Exploración y Segmentación de la Población en Habitabilidad de Calle: Un Enfoque de Aprendizaje No Supervisado

Resumen

La población en situación de calle enfrenta desafíos significativos que van más allá de la atención inmediata en alimentación, refugio e higiene. Es fundamental comprender las causas subyacentes y los problemas actuales para desarrollar soluciones efectivas. En Colombia, el multi lleva a cabo censos a esta población, recolectando grandes volúmenes de datos complejos y multidimensionales. Sin embargo, el uso de estos datos para la toma de decisiones puede verse limitado si se presta atención a variables que no son realmente relevantes. Este proyecto tiene como objetivo utilizar la información del censo del DANE de 2021 y aplicar dos métodos de aprendizaje no supervisado para descubrir patrones ocultos en los datos. El primer método consiste en la reducción de dimensionalidad mediante el Análisis de Componentes Principales (PCA), con el fin de identificar las variables más relevantes y concentrarse en ellas. A partir de esta reducción, se aplicará un algoritmo de agrupamiento (clustering) que permita clasificar a la población en grupos con características similares. Este enfoque permitirá centrar la atención en las variables más importantes y, al agrupar a la población en función de características comunes, facilitará la toma de decisiones más precisas, mejorando la efectividad de los programas de atención, prevención y rehabilitación, entre otros.

Introducción

Según el boletín de prensa No 410 de 2022 del Ministerio de Salud y Protección Social, en Colombia más de 34.000 personas viven en situación de calle, lo que constituye un desafío importante para la sociedad. Estas personas no solo enfrentan la falta de recursos esenciales como alimentación, refugio y atención médica, sino que también están atrapadas en un ciclo de exclusión social que perpetúa su condición. Ante este reto, el Departamento Nacional de Estadística (DANE) ha llevado a cabo varios censos, como el Censo de Habitantes de Calle 2021 (CHC 2021), con el propósito de caracterizar a esta población mediante una extensa gama de variables, que incluyen la edad, el género, el estado de salud, y las razones de permanencia en habitabilidad de calle. No obstante, surge la pregunta: ¿cómo asegurar que el vasto conjunto de datos recopilados por el DANE sea utilizado de manera eficiente para el diseño de políticas efectivas a nivel interinstitucional? El volumen y la diversidad de información podrían dispersar la atención hacia variables menos relevantes, dificultando la creación de soluciones focalizadas.

Este problema de interés surge en el contexto organizacional de entidades gubernamentales y no gubernamentales que trabajan en la formulación de políticas y programas de atención para la población en situación de calle, por ende los clientes potenciales de este estudio incluyen el Ministerio de Salud y Protección Social, alcaldías y gobernaciones, así como ONGs que se dedican a la atención y rehabilitación de esta población, estas entidades necesitan datos precisos y relevantes para diseñar intervenciones efectivas que aborden las necesidades específicas de las personas sin hogar.

Este proyecto pertenece al área del aprendizaje no supervisado, ya que se enfoca en descubrir patrones ocultos en los datos sin la necesidad de etiquetas predefinidas buscando dar respuesta a esa cuestión aplicando métodos como el Análisis de Componentes Principales (PCA) para la reducción de dimensionalidad y el uso de algoritmos de clustering para la agrupación de datos. La motivación detrás de este enfoque radica en identificar los patrones clave dentro de la población en situación de calle, proporcionando a las instituciones datos relevantes para el diseño de programas específicos, lo cual puede ser valioso para la formulación de políticas públicas y programas de intervención social. De este modo, se espera abordar de manera más precisa las necesidades de cada grupo identificado, lo que podría traducirse en intervenciones y programas de atención mucho más efectivos.

