CARACTERIZACIÓN DE ESTUDIANTES DESPLAZADOS CON CLUSTERING

Caracterización de estudiantes en situación de desplazamiento en Santander a través de técnicas de clustering.

Leonardo Alquichire Alquichire

Trabajo para optar por el título de ingeniero industrial

Director
M.Sc. Edgar Eduardo Córdoba Sarmiento

Codirector
M.Sc Dilan Jhoanny Mogollon Carreño

Universidad Industrial de Santander
Facultad de Ingenierías Fisicomecánicas
Escuela de Estudios Industriales y Empresariales
Bucaramanga
2024

Dedicatoria

A Dios, porque sin Él, toda esta historia aún sería un sueño sin realizar, por brindarme la sabiduría, fuerza y disciplina necesaria para culminar esta etapa tan importante.

A mi familia, por ser un pilar y por brindarme un brazo en donde descansar siempre que me encontraba desmotivado, pero también por poner mis pies sobre la tierra cuando todo marchaba bien.

A mis amigos, los ingenieros del balón pie, porque jugaron un papel importante brindándome consejos, risas y esparcimiento, todo en el momento indicado.

A mis amigas, por brindarme compañía, por permitirme escuchar y ser escuchado, por hacer más amena mi etapa de foráneo primíparo.

Agradecimientos

A la universidad industrial de Santander, por acogerme en sus instalaciones, por educarme como profesional y como persona, por permitirme retribuirle a la sociedad lo que tanto me ha dado.

A los profesores que han contribuido en mi formación, enseñando con vocación, despertando interés en sus áreas.

A la ESSA y la sociedad San Vicente de Paul, y su programa de becas: buena energía para tu proyecto de vida. Fueron de gran ayuda para cumplir esta meta.

A mis mejores amigos: Vladimir y Diana. Por todo y más, porque en este punto, las palabras se quedan cortas.

Tabla de contenido

Lista de tablas

Lista de figuras

Lista de apéndices

Introducción

El desplazamiento forzado en Colombia, una consecuencia directa del conflicto armado interno ha impactado profundamente el acceso a la educación de los niños, quienes conforman una de las poblaciones vulnerables en este contexto. Según ACNUR (2016), Colombia registra la mayor cantidad de personas desplazadas contra su voluntad en el mundo, con cifras que superaron los 6.5 millones en 2016. Este fenómeno no solo expulsa a las familias de sus territorios, sino que también vulnera derechos fundamentales, como la educación, generando exclusión y dificultades de adaptación para los menores en los contextos escolares (Guerra Lozano et al., 2021).

La exclusión educativa que enfrentan los niños desplazados no se limita a su condición económica o social, sino que también responde a barreras culturales y emocionales que dificultan su integración. En muchos casos, deben enfrentarse a discriminación, desigualdades en el acceso a recursos educativos y retos asociados a la adaptación a sistemas educativos ajenos a sus experiencias previas. Como señala Hernández Burgos & Murillo Estepa (2015) estos niños enfrentan desafíos de aceptación que deben soportar durante los primeros meses, además de un nivel académico superior y diferente al que conocían.

En este estudio, se propone abordar esta problemática desde una perspectiva analítica, utilizando técnicas de clustering para caracterizar a los estudiantes desplazados en el departamento de Santander utilizando una base de datos pública conseguida en la página de datos abiertos del gobierno de Colombia. El objetivo es identificar patrones y características clave de esta población que permitan generar información relevante para la planificación y toma de decisiones en los organismos educativos. Con ello, se busca promover estrategias más efectivas de inclusión y equidad educativa, fortaleciendo la integración y el desarrollo de estos estudiantes en el sistema escolar.

1. Planteamiento del problema

El desplazamiento forzado es uno de los fenómenos sociales más críticos en Colombia, resultado del prolongado conflicto armado que ha afectado a millones de personas. Santander no ha sido ajeno a esta problemática. Según el Registro Único de Víctimas (RUV), hasta el 31 de agosto de 2024 se han registrado 265.080 eventos victimizantes en el departamento, de los cuales el 81,84% corresponde al desplazamiento forzado. Este fenómeno genera barreras significativas para el acceso a derechos fundamentales, como la educación. Los niños desplazados enfrentan desafíos que afectan su desempeño académico, su integración social y, en general, su desarrollo integral.

A pesar de la gravedad del problema, existe una carencia de estudios específicos que analicen las características de esta población desde una perspectiva educativa. Actualmente, los esfuerzos para atender a los estudiantes en situación de desplazamiento se ven limitados por la falta de información detallada que permita segmentar y priorizar estrategias basadas en evidencia. Como señala Guerra Lozano et al. (2021), los estudiantes desplazados enfrentan exclusión y discriminación, además de dificultades para adaptarse a entornos educativos que no consideran sus condiciones particulares. Esta situación limita no solo su potencial académico, sino también su capacidad para superar las desigualdades que derivan del desplazamiento.

En este contexto, el presente proyecto busca segmentar y analizar a la población estudiantil en situación de desplazamiento en Santander mediante el uso de técnicas de clustering. Estas herramientas permiten identificar patrones y características clave en los datos, lo que a su vez facilita el diseño de políticas educativas más focalizadas y efectivas. Según Anuradha et al. (2015), "los algoritmos de clustering permiten clasificar a los estudiantes en grupos bien definidos para comprender sus comportamientos y estilos de aprendizaje, y así predecir su desempeño

académico" (p. 47). Este enfoque puede permitir en el futuro a las instituciones gubernamentales optimizar la asignación de recursos, establecer prioridades en las zonas más vulnerables y promover una mayor equidad en la educación.

