**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 Departamento Nacional de Estadística (DANE) 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 proyecto busca dar respuesta a esa cuestión aplicando métodos de aprendizaje no supervisado, 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. 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.

https://www.adelante-i.eu/red-calle#descr

**Revisión preliminar de la literatura**

Se realizó una búsqueda de estudios donde se hubiera utilizado PCA en el contexto de la población habitante de calle, encontrando los siguientes:

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

En el caso del primer estudio utilizaron PCA para identificar las causas principales de morbimortalidad para la población habitante de calle de Estados Unidos, analizando múltiples estudios (papers) previamente publicados. El estudio encontró correlaciones entre comorbilidades como enfermedades cardiacas, cáncer, diabetes y el uso de sustancias psicoactivas. A pesar de usar una aproximación similar al presente proyecto (PCA), el contexto del primer estudio difiere en que el análisis se enfocó en revisar estudios previos para construir un metaanálisis mientras que para este caso se plantea usar datos obtenidos a través del CHC 2021 directamente, además de la diferencia en la ubicación de la población (Estados Unidos y Colombia).

1. THE IMPACT OF HOMELESSNESS IN SOCIAL VULNERABILITY ASSESSMENT: A CASE STUDY OF AUSTIN, TEXAS - Lasode, M. K. (2019)

En cuanto al segundo estudio, éste se centra en establecer un índice de vulnerabilidad a partir de diferentes variables demográficas y socioeconómicas obtenidas de datos de censo en Austin-Texas, para lo cual se utilizó PCA obteniendo agrupaciones de las variables en los componentes principales. Los componentes fueron utilizados para evaluar la vulnerabilidad social, destacando la importancia de incluir la identificación de la población habitante de calle en la evaluación a la hora de calcular el índice. Teniendo en cuenta esta aproximación y la que se plantea en el presente estudio, se encuentra una diferencia en el uso deseado para los componentes obtenidos por PCA, ya que para el estudio mencionado los componentes se utilizaron como insumo para el modelo de vulnerabilidad mientras que para nuestro caso se desea utilizarlos para identificar patrones y factores directamente de los datos del censo, sin llegar a emplearlos en un modelo más amplio.

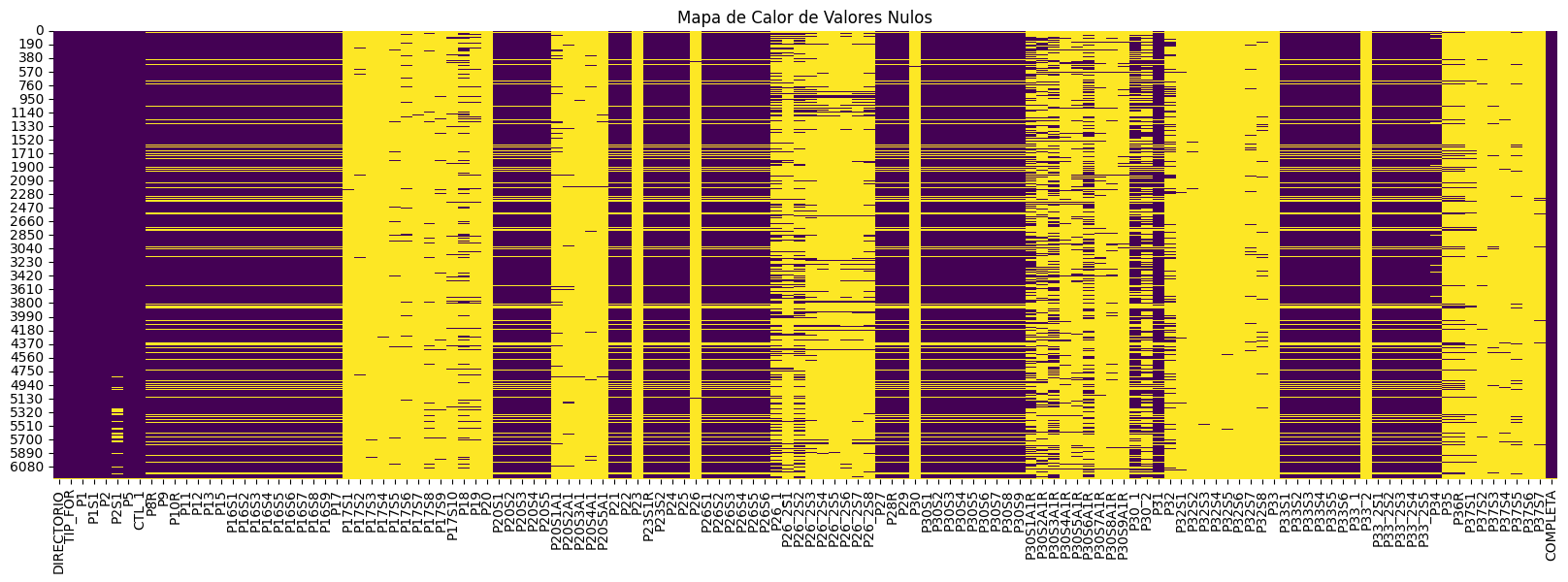
**Descripción de los datos**

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.