Además, se pretende explorar la utilidad de los modelos de aprendizaje no supervisado en contextos como este, donde los programas de atención suelen basarse en la experiencia. El programa de la Unión Europea Adelante (Facilidad para la Cooperación Triangular UE-ALC) busca fomentar las relaciones entre América Latina y el Caribe con Europa. Uno de sus proyectos, conocido como Red Calle, se enfoca en abordar problemáticas de la población en situación de calle. Este proyecto es un ejemplo de

colaboración entre la experiencia europea en políticas públicas y los gobiernos de Brasil, Chile, Colombia, Costa Rica, Paraguay y Uruguay. A través de este diálogo birregional, se identifican y adaptan las mejores prácticas a las condiciones socioculturales de cada país. No obstante, se observa que estos proyectos a menudo no incluyen la recolección sistemática de datos para tomar decisiones basadas en evidencia. Incorporar el análisis de datos en estos programas podría fortalecerlos al proporcionar señales claras y patrones que permitan focalizar y mejorar la efectividad de las intervenciones.

Este estudio aplica el Análisis de Componentes Principales (PCA) al Censo de Habitantes de Calle 2021 (CHC 2021) en Colombia, identificando factores clave como la salud y el uso de sustancias, en línea con investigaciones internacionales como las de Ngo y Turbow (2019) y Lasode (2019). A diferencia de estos estudios, que revisaron literatura previa o evaluaron la vulnerabilidad social, nuestro enfoque utiliza directamente datos del censo, proporcionando una perspectiva específica y localizada. Una limitación del estudio es la dependencia de la calidad y alcance de los datos del CHC 2021, lo que podría afectar la generalización de los resultados. Recomendamos complementar el PCA con análisis cluster para una segmentación más detallada, y mejorar la recopilación de datos en futuros censos, lo que podría enriquecer el análisis y ofrecer insights más profundos para el diseño de políticas efectivas dirigidas a esta población vulnerable.

Materiales y Métodos

Descripción de los datos

Los datos se tomaron del catálogo central de datos del Departamento Administrativo Nacional de Estadística (DANE), entidad colombiana que produce y comunica información estadística, específicamente del Censo de Habitantes de Calle – CHC – 2021.

La información disponible cuenta con 130 variables y 6250 registros. Todas las variables son de tipo numérico real. Al ser un censo, la información registrada proviene de preguntas realizadas a la población en estudio, por lo que la mayoría pueden clasificarse como variables categóricas, algunas de respuesta múltiple, otras binaria. No obstante, todas se clasifican como numéricas porque cada categoría fue asignada a un valor numérico, durante la creación de la fuente.

Se cuenta con variables de identificación y ubicación del encuestado(a), demográficas como la edad y el género, otras que describen la condición de vida actual como el lugar donde duerme y el tiempo que lleva viviendo en la calle. De salud y discapacidad, consumo de sustancias psicoactivas, la percepción de seguridad y vulnerabilidad, acceso a servicios y ayuda, y otras de educación y tipo de empleo.

Se identifico que las variables que cuentan con una gran cantidad de valores nulos son las orientadas a conocer si el encuestado(a):

- Ha tenido alguna enfermedad, accidente o problema de salud.
- Ha sido diagnosticado con alguna enfermedad.
- Consume una sustancia psicoactiva específica.
- Conoce de programas de atención para habitantes de calle.

De las estadísticas descriptivos podemos decir que la población en situación de calle:

- Tiene una edad máxima de 75 años, mínima de 15 años, promedio de 41 años y el 50% de la población tiene hasta 39 años.
- Lleva en esta situación máximo 60 años, como mínimo menos de un año, en promedio 10 años y el 50% de la población ha permanecido en esta condición por 7 años.

Respecto a la correlación entre variables, se observaron algunas correlaciones, especialmente debido al tipo de pregunta, por ejemplo: el grupo de variables P17, que principalmente quiere conocer: en los últimos 30 días, ¿usted tuvo alguna enfermedad, accidente, problema odontológico o algún otro problema de salud?, profundizan en opciones para responder la pregunta:

- ¿Cuál(es): Lesión, intoxicación o envenenamiento causado por accidente?
- ¿Cuál(es): Lesión intencional por parte de terceros?

- ¿Cuál(es): Problema mental o emocional?, etc.

