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

En este proyecto, se desarrollará una solución basada en datos para CARE, una ONG internacional dedicada a combatir la pobreza y la injusticia social en Colombia. A pesar de sus esfuerzos durante una década, el impacto de las intervenciones de CARE ha sido limitado. Para mejorar la eficacia de estas acciones, se empleará el estudio "Medición de Pobreza Monetaria y Desigualdad 2023" del DANE, que ofrece datos detallados sobre las características económicas y demográficas de los hogares colombianos. Utilizando técnicas de clustering, se agruparán los hogares de acuerdo con sus características, lo que permitirá identificar patrones emergentes en pobreza y desigualdad.

El objetivo principal del proyecto es segmentar los hogares de manera precisa, lo que permitirá a CARE diseñar e implementar estrategias de intervención más específicas y efectivas. Esta segmentación mejorará significativamente el impacto de los programas de CARE, contribuyendo a la reducción de la pobreza y desigualdad en el país. Además, el proyecto incluirá una revisión de la literatura relevante, un análisis de los datos mediante estadísticas descriptivas y visualizaciones, y una propuesta metodológica que detallará el uso del aprendizaje no supervisado en el proceso.

2. Introducción

Con presencia en Colombia desde hace 10 años, CARE ha llevado a cabo intervenciones en comunidades donde la desigualdad económica y la pobreza son persistentes. Una intervención se refiere a las acciones o programas diseñados para resolver problemas específicos o mejorar situaciones particulares. En el caso de CARE, estas intervenciones pueden incluir apoyo financiero, programas educativos, iniciativas de salud, y proyectos comunitarios, todos orientados a ayudar a los grupos más necesitados. La eficacia de estas intervenciones depende de su capacidad para adaptarse a las necesidades y características específicas de cada comunidad.

Para mejorar el impacto de sus programas, CARE se enfrenta al desafío de identificar y segmentar los hogares en Colombia de manera más precisa. El estudio de Medición de Pobreza Monetaria y Desigualdad (DANE, 2023) detalla las características económicas y demográficas de los hogares, lo que permite descubrir patrones ocultos relacionados con la pobreza y la desigualdad.

En esta perspectiva, buscamos responder a la siguiente pregunta de investigación: *¿Cómo se puede analizar datos económicos y demográficos para identificar y segmentar los hogares en Colombia, para que CARE pueda diseñar intervenciones más personalizadas y efectivas que aborden específicamente las variaciones en pobreza y desigualdad en las comunidades?*

Para abordar este desafío, emplearemos técnicas de clustering, una metodología de aprendizaje no supervisado que permite descubrir patrones ocultos en los datos que no cuentan con una variable de respuesta. Al hacerlo, buscamos segmentar eficazmente los hogares y mejorar la capacidad de CARE para diseñar e implementar estrategias de intervención más ajustadas a las realidades específicas de cada comunidad, maximizando el impacto de sus programas y contribuyendo significativamente a la erradicación de la pobreza y la desigualdad en Colombia.

3. Revisión preliminar de antecedentes en la literatura

Diversas organizaciones nacionales e internacionales han realizado análisis sobre las variables que determinan la pobreza y sus características. La Comisión Económica para América Latina y el Caribe (CEPAL), una de las principales entidades en este ámbito. En sus publicaciones *Herramientas para el análisis de las desigualdades y del efecto redistributivo de las políticas públicas* (Atuesta, 2018) y *Pobreza infantil en América Latina y el Caribe* (CEPAL, 2010), la CEPAL proporciona evalúa desigualdades socioeconómicas y la distribución de los recursos utilizando modelos econométricos y análisis de la incidencia del beneficio y del impuesto. Aunque la CEPAL no utiliza técnicas de aprendizaje no supervisado (ANS), su enfoque podría beneficiarse de estas herramientas para segmentar eficientemente a las poblaciones.

