodos más complejos como *DBSCAN* que generaron mayor dispersión.

2. Introducción

CARE ha estado presente en Colombia durante 10 años, implementando intervenciones en comunidades afectadas por la pobreza y la desigualdad. Estas intervenciones incluyen apoyo financiero, programas educativos, de salud y proyectos comunitarios dirigidos a los grupos más vulnerables. La clave para el éxito de estas acciones radica en su capacidad de adaptarse a las necesidades específicas de cada comunidad. Para mejorar el impacto de sus iniciativas, CARE enfrenta el reto de identificar y segmentar los hogares en Colombia de manera precisa, lo cual requiere un análisis detallado de las características económicas y demográficas.

Con base en el estudio de Medición de Pobreza Monetaria y Desigualdad del DANE (2023) y apoyándonos en técnicas de clustering, aplicamos métodos de segmentación de hogares para descubrir patrones ocultos en los datos y mejorar la precisión de las intervenciones. La segmentación basada en clustering ya ha demostrado ser efectiva en estudios internacionales sobre pobreza, y este enfoque permite adaptar mejor los programas de asistencia a las características específicas de cada grupo.

Una de las principales limitaciones fue la alta dimensionalidad del dataset.Para abordar esta situación, se realizó una selección de características clave y reducción de la dimensionalidad, lo que permitió reducir el costo computacional sin comprometer la calidad del análisis. El análisis mostró que K-medoides fue la técnica más adecuada, proporcionando el mejor balance entre simplicidad, cohesión interna y facilidad de interpretación. Con este enfoque, se recomienda a CARE utilizar la segmentación generada para personalizar mejor sus intervenciones, lo que les permitirá maximizar su impacto en la reducción de la pobreza y la desigualdad en Colombia.

3. Materiales y Métodos

Datos utilizados: El presente análisis utilizó datos del estudio "Medición de Pobreza Monetaria y Desigualdad 2023" del DANE, de los cuales se obtuvo un DataFrame con 277.158 registros y 30 columnas (Diccionario de datos). Estas columnas abarcan una variedad de variables categóricas y numéricas que describen las características socioeconómicas de los hogares en Colombia. Las variables categóricas incluyen tipo_area, region, y tenencia_viv, que representan aspectos geográficos y de vivienda. Las variables numéricas abarcan desde ingresos y composición del hogar hasta indicadores de pobreza, así como información sobre la situación laboral y educativa del jefe de hogar.

Proceso de limpieza y preparación de datos: Dado el tamaño del DataFrame, surgió la necesidad de optimizar los tiempos de ejecución y reducir el costo computacional. Para ello, se realizó una selección de características clave, eliminando columnas redundantes o con baja variabilidad. Las columnas con alta frecuencia en un único valor, que no aportaban información diferenciada, fueron descartadas tras calcular la frecuencia del valor más común en cada columna.

Posteriormente, se aplicó un análisis de correlación y multicolinealidad para eliminar variables numéricas altamente correlacionadas. Además, se utilizó el algoritmo Random Forest para calcular la importancia de las características, conservando aquellas más relevantes en relación con pobreza, ingresos y demografía. Entre las variables retenidas se incluyen indicadores económicos clave como ingreso_per_capita, num_pobres, y total_ingresos_hogar, así como variables demográficas como tipo_area y tenencia_viv.

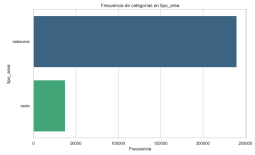
Estadísticas descriptivas: En cuanto a la composición de los hogares, el 86.49% de los hogares se ubican en áreas urbanas (Figura 1a), especialmente en ciudades como Medellín, Cali y Bogotá, mientras que el 13.51% están en zonas rurales (Figura 1b). Un 40.49% de los hogares vive en arriendo, y un 30.91% posee vivienda formal (Figura 1c). El 71% de los hogares no se consideran pobres (Figura 1d), y el 90% no están en situación de indigencia (Figura 1e). En términos de afiliación a la seguridad social, el 90% de los hogares están afiliados a algún sistema (Figura 1f). El 58% de los jefes de hogar están en situación de trabajo, y el 37.22% no cotizan a pensiones (Ver Figura 1i).