Para revisar la cantidad de información disponible para cada variable se utiliza un gráfico de calor en donde el color amarillo representa la cantidad de valores nulos y el morado, no nulos.



Cómo se puede observar en la imagen, hay variables que cuentan con una gran cantidad de valores nulos como las orientadas a conocer si el encuestado(a):

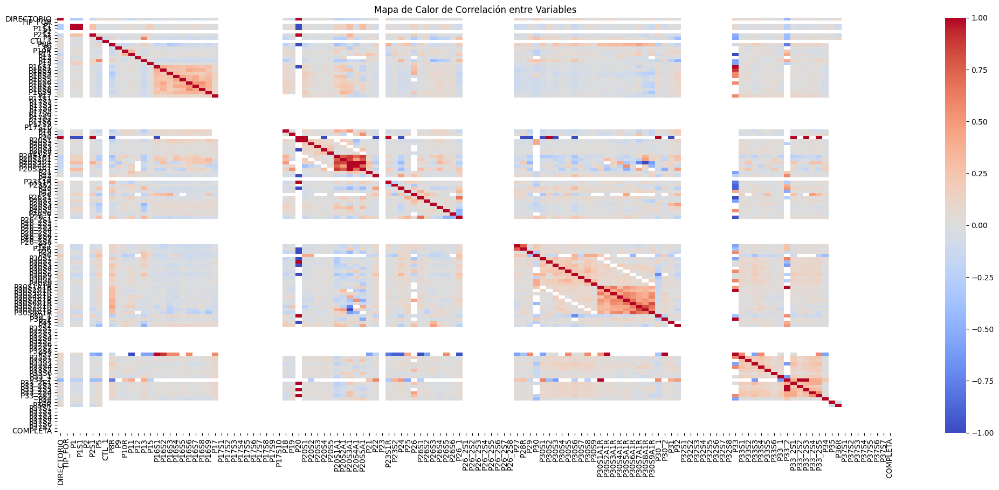
* Ha tenido alguna enfermedad, accidente o problema de salud.
* Ha sido diagnosticado con alguna enfermedad.
* Consume una sustancia psicoactiva específica.
* Conoce de programas de atención para habitantes de calle.

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, con base en el mapa de calor que se presenta a continuación, se pueden observar 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?](https://microdatos.dane.gov.co/index.php/catalog/720/variable/F3/V279?name=P17S1)
* ¿[Cuál(es):](https://microdatos.dane.gov.co/index.php/catalog/720/variable/F3/V279?name=P17S1) Lesión intencional por parte de terceros?
* [¿Cuál(es): Problema mental o emocional?, etc.](https://microdatos.dane.gov.co/index.php/catalog/720/variable/F3/V281?name=P17S3)



Se realizan algunos gráficos de variables para observar su comportamiento:

|  |  |
| --- | --- |
|  | Gráfico, Gráfico de barras  Descripción generada automáticamente |

Con base en las gráficas y 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.

**Propuesta metodológica:**

Para este proyecto se utilizarán dos técnicas de aprendizaje no supervisado: el Análisis de Componentes Principales (PCA) y el algoritmo de K-means. PCA se utilizará para la reducción de dimensionalidad de los datos, mientras que K-means se aplicará posteriormente para la segmentación de la población estudiada en grupos homogéneos. Esta combinación permitirá no solo identificar los factores clave que afectan a la población sin hogar en Colombia, sino también clasificar a las personas en grupos según características similares, lo cual puede ser valioso para la formulación de políticas públicas y programas de intervención social

El uso combinado de PCA y K-means es adecuado debido a la naturaleza compleja y multidimensional del conjunto de datos del Censo de Habitantes de Calle. PCA ayudará a reducir la cantidad de variables a un conjunto más manejable de componentes principales, preservando la mayor parte de la variabilidad en los datos. Esto simplificará el análisis posterior y reducirá el ruido de datos menos relevantes. Luego, K-means se utilizará para agrupar a la población en segmentos distintos basados en los componentes principales, facilitando la identificación de patrones dentro de la población.

Se propone seguir las siguientes etapas en el desarrollo de la metodología:

* Preparación de Datos: La base de datos será revisada para manejar datos faltantes y realizar las transformaciones necesarias para adecuar los datos a un formato utilizable para PCA y K-means. Esto incluye la estandarización de las variables para garantizar que cada una contribuya equitativamente al análisis.
* Aplicación de PCA: Utilizando Python y bibliotecas como scikit-learn, se calcularán los componentes principales del conjunto de datos. Se utilizarán gráficos de sedimentación (scree plots) y el criterio de Kaiser para determinar el número óptimo de componentes a retener.
* Segmentación con K-means: Los componentes principales resultantes de PCA serán utilizados como input para el algoritmo K-means. El número de clusters será determinado utilizando métodos como el codo y la silueta (silhouette score) para garantizar una segmentación efectiva.
* Interpretación de Resultados: Los resultados de PCA y K-means serán interpretados para identificar patrones clave en la población de estudio. Esto ayudará a comprender mejor las características y necesidades de los diferentes segmentos de la población sin hogar.
* Validación y Ajuste del Modelo: Los resultados iniciales serán revisados y, si es necesario, se realizarán ajustes en los modelos de PCA y K-means para mejorar su precisión.

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>
* 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>