En el contexto colombiano, el Departamento Administrativo Nacional de Estadística (DANE) es clave en la recolección y análisis de datos estadísticos para la formulación de políticas públicas. Su estudio de Pobreza y Desigualdad (DANE, 2023) proporciona datos sobre pobreza multidimensional utilizando los factores de expansión derivados del Censo Nacional de Población y Vivienda, los cuales son valiosos para enriquecer la información del proyecto. Asimismo, el Departamento Nacional de Planeación (DNP) utiliza estos datos para desarrollar y evaluar políticas públicas. En su informe *Índice de Pobreza Multidimensional por Departamento* (DPN, 2019), ofrece un analisis detallado de la pobreza detallados por región, aunque no emplea técnicas de ANS.

En el ámbito académico, estudios como la *Identificación de tipología de pobreza multidimensional a través del enfoque de cluster probabilístico* (Nalbarte, n,d), aplican técnicas de clustering, un enfoque de aprendizaje no supervisado, para clasificar la población en diferentes tipologías de pobreza basadas en múltiples indicadores. Este enfoque permite una segmentación más precisa. Otro estudio relevante es *Desigualdad multidimensional de los hogares: tipos de hogares y variables predictoras* (Di Capua, nd), donde se aplican técnicas de análisis multivariado para identificar distintos tipos de hogares según variables socioeconómicas, utilizando modelos econométricos para encontrar predictores significativos de la desigualdad. La ventaja de emplear ANS frente a los análisis meramente descriptivos es que permite identificar patrones complejos en los datos sin supervisión, facilitando la clasificación precisa de grupos y el diseño de intervenciones más efectivas. Este enfoque mejora el análisis de la pobreza y la desigualdad al ofrecer una segmentación más detallada y una mejor focalización de las intervenciones.

4. Descripción de los datos

A continuación, se presenta un resumen con las principales estadísticas descriptivas de los datos provenientes del Estudio de *Medición de Pobreza Monetaria y Desigualdad* (DANE,2'23). El conjunto de datos se compone de dos dataframes:

Hogares (d*f_hogares*): Este DataFrame proporciona información sobre las características económicas y sociales de los hogares encuestados. Incluye datos sobre la vivienda, como el tipo de propiedad y los pagos de arriendo, así como indicadores económicos clave, como el ingreso total y per cápita de la unidad de gasto. También contiene las líneas de pobreza e indigencia, que determinan los umbrales de pobreza y permiten clasificar los hogares según su nivel de pobreza y pobreza extrema. *df_hogares* tiene 277,158 registros y 20 columnas. Cada columna representa una variable específica del estudio de pobreza monetaria y desigualdad. La descripción de las columnas y su tipo de dato se puede consultar en: Diccionario de datos Dataframe Hogares