En cuanto a las variables numéricas, la mayoría de los hogares reporta 1 o ningún miembro en situación de pobreza o pobreza extrema. En promedio, cada hogar tiene 3 personas. Los ingresos promedio de los hogares son de 2.481.285 COP, con ingresos por subsidios que son bajos en general. Existe una marcada disparidad económica, ya que algunos hogares reportan ingresos considerablemente más altos que otros. El ingreso per cápita refleja esta desigualdad, mostrando una distribución sesgada hacia la izquierda.

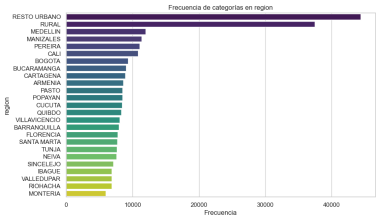
Reducción de dimensionalidad: Finalmente, se aplicó un muestreo aleatorio estratificado del 20% para garantizar la representatividad de la población, preservando las combinaciones relevantes de las variables categóricas. Sobre este dataset reducido y escalado, se aplicó Análisis de Componentes Principales (PCA) para reducir la dimensionalidad, manteniendo el 95% de la varianza original.

Algoritmo de Clustering: Se utilizaron los siguientes algoritmos de clustering: K-means, K-medoides, Clustering Jerárquico, y DBSCAN. En cada caso, se realizaron validaciones cruzadas para ajustar los parámetros, incluyendo el número de clusters (determinados mediante

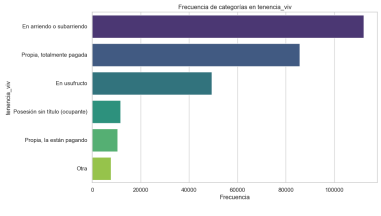
el método del codo y el índice de Silhouette), y en el caso de algunos algoritmos, se probaron distintas métricas de distancia.



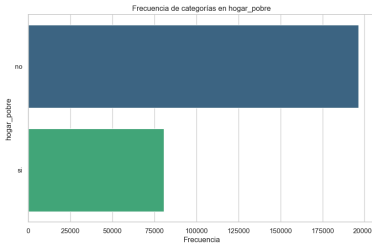
a. Hogares por tipo de área



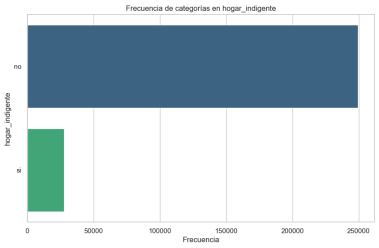
b. Hogares por región



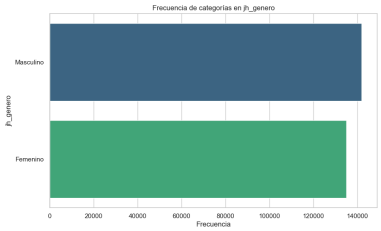
c. Hogares por tenencia de vivienda



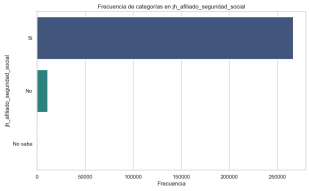
d. Hogares en condición de pobreza



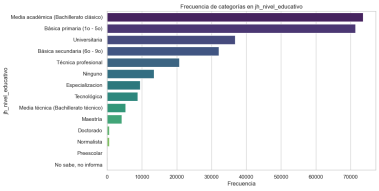
e. Hogares en condición de pobreza extrema



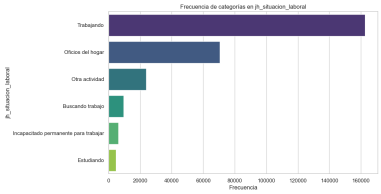
f. Jefes de hogar por género



g. Jefes hogares en seguridad social



h. Jefes de hogares por nivel educativo



i. Jefes de hogares por situación laboral

Figura 1. Distribución en variables categóricas. Fuente: elaboración propia.

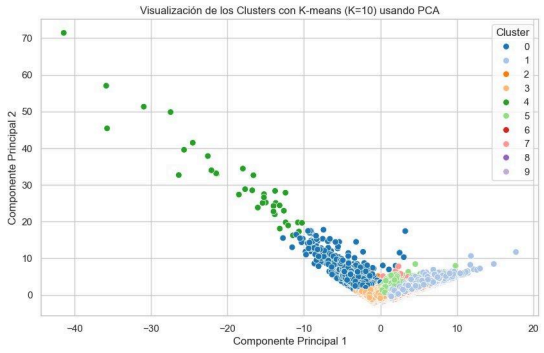
4. Resultados y discusión

En esta sección, se implementaron y compararon cuatro algoritmos de clustering: K-means, K-medoides, Clustering Jerárquico y DBSCAN, con el objetivo de segmentar los hogares en función de sus características económicas y demográficas. Los resultados fueron evaluados mediante el índice de Silhouette, la inercia y la varianza explicada. Los detalles completos de los resultados se pueden consultar en el repositorio de GitHub del proyecto, en el archivo segmentacion.ipynb.