Con base en la exploración de los datos se puede decir de la población encuestada:

- La mayoría vive actualmente en el departamento de Norte de Santander (54) representando un 19.53%, seguido de Bolívar (13) con un 13.63%.
- En gran proporción son hombres (88.79%).
- Duermen principalmente en la calle (puente, andén, parque, alcantarilla, carreta etc.) (78.08%), otros en un dormitorio (hotel, paga diario, inquilinato, residencia, camarote) (14.77%) y otros en una institución (7.15%).
- La principal razón por la que comenzaron a vivir en la calle es por consumo de sustancias psicoactivas (33.48%), seguido de conflictos o dificultades familiares (25.70%) y dificultades económicas (15.42%), entre otros.
- El nivel educativo más alto que tiene la mayoría es básico primaria (37.13%), seguido por preescolar (17.69%), entre otros.
- Principalmente reconocer su orientación sexual como Heterosexual (91.31%).
- La mayoría (más del 70%) no recibe ayuda, y los que la reciben, principalmente la reciben de su familia.
- Principalmente consumen cigarrillos, marihuana y basuco, y no consumen Heroína y Pepas.

Las etiquetas relacionadas a cada categoría y opción son tomadas directamente del lugar desde donde se descarga la fuente de datos, más no vienen incluidas en los datos.

Limpieza de datos

Se realizaron las siguientes acciones para controlar los valores nulos sobre la fuente de datos del Censo de Habitantes de Calle (CHC 2021):

- En primer lugar, se calculó el porcentaje de valores nulos por cada fila/registro y fueron eliminadas aquellas con un porcentaje mayor al 50%.
- Se realizó un chequeo una a una de las columnas/variables de la fuente y se aplicaron las siguientes acciones:
    - Para las columnas que hacían referencia a valores binarios en donde solo se estuvieran registrando las respuestas afirmativas 1 (Ejemplo: ¿Consume cierta sustancia?) se imputó el valor 0.
    - Para el caso de las variables asociadas a edades de consumo donde el entrevistado no tenía consumo de la sustancia se imputó el valor -1.
    - Para la variable de identidad de género se imputó el valor -1 haciendo referencia a una respuesta no registrada.
    - Para las variables que tenían más del 90% de valores nulos se eliminó la columna.

Finalmente, el resultado fue un dataset con 112 variables y 5341 registros.

Resultados y Discusión.

Implementación de algoritmo

En la implementación del análisis de componentes principales (PCA), se prepararon los datos utilizando un ColumnTransformer para estandarizar las variables numéricas y aplicar codificación One-Hot a las variables categóricas. Posteriormente, se configuró PCA sin un límite predefinido de componentes para explorar la estructura completa de los datos. Al analizar la varianza explicada por cada componente y aplicar el método de Kaiser, se determinó que los primeros 78 componentes eran significativos, ya que cada uno de ellos poseía valores propios superiores al promedio. Estos 78 componentes explicaron aproximadamente el 88% de la variabilidad total de los datos, proporcionando una comprensión profunda y amplia de las principales influencias y patrones subyacentes en el conjunto de datos. Este enfoque permitió una reducción efectiva de la dimensionalidad mientras se retenía la mayoría de la información crítica contenida en los datos originales.
Los primeros 5 componentes que agrupan cerca del 30% de la varianza se asocian a las siguientes dimensiones de la encuesta según sus loadings:

- PC1: Edades de inicio de consumo de sustancias y consumo actual (basuco, marihuana, cigarrillo), consumo de sustancias como razón principal de seguir viviendo en la calle.
- PC2: Edades de inicio de consumo de sustancias y consumo actual (pepas, alcohol, cocaina), recepción de ayuda de alimentación. Edad del entrevistado.
- PC3: Tiempo de habitabilidad de calle, Edad, Edad de inicio de consumo de basuco, Ningún contacto con familiares, basuco como sustancia principal.
- PC4: Edad de inicio de consumo de alcohol y cigarrillo, consumo de alcohol y cigarrillo, no consumo de basuco.
- PC5: Edad de inicio de consumo de marihuana y cigarrillo, consumo actual de marihuana y cigarrillo, consumo principal de marihuana.