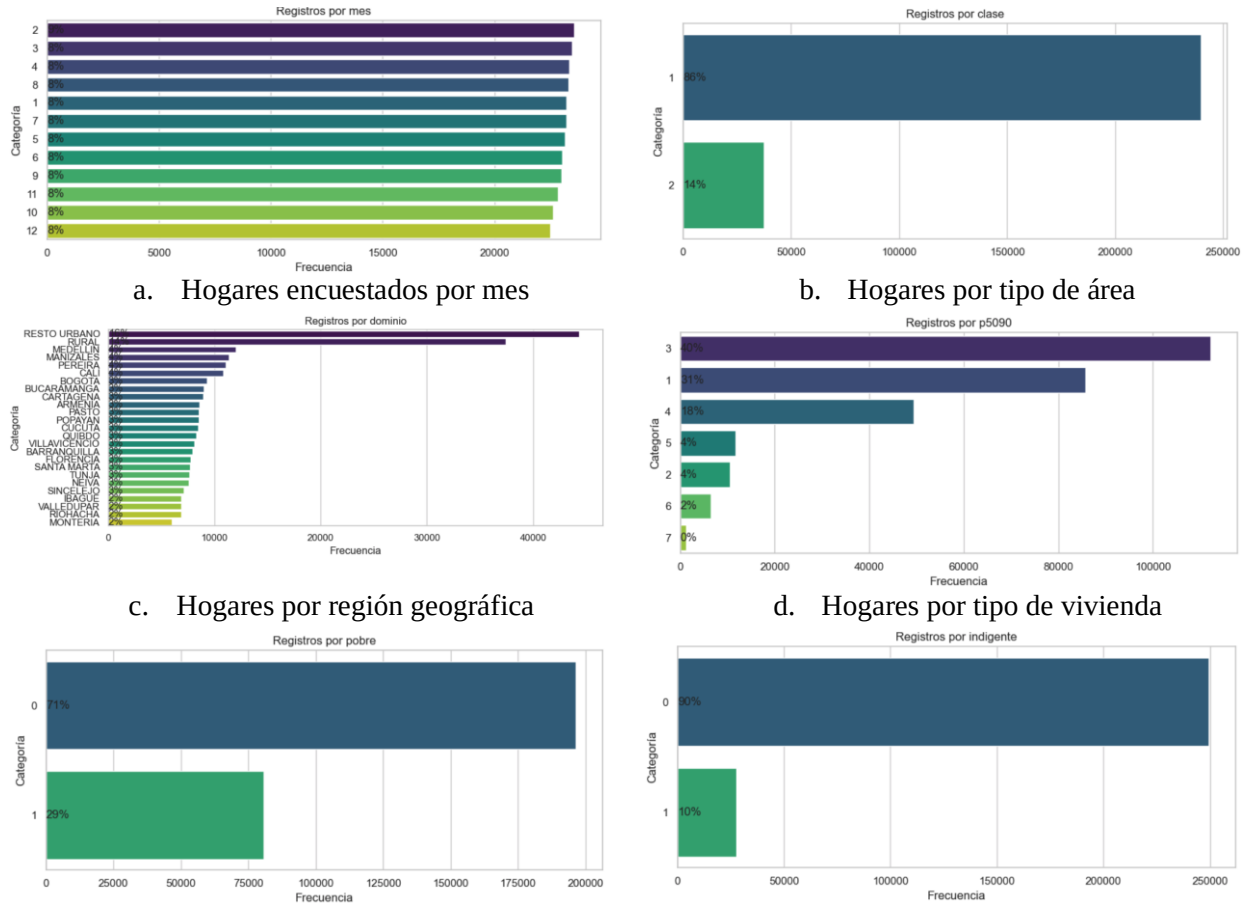
Personas (d*f_personas*): Este DataFrame ofrece información detallada sobre la situación socioeconómica y laboral de los individuos dentro de un hogar. Las variables incluyen datos personales como sexo y edad, parentesco con el jefe del hogar, afiliación y régimen de seguridad social en salud, nivel educativo y ocupación. También se detallan los ingresos y beneficios recibidos, como salario, bonificaciones, subsidios y otras formas de compensación. Además, se documenta la duración y tipo de empleo, si el individuo tiene

un segundo trabajo y si ha recibido o desea recibir más horas de trabajo. Este dataframe permite una comprensión profunda de la situación económica y laboral de los encuestados.

df_personas contiene 814,665 filas y 130 columnas, representando una base de datos extensa con un gran volumen de registros. La descripción de las columnas y su tipo de dato se puede consultar en: Diccionario de datos Dataframe Personas.

El detalle completo del proceso de carga, preprocesamiento de datos, y análisis estadístico de los DataFrames, incluyendo las técnicas empleadas para la limpieza de datos y el manejo de valores nulos, se encuentra disponible en el repositorio de GitHub del Proyecto. Allí, se proporcionan scripts y documentación detallada que permiten replicar los análisis y comprender las decisiones tomadas en cada etapa del proyecto.