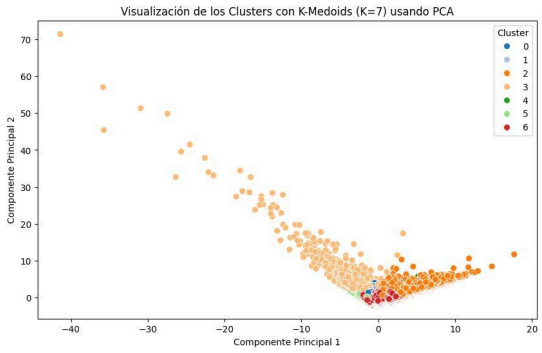
K-means: Se probaron diferentes números de clusters (2 a 30). El método del codo sugirió que 10 clusters era el número óptimo, capturando el 62% de la varianza, mientras que el índice de Silhouette alcanzó su valor máximo con 22 clusters, explicando el 78% de la varianza. Aunque 22 clusters proporcionaron una segmentación más precisa, K=10 logró un equilibrio entre simplicidad y cohesión. La visualización de los clusters para K=10 se muestra en la Figura 2a.

K-medoides: Este algoritmo, que utiliza medoides en lugar de centroides, mostró ser menos sensible a outliers. Se probaron diferentes valores de K y métricas de distancia (Euclidean, Manhattan, Cosine). Con la distancia Cosine, se obtuvieron los mejores resultados, con un índice de Silhouette de 0.2065 y 65% de varianza explicada. K=7 mostró ser la mejor opción, dado que ofreció un buen balance entre simplicidad y cohesión interna. Los clusters resultantes fueron más claros y concentrados, facilitando su interpretación. Los resultados de este método se ilustran en la Figura 2b.

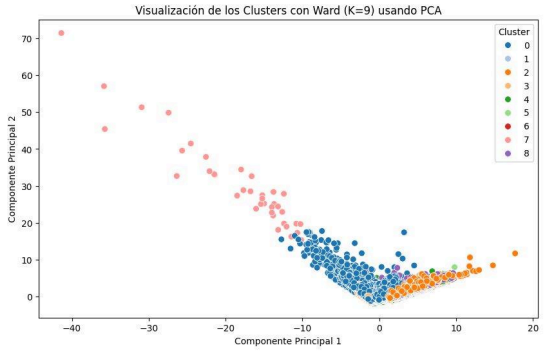
Clustering Jerárquico: Se probaron varios métodos de enlace (Ward, Complete, Average) y métricas de afinidad. El método de Ward con distancia Euclidiana mostró el mejor rendimiento, con un índice de Silhouette de 0.353 para 29 clusters. Sin embargo, se recomienda la solución de 9 clusters para mantener la simplicidad, ya que los resultados con K=29 resultaron demasiado complejos para una implementación efectiva. Los resultados visuales se pueden observar en la Figura 2c.



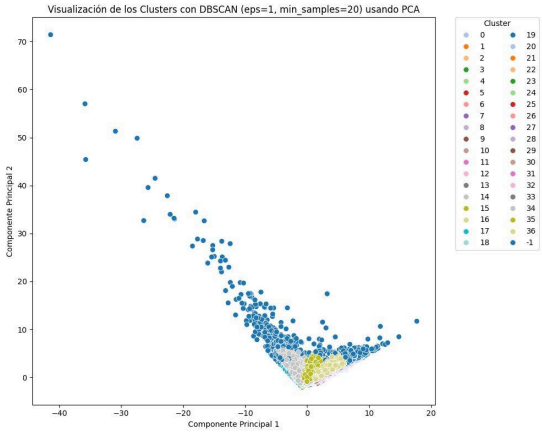
a. K-medias



b. K-medoides



c. Clustering Jerárquico



d. BSCAN

Figura 2. Resultados de Segmentación de los algoritmos de Clustering. Fuente: elaboración propia.