Implementación del Clustering Jerárquico Aglomerativo:

1. **Preprocesamiento y Reducción Dimensional (PCA)**:
    - Antes de aplicar el clustering, se preprocesaron los datos utilizando técnicas de normalización y transformación. Se aplicó PCA para reducir las dimensiones del dataset, seleccionando las primeras 78 componentes principales, que explican la mayor varianza.
2. **Aplicación del Algoritmo Jerárquico**:
    - Se utilizó la técnica de clustering aglomerativo con el método de enlace de Ward y la distancia euclidiana. Esta técnica comienza agrupando los datos más similares y va formando clusters más grandes hasta que todos los datos se encuentran en un solo grupo.
3. **Visualización mediante Dendrograma**:
    - Se generó un dendrograma para visualizar la jerarquía de los agrupamientos. A partir de este dendrograma, se seleccionó un punto de corte (distancia de 120) que permitió dividir los datos en dos grupos grandes.
4. **Interpretación de Resultados**:
    - Se obtuvieron dos grandes clusters, lo que nos permitió analizar cómo los datos se agrupan de manera natural y observar las relaciones entre individuos. El dendrograma nos dio una visión clara de la estructura jerárquica entre los datos.

Implementación de K-Means:

1. **Preprocesamiento y Reducción Dimensional (PCA)**:
    - Al igual que en el clustering jerárquico, los datos fueron preprocesados y reducidos dimensionalmente mediante PCA, seleccionando nuevamente las primeras 78 componentes.
2. **Selección del Número de Clusters (Codo y Silhouette)**:
    - Se utilizó el Método del Codo y el Índice de Silhouette para determinar el número óptimo de clusters. En este paso, se iteró entre 2 y 6 clusters, evaluando la inercia (varianza intra-cluster) y el índice de Silhouette para medir la cohesión de los grupos.
    - Estos métodos ayudaron a identificar que una configuración de 3 o 4 clusters era adecuada para los datos, ya que mostraban una mejor separación y compacidad.
3. **Aplicación de K-Means**:
    - Con el número de clusters determinado, se aplicó el algoritmo K-Means con 3 y 4 clusters. K-Means agrupó los datos en clusters homogéneos basados en las características seleccionadas, lo que permitió una clasificación clara de los individuos en función de sus patrones de comportamiento.
4. **Visualización y Evaluación**:
    - Se visualizó la agrupación de los clusters en un gráfico bidimensional utilizando las dos primeras componentes principales de PCA. Esto permitió una representación gráfica de los clusters formados por K-Means.
    - Además, se evaluó el rendimiento de los clusters utilizando el índice de Silhouette, para asegurarnos de que los grupos eran bien diferenciados.
5. **Análisis de las Estadísticas de los Clusters**:
    - Finalmente, se obtuvieron estadísticas descriptivas para cada cluster, como la edad promedio y el consumo de sustancias, lo que permitió una interpretación detallada de

las diferencias entre los grupos. Estas estadísticas ayudaron a definir características específicas para cada cluster, destacando los patrones de consumo de sustancias y las características demográficas de cada grupo.

Justificación de la elección del algoritmo

Como se menciona anteriormente para abordar el problema de la población en situación de calle en Colombia, se han seleccionado las siguientes técnicas de aprendizaje no supervisado: el Análisis de Componentes Principales (PCA), el algoritmo de clustering K-means y jerárquico aglomerativo.

En cuanto a la reducción de dimensionalidad con PCA, el conjunto de datos contiene 112 variables lo que dificulta la identificación de patrones significativos, para esto el PCA permite reducir la dimensionalidad de los datos, preservando la mayor parte de la variabilidad y eliminando el ruido de las variables menos relevantes, facilitando el enfoque en los factores que realmente afectan a la población en situación de calle, aunque este algoritmo tiene limitaciones a la hora de interpretar los resultados y puede que exista el riesgo de perder información relevante al reducir la dimensionalidad.

