Resumen

Mediante el uso de técnicas de aprendizaje no supervisado y a partir simplificación de los datos sin perder información relevante se intenta identificar patrones definidos relacionados a los datos de vivienda referentes o ligados con el índice de pobreza multidimensional en Colombia capturados en la Encuesta Nacional de Calidad de Vida del DANE 2022 para la identificación e implementación de políticas públicas de reducción de pobreza dirigidas a las poblaciones vulnerables resultantes del estudio.
Los datos contenidos se dividen en 93161 observaciones con 35 variables de consultas resultantes de encuestas en el territorio nacional, sin embargo, después del tratamiento inicial de los datos, el tamaño de los datos se redujo a 58772 filas y 29 de los cuales podemos ver directamente relacionados a nivel de multidimensionalidad de pobreza con factores de sanidad, materiales de construcción y seguridad.
Gracias a los resultados dados por el análisis de componentes principales encontramos una real importancia de factores ambientales y materiales de construcción relacionadas con las viviendas correlacionadas directamente con la estimación de la pobreza multidimensional. Aunque se intenta crear un índice unidimensional, se elige utilizar la matriz truncada para una mayor interpretabilidad de los datos.

Introducción

Según estimaciones de la ONU para el año 2022 alrededor de 1200 millones de personas residentes de países en vía de desarrollo se encuentra en estado de pobreza multidimensional, estas cifras fueron calculadas en un esfuerzo conjunto entre el programa de desarrollo de las Naciones Unidas (UNDP) y la Iniciativa de Pobreza y Desarrollo Humano de la Universidad de Oxford. El índice de pobreza multidimensional nos ayuda, a determinar la incidencia de la pobreza, así mismo como su intensidad y es una medida que agrega, educación salud y estándares de vida: 1. Educación, compuesto por los años de escolaridad y asistencia al colegio, 2. Salud, compuesto por nutrición y mortalidad infantil y 3. Estándares de Vida, compuesto por combustible para cocinar, Vivienda, sanidad, agua potable, electricidad y bienes.
En Colombia, en mayo de 2023, el Departamento Administrativo Nacional de Estadísticas (DANE) publicó un informe que mostraba una reducción de alrededor de 300 puntos básicos en el índice de pobreza multidimensional en comparación con el año anterior. La incidencia de la pobreza se redujo en general en un 3.1%, siendo mayor en áreas rurales y centros poblados. Las regiones con las tasas de pobreza más altas fueron la Caribe y la Pacífica (sin Valle del Cauca), En cuanto a la contribución de los componentes del índice, la educación representó el 35% y la vivienda el 13.4% en 2022.
Con miras a la identificación, creación e implementación de políticas públicas dirigidas a la reducción generalizada de la pobreza en Colombia, se nos ha contratado por medio de una ONG aliada al departamento de Prosperidad Social de Colombia para realizar un desarrollo utilizando aprendizaje no supervisado basados en la encuesta nacional de calidad de vida del DANE para Colombia realizada en 2022 específicamente de la data recolectada de datos de vivienda, que sea capaz de identificar si en la población nacional y basados en el índice de vivienda adecuada (Adequate Hoising index) propuesto por el banco mundial, se identifican componentes o clústeres específicos que sean relacionados a estimaciones de nivel de pobreza multidimensional en el territorio Colombiano para que estos mismos sean atacados de manera proactiva para mejorar la calidad de vida y reducción de pobreza.
Para establecer políticas, es importante entender que este índice de pobreza construye el perfil de cada hogar y las personas que lo habitan basado en 10 criterios específicos que se ubican entre 0 y 1 en el cual a mayores valores mayor estado de pobreza.
De la mano del índice multidimensional de pobreza, tenemos el Índice de vivienda Adecuada, este índice fue introducido por el banco mundial en el cual se estima cuales debería ser los criterios para considerarse una vivienda adecuada, El primer criterio introducido es el de acceso al agua, con concordancia al programa conjunto de monitoreo de la OMS y de la UNICEF(JMP). El segundo criterio considerado es acceso a sanidad. El tercer criterio está definido como un adecuado espacio para vivir. Como quinto criterio tenemos la seguridad de la tenencia de la vivienda contra situaciones