DBSCAN: Este algoritmo basado en la densidad identificó 37 clústeres (Figura 2d) con un índice de Silhouette de 0.3705, y un número razonable de outliers (1572). Aunque DBSCAN fue competitivo, la cantidad de clusters generó complejidad en la interpretación. A pesar de ofrecer una segmentación precisa, la implementación de este método en un entorno práctico sería menos manejable.

Tras la evaluación, se concluye que *K-medoides con K=7* fue el método más adecuado, ya que logró un índice de Silhouette de 0.375, balanceando simplicidad y cohesión interna. La solución generó clusters claros y manejables, a diferencia de otros métodos que presentaban demasiada dispersión. Aunque DBSCAN y Clustering Jerárquico ofrecieron resultados interesantes, su complejidad los hizo menos prácticos para implementar.

Limitaciones y Estudios Futuros

La principal limitación fue la alta dimensionalidad del dataset original, lo que requirió técnicas de reducción de dimensionalidad y selección de características. En estudios futuros, se podrían explorar algoritmos híbridos o técnicas de deep learning para mejorar la segmentación. Además, se podría ampliar la validación cruzada para afinar aún más los parámetros utilizados.

Conclusión

El proyecto ha desarrollado una solución de segmentación efectiva para los hogares colombianos, utilizando datos de pobreza y desigualdad. A través del análisis del estudio "Medición de Pobreza Monetaria y Desigualdad 2023" del DANE, se aplicaron técnicas de clustering para agrupar los hogares según sus características económicas y demográficas. Se evaluaron cuatro algoritmos: K-means, K-medoides, Clustering Jerárquico y DBSCAN, destacando que el método K-medoides con K=7 es el más adecuado. Este enfoque logró un índice de Silhouette de 0.375, evidenciando una cohesión significativa entre los clusters, lo que facilita su interpretación y permite a CARE implementar intervenciones más precisas y efectivas.

La segmentación obtenida no solo mejora la capacidad de CARE para diseñar estrategias adaptadas a las necesidades específicas de cada grupo de hogares, sino que también contribuye a maximizar el impacto de sus programas. Se recomienda que CARE adopte esta metodología en sus futuras intervenciones, asegurando que sus esfuerzos se centren en las comunidades que realmente necesitan apoyo, con el objetivo de reducir la pobreza y la desigualdad en Colombia de manera más eficaz.

Bibliografía

Atuesta, B., Mancero, X., & Tromben, V. (2018). Herramientas para el análisis de las desigualdades y del efecto redistributivo de las políticas públicas. Recuperado de: https://www.cepal.org/es/publicaciones/43678-herramientas-analisis-desigualdades-efecto-redistributivo-politicas-publicas

DANE. (2023). Gran Encuesta Integrada de Hogares – GEIH. Recuperado de: https://www.datos.gov.co/Estad-sticas-Nacionales/Gran-Encuesta-Integrada-de-Hogares-GEIH/mcpt-3dws/about_data

DANE. (2023). *Medición de Pobreza Monetaria y Desigualdad 2023*. Recuperado de https://microdatos.dane.gov.co/index.php/catalog/835

DANE. (2023). Pobreza y desigualdad. Departamento Administrativo Nacional de Estadística. https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-multidimensional

Departamento Nacional de Planeación (DNP). (2019). Índice de pobreza multidimensional por departamento 2019. Recuperado de https://colaboracion.dnp.gov.co/CDT/Prensa/Publicaciones/Publicaci%C3%B3n%20Ipm%20deptal.pdf

Di Capua, L., Brun, C., & Pellegrini, J. L. (n.d.). Desigualdad multidimensional de los hogares: Tipos de hogares y variables predictoras. Recuperado de https://ri.conicet.gov.ar/bitstream/handle/11336/180326/CONICET_Digital_Nro.416bdb5d-ff1c-4b33-9ae0-08826d9c328c_B.pdf?sequence=2&isAllowed=y

División de Desarrollo Social de la Comisión Económica para América Latina y el Caribe (CEPAL), & Centro Latinoamericano y Caribeño de Demografía (CELADE)-División de Población de la CEPAL. (2010). Pobreza infantil en América Latina y el Caribe. CEPAL. https://repositorio.cepal.org/server/api/core/bitstreams/9b920aaa-1840-471f-942f-3db8fa4faeb1/content

Nalbarte, L., Altmark, S., & Massa, F. (n.d.). Identificación de tipología de pobreza multidimensional a través del enfoque de cluster probabilístico. Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República. Recuperado de https://osf.io/nv962/download