Para la segmentación de los datos, el algoritmo de K-means es el adecuado, ya que este es conocido por su simplicidad y eficiencia computacional, lo que lo hace adecuado para manejar grandes volúmenes de datos como en este caso, este permite realizar ejercicios de clustering, agrupando los datos en clusters homogéneos basados en características similares, lo que es crucial para clasificar a la población en grupos con necesidades y características comunes. Asimismo, es útil para analizar segmentos jerárquicamente relacionados, lo que puede ayudar a la formulación de políticas y programas de intervención más específicos.

Conclusiones

Los análisis realizados con el Análisis de Componentes Principales (PCA) y técnicas de clustering sobre los datos del Censo de Habitantes de Calle 2021 en Colombia revelan que el inicio temprano en el consumo de sustancias como el basuco, la marihuana y el cigarrillo está estrechamente ligado a una vida prolongada en la calle. Estos hallazgos resaltan la importancia de las intervenciones preventivas que deberían dirigirse a jóvenes en riesgo, con el objetivo de reducir la incidencia del inicio temprano en el consumo de sustancias y su transición hacia la habitabilidad prolongada en la calle.

Además, se identifica la necesidad de políticas diferenciadas que aborden las diversas causas y condiciones que mantienen a las personas en la calle, especialmente aquellas relacionadas con la dependencia a sustancias adictivas. Los programas de rehabilitación y las estrategias de prevención deben ser multidimensionales, ofreciendo soporte médico, psicológico y social para facilitar la reintegración de estas personas a la sociedad. Mejorar la recolección y análisis de datos futuros fortalecerá la base para políticas públicas más efectivas, permitiendo intervenciones más precisas y fundamentadas que puedan impactar positivamente en la reducción de la población en situación de calle en Colombia.

Este trabajo demuestra que el uso de metodologías como el Análisis de Componentes Principales (PCA) y las técnicas de clustering no solo simplifica la complejidad de grandes conjuntos de datos, sino que también facilita la identificación de patrones relevantes en problemas de la vida real. En este caso, se ha logrado reducir el número de variables sin perder la esencia de la información, permitiendo crear agrupaciones que resaltan aspectos críticos de la población en situación de calle.

Repositorio en Github:https://github.com/Dsharlie/PROYECTO-ANP-GRUPO-27/tree/main

Bibliografía

- Ngo, A. N., & Turbow, D. J. (2019). Principal Component Analysis of Morbidity and Mortality among the United States Homeless Population: A Systematic Review and Meta-Analysis.

International Archives of Public Health and Community Medicine, 3(2). https://doi.org/10.23937/2643-4512/1710025

- Ester, M., H. P. Kriegel, J. Sander, and X. Xu. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, pp. 226–231.
- Adelante-I. (s.f.). Red Calle: Desarrollo de políticas de atención a las personas en situación de calle [Archivo Video]. Recuperado de: https://www.adelante-i.eu/red-calle#descr
- Lasode, M. K. (2019). The impact of homelessness in social vulnerability assessment: A case study of Austin, Texas. (Master's thesis, Texas State University). https://digital.library.txst.edu/items/de621cc0-11fc-4f2c-ab99-7049e1c339b9
- Ministerio de Salud y Protección Social. (2022, 30 de julio). Boletín de prensa No 410 de 2022: Declaración del ministro de Salud y Protección Social, Fernando Ruiz Gómez [Comunicado de prensa]. Recuperado de https://www.minsalud.gov.co/Paginas/Gobierno-Nacional-presento-Politica-Publica-Social-para-Habitantes-de-Calle-.aspx
- Departamento Administrativo Nacional de Estadística (DANE). (2021). Censo de Habitantes de Calle 2021: Diccionario de Datos. Recuperado de https://microdatos.dane.gov.co/index.php/catalog/720/data-dictionary/F3?file_name=CHC_2021