como el desalojo forzoso, y otras amenazas. Como sexto factor encontramos el acceso a la electricidad siguiendo la definición de la agencia internacional de energía. Finalmente, como el ultimo componente, el acceso a cocinar con combustibles limpios.
Adicionalmente, encontramos literatura científica cercana a el desarrollo de análisis relacionados a el aprendizaje no supervisado donde encontramos el desarrollo de un modelo a partir de K medias enfocado en la identificación de factores de pobreza en Malasia para así a partir de su identificación la creación de políticas de alivio, este articulo fue publicado en el año 2021 en la revista[1]. También se estudiaron artículos enfocados en componentes adicionales del índice multidimensional como lo es el tipo de combustible para cocinar como se presenta en el artículo[2], en el cual se intenta identificar perfiles específicos para la implementación de políticas de transiciones energéticas en hogares de bajos recursos. Finalmente, se exploró cómo el crecimiento urbano afecta la pobreza mediante la identificación de clústeres utilizando el método de estimación por K-medias[3].
Los tres artículos de investigación nos presentan resultados similares en cuanto los métodos de aplicación del aprendizaje no supervisado, resaltando especialmente la posibilidad de identificación de clúster específicos con características determinadas relacionales a la pobreza multidimensional, sin embargo dada la complejidad y la dimensiones que aborda el índice, se menciona que perfiles adicionales tienes que ser incluidos dentro de los cálculos de identificación, para atacar la pobreza en todos sus aspectos adicionalmente con perfiles de evolución temporal, para así poder determinar si las políticas decididas realmente cambian el componente de las agrupaciones durante su aplicación, por lo mencionado anteriormente nosotros consideramos que como punto de partida es necesario establecer desde forma granular el estudio de los componentes del índice de pobreza multidimensional, así generando a futuro una integración vertical donde podamos abordar todos los tres aspectos, educación, salud y calidad de vida en un solo conjunto integrado de políticas públicas.

Materiales y Métodos

Finalizada la revisión de literatura, y puesta en marcha el proceso de elaboración del modelo, y basándonos en la encuesta nacional de calidad de vida del DANE para Colombia realizada en 2022, específicamente en los datos recopilados por la encuesta "Datos de la Vivienda" que recopila los resultados de la encuesta realizada a 93,161 hogares colombianos, así como primera, mediante el atributo “df.shape”, se exploran las dimensiones del dataframe, este tiene 93161 filas o registros y 35 columnas o características. A través del atributo “df.size”, se explora el número total de datos.
A través del atributo “df.index”, se explora cómo están indexados los datos. Mediante le tributo “df.columns” se exploran los nombres de cada una de las columnas del dataframe. Luego de explorar los atributos básicos del dataframe y estructura de datos se concluye que ciertas columnas no aportan mucho al modelo, además de esto otras requieren ciertas transformaciones.
Se empieza entonces con un análisis a las variables binarias, se realizan analíticas descriptivas a través del atributo “.describe()”, así podemos entonces analizar estas estadísticas desde la perspectiva de proporción de unos y ceros que para el caso de estudio serían respuestas positivas y negativas respectivamente y con esto a través del cuadro generado para cada característica se puede analizar por ejemplo, para la variable “P8520S5” la cual toma el valor de “1” cuando el encuestado tiene presencia de acueducto en su hogar y “0” cuando no, se observa una media de 0.8991 lo cual desde la óptica del análisis de variables binarias se traduciría en que en promedio 89.91% de los hogares colombianos encuestados cuentan con acueducto, mientras que el 10.09% restante no, también puede observarse la desviación estándar de este valor que es 0.3010 y nos da una idea de la variabilidad de los datos alrededor de la media, la función también nos arroja valores mínimos, máximos y cuartiles, sin embargo al ser variables binarias son estadísticas que quizá no nos brindan mucha información para nuestro estudio.

---

[1] Abdul Rahman M, Sani NS, Hamdan R, Ali Othman Z, Abu Bakar A. A clustering approach to identify multidimensional poverty indicators for the bottom 40 percent group. PLoS One. 2021 Aug 2;16(8):e0255312. doi: 10.1371/journal.pone.0255312. PMID: 34339480; PMCID: PMC8328299.

[2] André Paul Neto-Bradley, Rishika Rangarajan, Ruchi Choudhary, Amir Bazaz, A clustering approach to clean cooking transition pathways for low-income households in Bangalore, Sustainable Cities and Society, Volume 66, 2021, 102697, ISSN 2210-6707

[3] Philippe Apparicio, Mylène Riva et Anne-Marie Séguin, « A comparison of two methods for classifying trajectories: a case study on neighborhood poverty at the intra-metropolitan level in Montreal », *Cybergeo: European Journal of Geography* [En ligne], Espace, Société, Territoire, document 727, mis en ligne le 04 juin 2015, consulté le 26 septembre 2023.