Los hogares analizados contienen las siguientes características:



a. Hogares encuestados por mes

b. Hogares por tipo de área

c. Hogares por región geográfica

d. Hogares por tipo de vivienda

e. Hogares en condición de pobreza

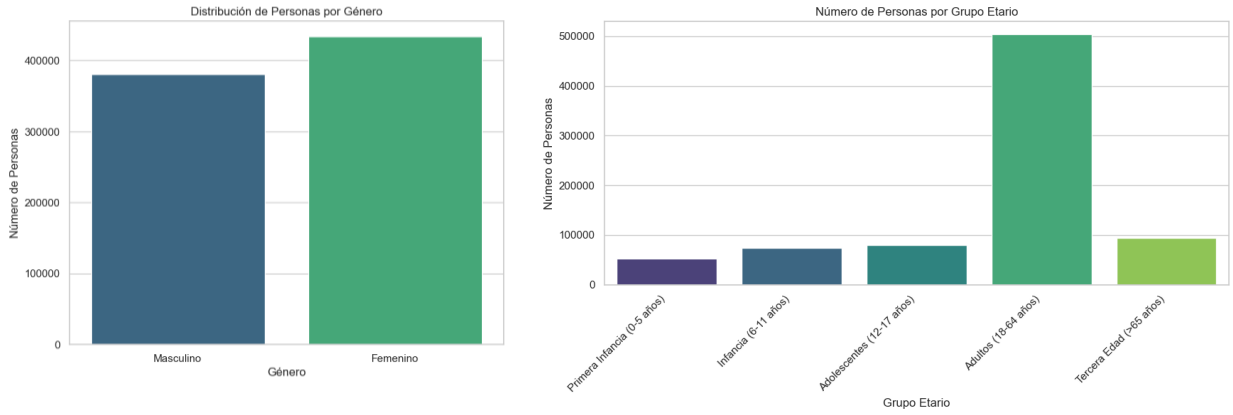
f. Hogares en condición de pobreza extrema

Figura 2. Principales descriptivos de hogares. Fuente: elaboración propia

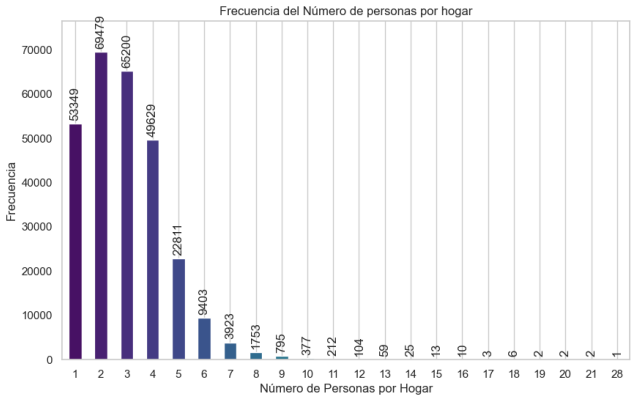
- Las encuestas fueron realizadas en los 12 meses del año 2023, con una participación similar cada mes (Figura 2a)
- El 86% de las encuestas se aplicaron en tipo de área de cabecera (clase), que corresponde al lugar en donde se ubica la sede administrativa de un municipio (Figura 2b)
- Las áreas principales donde se aplicaron fueron resto urbano (16%), rural (14%) y ciudades principales como Medellín, Manizales, Pereira, Cali y Bogotá (Figura 2c).

- El 40% de los hogares viven en arriendo, 31% tiene hogar propio totalmente pagada, 18% viven en usufructo y un 4% posee una casa, pero sin título (Figura 2d).
- El 71% de los hogares encuestados no viven en condición de pobreza (Figura 2e)
- El 10% de los hogares encuestados viven en condición de indigencia (Figura 2f)

En cuanto a las personas tenemos las siguientes características:



a. Distribución por genero b. Distribución por grupos etarios



c. Número de personas por hogar

Figura 3. Características de la población. Fuente: elaboración propia.

- Predomina ligeramente el género femenino (433,666) con un 53.2% frente al género masculino (380,999) con 46,8% (Figura 3a). Esto sugiere que, en el conjunto de datos, las mujeres representan una mayor proporción de la población.
- El análisis de la distribución por grupos etarios (Figura 3b) revela lo siguiente: la mayoría de la población se encuentra en el grupo de *Adultos* (18-64 años), lo que indica una fuerte base laboral y económica. Los grupos de *Primera Infancia* (0-5 años) e *Infancia* (6-11 años) también tienen una representación significativa, sugiriendo la necesidad de invertir en educación y salud infantil. Los *Adolescentes* (12-17 años) representan una porción importante, destacando la importancia validar contra información de acceso a servicios educativos. Aunque la *Tercera Edad* (>65 años) tiene una representación menor, sigue siendo crucial visualizar sus necesidades de salud y bienestar. Este perfil demográfico puede guiar la asignación de recursos y el diseño de intervenciones específicas para cada grupo etario.
- En cuanto a la distribución de personas por hogar, la Figura 3c muestra una predominancia de hogares pequeños (La distribución está claramente sesgada hacia la izquierda, con una concentración alta en el

rango de 1 a 4 personas por hogar) en el conjunto analizado, con una disminución gradual en la frecuencia de hogares a medida que aumenta el número de personas.

5. Propuesta metodológica.

Para llevar a cabo la segmentación de los hogares en Colombia, proponemos utilizar el algoritmo de K-means como técnica principal. El algoritmo K-means se elige para este proyecto debido a su eficiencia en el manejo de grandes volúmenes de datos, con 277,158 registros en df_hogares y 814,665 en df_personas y más de variables de datos. Su capacidad para procesar grandes conjuntos de datos de manera rápida y efectiva es crucial para manejar la magnitud y complejidad de los datos. Además, K-means proporciona resultados fácilmente interpretables, lo que facilita la comprensión de los clusters y su aplicación en el diseño de intervenciones específicas. Su escalabilidad permite una segmentación precisa sin comprometer el rendimiento, lo que es esencial dado el número de variables y registros presentes en los DataFrames.

5.1. Algoritmos y Técnicas Alternativas

K-medoides podría ser útil por su robustez frente a valores atípicos, ya que utiliza puntos de datos reales como centros de los clusters, lo que puede reflejar mejor los perfiles centrales de los hogares. DBSCAN, por su capacidad para identificar clusters de forma arbitraria y manejar datos con ruido, podría revelar patrones complejos y estructuras no esféricas en los datos, proporcionando una visión más detallada de la desigualdad y pobreza. El clustering jerárquico, aunque más costoso computacionalmente, ofrece una visión detallada de la estructura de los datos y las relaciones entre grupos, facilitando la identificación de subgrupos dentro de clusters principales y ayudando a entender mejor las variaciones entre hogares.

5.2. Procedimiento a seguir

- Preparación de Datos: Limpieza y preprocesamiento de df_hogares y df_personas para asegurar la calidad y consistencia de los datos. Esto incluirá la normalización y la selección de variables relevantes para el clustering.
- Aplicación del Algoritmo K-means: Determinación del número óptimo de clusters mediante el método del codo y el índice de Silhouette. Ejecución de K-means para segmentar los hogares y los individuos.
- Evaluación y Validación: Análisis de los resultados obtenidos, incluyendo la interpretación de los clusters y la validación de la estabilidad y coherencia de los clusters.
- Desarrollo de Herramienta Analítica: Creación de una herramienta analítica basada en los resultados del clustering, que proporcionará a CARE una interfaz intuitiva para explorar los datos segmentados. Esta herramienta permitirá a los usuarios visualizar los segmentos de hogares e individuos, analizar sus características específicas y obtener recomendaciones basadas en datos para diseñar intervenciones personalizadas. La herramienta incluirá funcionalidades para la generación de informes en función de los patrones identificados en los datos.

6. 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