A nivel general, en todas las estadísticas se observa una predominancia de media alta en los valores lo cual se traduce a que la mayoría de los encuestados respondieron “si” para estas preguntas, en la mayoría por encima del 80%.
Finalmente, se evalúa la relación lineal entre las variables categóricas a través de 3 elementos: la matriz de correlaciones, el mapa de calor y el correlograma. A través de la matriz de correlaciones, se expresa a través de una tabla la relación lineal entre los pares posibles de variables, se ve reflejado de forma gráfica en el mapa de calor el cual toma valores cercanos a amarillo a medida que va aumentando la relación lineal positiva entre los variables y a morado cuando aumenta la relación lineal negativa. En las variables, puede observarse que no hay una predominancia marcada de relaciones lineales entre las variables, sin embargo, hay algunos pares de variables con cierta predominancia a la linealidad (aunque no muy fuerte), por ejemplo se puede observar cierta tendencia a la linealidad negativa entre las variables “P8520S4A1” y “P4015” con un indicador de correlación de -0.34, se observa también cierta tendencia a correlación lineal positiva entre por ejemplo las variables “P5661S4” y “P5661S2” con un índice de correlación de 0.47.
Dado la complejidad y profundidad de los análisis mencionados anteriormente, como parte de visualización e interactividad, en el Anexo del código podemos encontrar las gráficas y estadísticas resultado de las transformaciones de los datos.
En el proceso de selección de la metodología de agrupación más apropiada para nuestro análisis, optamos por el algoritmo K-Medias debido a su capacidad para distribuir de manera efectiva nuestra muestra en grupos coherentes. Esta elección se basó en una evaluación exhaustiva de múltiples algoritmos de clustering que mostraremos a continuación:

Algoritmo K-Means: Se llevó a cabo un análisis de clustering utilizando el algoritmo K-Means, cuyo objetivo principal es agrupar el conjunto de datos en clústeres basados en similitud. Se exploraron diversos valores de k (número de clústeres) y se evaluaron dos métricas clave: la varianza intra-cluster y el coeficiente de Silhouette. Estos criterios condujeron a la determinación del número óptimo de clústeres, que resultó ser 4, considerado apropiado para abordar nuestra pregunta de investigación.

Algoritmo K-Medoides: Posteriormente, se realizó otro análisis de clustering utilizando el algoritmo K-Medoides, el cual, en lugar de centroides, emplea medoides como representantes de los grupos de datos similares. Se aplicó este algoritmo con k = 4, seleccionado en función del coeficiente de Silhouette. Vale la pena destacar que este algoritmo requirió un considerable poder computacional debido al gran volumen de registros y a la alta dimensionalidad de la base de datos. La distribución de puntos en los clústeres resultantes se evaluó en términos del coeficiente de Silhouette y la varianza intra-cluster. Si bien el coeficiente de Silhouette indicó cierta coherencia en la agrupación, aún se percibió la necesidad de una mayor precisión.

Algoritmo de DBSCAN: Finalmente, se empleó el algoritmo DBSCAN, basado en la densidad de puntos cercanos, para llevar a cabo un análisis de clustering. A diferencia de los algoritmos anteriores, DBSCAN no requiere la predefinición del número de clústeres, pero sí necesita ajustes como el radio del vecindario y la cantidad mínima de muestras. Estos ajustes se optimizaron mediante técnicas como "NearestNeighbors" y "KneeLocator". Sin embargo, este algoritmo generó un exceso de clústeres que no se consideraron adecuados para el tipo de análisis realizado con los datos.

Resultados y Discusión

Para darle solución a nuestra pregunta problema, implementamos múltiples algoritmos de reducción de dimensionalidad y clustering, con el fin de poder comprender las agrupaciones subyacentes de los datos y a través de la reducción de dimensionalidad, tener la capacidad de comprender de manera más precisa las principales características que lo componen y de esta forma identificar el nivel de pobreza multidimensional de los mismos. Teniendo en cuenta lo anteriormente mencionado, este proceso se llevó a cabo de la siguiente manera:

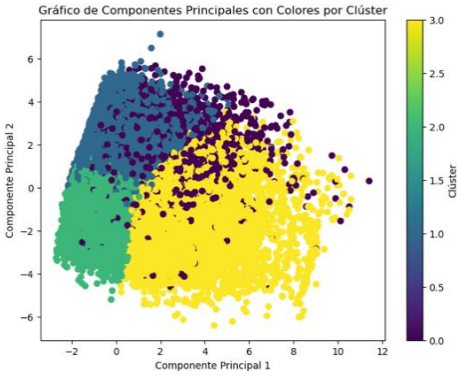
Inicialmente, se procedió a escalar los datos recopilados, ya que algunos de ellos estaban en diferentes magnitudes. Este paso es esencial para asegurarse de que todos los datos tengan la misma escala y facilitar un análisis de clustering más adecuado.
Dado que nuestro conjunto de datos contiene variables categóricas y variables ficticias (dummy), se desaconseja el empleo de Análisis de Componentes Principales (PCA), ya que esta técnica está diseñada para el tratamiento de variables continuas. En su lugar, se persigue el objetivo de reducir la dimensionalidad mediante la utilización de una matriz truncada, lograda a través de la Descomposición en Valores Singulares (SVD).
Se realizó una descomposición en valores singulares para evaluar la posibilidad de reducir las 23 dimensiones originales a través de una matriz truncada, manteniendo al menos el 90% de la variabilidad de los datos. Tras varias iteraciones, se optó por retener 18 componentes, que explican el 91.32% de la varianza. Sin embargo, estos resultados no cumplieron con el objetivo de crear un índice unidimensional para describir la calidad de los hogares colombianos en términos sociodemográficos. Sin embargo, se opta por usar la matriz truncada al momento de implementar los algoritmos de clustering para tener un componente adicional de interpretabilidad de los datos.

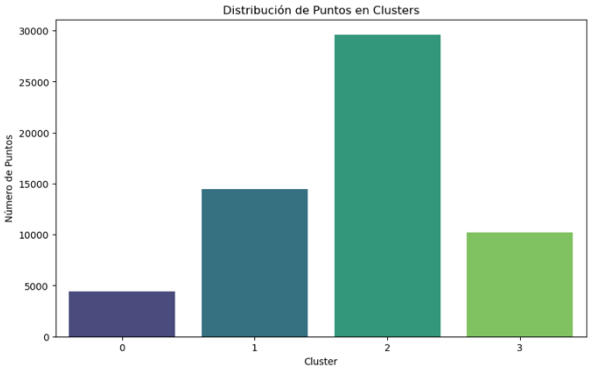
Loadings Componente #1	
Variable	Loading
P5661S2	0.372669
P5661S4	0.356238
P5661S3	0.355978
P5661S9	0.342411
P5661S7	0.342255
P5661S6	0.337645
P5661S5	0.292216
P5661S1	0.264776

Loadings Componente #2	
Variable	Loading
P4567	0.400713
P4015	0.371219
P4005	0.271886
P4065S1	0.112865
P5661S7	0.094615
REGION	0.079765
P4065S3	0.056257
P4065S2	0.048700

Loadings Componente #3	
Variable	Loading
CANT_HOGARES_VIVIENDA	0.704409
CANT_HOG_COMPLETOS	0.704409
P8520S5	0.034293
P8520S3	0.027850
P4065S1	0.027034
P4015	0.025940
P4065S3	0.018258
P4065S2	0.017004

CP1 parece estar relacionado con factores ambientales o de calidad de vida, mientras que CP2 podría estar relacionado con los materiales de construcción de las viviendas. Estas interpretaciones se basan en las cargas de las variables en cada componente principal.
El algoritmo K-Medias demostró ser especialmente adecuado para nuestros objetivos de investigación, ya que permitió la formación de clústeres que capturaron de manera efectiva la estructura subyacente de nuestros datos. Una de las ventajas clave de K-Medias radica en su simplicidad conceptual y su eficiencia computacional, lo que lo convierte en una elección práctica para conjuntos de datos de mayor tamaño.





Aunque consideramos alternativas como el algoritmo K-Medoides y DBSCAN, observamos que K-Medias logró una distribución más precisa de los datos en clústeres significativos, lo que facilitó una interpretación más clara de los patrones en nuestros datos. Si bien cada algoritmo tiene sus propias ventajas, como la robustez de K-Medoides y la adaptabilidad de DBSCAN a la densidad de puntos, la elección de K-Medias se justifica por su capacidad para abordar específicamente nuestros objetivos de investigación y la calidad de los resultados obtenidos.

En el contexto de este informe, es importante destacar que, si bien las técnicas de aprendizaje no supervisado, como K-Means y DBSCAN, junto con la reducción de dimensionalidad a través de PCA y SVD, ofrecen herramientas valiosas para abordar la identificación de patrones relacionados con la pobreza multidimensional en datos de vivienda, existen algunas potenciales limitaciones en su implementación que merecen discusión.

Primero, la calidad de los resultados depende en gran medida de la elección adecuada de parámetros y del preprocesamiento de datos. La selección incorrecta de k en K-Means o de los hiperparámetros en DBSCAN, por ejemplo, podría llevar a agrupaciones subóptimas. Además, la elección de la cantidad de componentes principales en PCA o SVD requiere una cuidadosa consideración para equilibrar la reducción de dimensionalidad con la preservación de la información esencial.

Otra limitación para considerar es la interpretación de los clústeres o componentes principales resultantes. Si bien estas técnicas pueden revelar patrones ocultos en los datos, la interpretación precisa de estos patrones puede requerir análisis adicionales y conocimiento de dominio para contextualizarlos adecuadamente. En cuanto a análisis y estudios posteriores, sería beneficioso profundizar en la validación de clústeres y la evaluación de su estabilidad a lo largo del tiempo. Además, se pueden realizar análisis específicos sobre la relación entre las características de vivienda identificadas y los indicadores de pobreza multidimensional. También se podría considerar la inclusión de otras variables socioeconómicas para un análisis más completo.

Conclusiones

Este proyecto se enfoca en abordar el desafío de identificar la pobreza multidimensional en Colombia utilizando datos de vivienda de la Encuesta Nacional de Calidad de Vida del DANE en 2022. El índice de pobreza multidimensional es esencial para comprender la situación socioeconómica, cubriendo aspectos como educación, salud y estándares de vida. Según estimaciones de la ONU, más de 1.2 mil millones de personas en países en desarrollo viven en pobreza multidimensional, y Colombia no es una excepción.

El estudio utiliza técnicas de aprendizaje no supervisado y reducción de dimensionalidad para identificar patrones relacionados con la pobreza multidimensional en los datos de vivienda, clasificando viviendas en niveles de riesgo social. Se analizan múltiples algoritmos de clustering y reducción de dimensionalidad, y se opta por el algoritmo K-Medias debido a su eficiencia y capacidad para identificar patrones coherentes en los datos.

El análisis de componentes principales revela la importancia de factores ambientales y materiales de construcción en la pobreza multidimensional. Aunque se intenta crear un índice unidimensional, se elige utilizar la matriz truncada para una mayor interpretabilidad de los datos.

A pesar de las ventajas de las técnicas utilizadas, se reconocen limitaciones, como la elección de parámetros y la interpretación de resultados. Se sugieren análisis futuros que profundicen en la validación de clústeres, la relación entre las características de vivienda y la pobreza, y la inclusión de variables socioeconómicas adicionales, dicho esto, este estudio proporciona una base sólida para comprender la pobreza multidimensional en Colombia y ofrece valiosas perspectivas para la toma de decisiones y políticas públicas, destacando la eficacia del algoritmo K-Medias en la identificación de patrones significativos en los datos de vivienda.

Bibliografía

- Organización de las Naciones Unidas (ONU). (2022). Informe sobre Desarrollo Humano 2022. Nueva York: ONU.
- Departamento Administrativo Nacional de Estadísticas (DANE). (2023, mayo). Boletín Técnico de Pobreza Multidimensional en Colombia 2022. Bogotá, Colombia: DANE.
- UNDP: Programa de las Naciones Unidas para el Desarrollo. (2022). Informe sobre Desarrollo Humano 2022. Nueva York: Autor.
- Iniciativa de Pobreza y Desarrollo Humano de la Universidad de Oxford: Oxford Poverty & Human Development Initiative. (2022). MPI Database. Oxford: Autor.
- PNUD: Programa de las Naciones Unidas para el Desarrollo. (2010). Manual de medición de la pobreza multidimensional. Nueva York: Autor.
- Banco Mundial: Banco Mundial. (2005). Adequate Housing: Concepts, indicators, and policies. Washington, DC: Autor.
- Abdul Rahman M, Sani NS, Hamdan R, Ali Othman Z, Abu Bakar A. A clustering approach to identify multidimensional poverty indicators for the bottom 40 percent group. PLoS One. 2021 Aug 2;16(8):e0255312. doi: 10.1371/journal.pone.0255312. PMID: 34339480; PMCID: PMC8328299.
- André Paul Neto-Bradley, Rishika Rangarajan, Ruchi Choudhary, Amir Bazaz, A clustering approach to clean cooking transition pathways for low-income households in Bangalore, Sustainable Cities and Society, Volume 66, 2021, 102697, ISSN 2210-6707
- Philippe Apparicio, Mylène Riva et Anne-Marie Séguin, « A comparison of two methods for classifying trajectories: a case study on neighborhood poverty at the intra-metropolitan level in Montreal », Cybergeo: European Journal of Geography [En ligne], Espace, Société, Territoire, document 727, mis en ligne le 04 juin 2015, consulté le 26 septembre 2023. URL : http://journals.openedition.org/cybergeo/27035 ; DOI : https://doi.org/10x`x`.4000/cybergeo.27035