La caracterización de esta población podría llegar a identificar variables latentes que contribuyan no solo a mejorar la planificación educativa, sino que también a sentar las bases para generar intervenciones que respondan directamente a las necesidades de los estudiantes desplazados. De esta forma, se espera lograr un impacto positivo en términos de inclusión, garantizando que todos los niños, sin importar su condición, puedan acceder a una educación de calidad y desarrollarse plenamente en un entorno que promueva la equidad.

2. Objetivos

2.1 Objetivo general

Caracterizar la población estudiantil en situación de desplazamiento en Santander mediante técnicas de clustering, con el propósito de generar información que fortalezca la planificación y toma de decisiones en los organismos educativos, promoviendo la equidad e inclusión educativa.

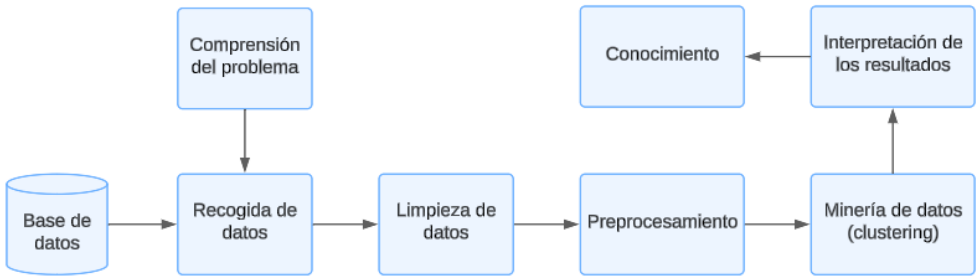
2.2 Objetivos específicos

- Seleccionar la técnica de clustering más adecuada para la base de datos de estudiantes en situación de desplazamiento en Santander a través de una revisión de literatura.
- Aplicar la técnica de clustering seleccionada a la base de datos segmentando los estudiantes en situación de desplazamiento en Santander.
- Analizar los resultados obtenidos del clustering para identificar características y tendencias dentro de la población estudiantil.
- Desarrollar un artículo científico de carácter publicable que presente los hallazgos y conclusiones derivadas del estudio realizado.

3. Metodología

La metodología KDD está asociada al machine learning o aprendizaje automático, es un proceso cuyo objetivo es la extracción de conocimiento de las bases de datos. Este proceso inicia desde la comprensión del problema, pasando por la recolección de los datos para proceder con su limpieza y posterior preprocesamiento, luego se aplica la minería de datos, que, en este caso, está asociada al clustering específicamente, para terminar con el análisis e interpretación de los resultados y generación del conocimiento (Charte, 2020). Teniendo en cuenta lo dicho, las fases se pueden ver de manera gráfica en la Figura 1.

Figura 1
Fases de la metodología KDD.



Nota: Adaptado de Charte (2020).

3.1 Comprensión del problema

En este primer paso, lo que se realiza es la comprensión del contexto de la problemática, identificación de los objetivos y búsqueda de la base de datos sobre la que se trabajará. (Charte, 2020).

Para este proyecto, la base de datos que será utilizada fue obtenida de la plataforma nacional de datos abiertos de Colombia (datos.gov.co) y tiene como nombre “ESTUDIANTES EN SITUACIÓN DE DESPLAZAMIENTO EN SANTANDER”, la cual a fecha de elaboración de

este documento tiene su última actualización el 29 de julio de 2024 (Gobernación de Santander, 2022).

Es una base de datos con alrededor de 18100 registros y 22 campos en donde se pueden encontrar variables de tipo categóricas (como tipo de desplazamiento, etnia o discapacidad), discretas (como la edad o el grado escolar) y binarias (como el género, visto desde un punto de vista cuantitativo).

Además, resulta necesario entender el panorama de las técnicas utilizadas en problemas similares por medio de una revisión de literatura. Esta, debe ser enfocada a poblaciones vulnerables o que presenten una situación de manera que la caracterización de las personas pueda ser una posible solución, o una herramienta que facilite el camino hacia ella. Es importante que la revisión sea dirigida hacia investigaciones que hayan usado clustering como técnica de minería de datos, para de esta forma listar los algoritmos que tienden a ser utilizados y poder comparar con respecto a las características de la base de datos que se utilizará, cuál técnica puede llegar a ser más efectiva para el presente trabajo, y de esta forma, seleccionarla.

3.2 Limpieza de datos

En el proceso de recolección de datos, así como en su codificación y transmisión es común que se cometan errores. Estos, a su vez, se pueden presentar de dos tipos. El primero está relacionado a que los datos son erróneos, por lo que proporcionarán ruido al análisis. Por otro lado, el segundo es la pérdida de algunos datos, por lo que dejaría vacíos en el conjunto de datos.

Esto, se resuelve mediante la eliminación o consideración del ruido dependiendo de la técnica de clustering utilizada o la imputación de datos según sea el caso (Charte, 2020).

3.3 Preprocesamiento

Dado que no todos los datos recogidos en la base de datos serán útiles, en esta fase se lleva a cabo el proceso de selección de variables relevantes. En esta fase se busca identificar las variables que aportarán al desarrollo del problema, eliminando aquellas que no lo hagan o que brinden información repetida y que por lo tanto entorpecen el proceso de extracción de conocimiento. Esta etapa se conoce como reducción de dimensionalidad (Charte, 2020).

Además, en esta fase también se realiza, en caso de ser necesario por el comportamiento de las variables originales, la transformación de los datos que tiene pasos como la normalización, escalado y discretización de estos, los cuales conseguirán que el algoritmo consiga una mejor efectividad, y por tanto brinde unos mejores resultados (Charte, 2020).

3.4 Desarrollo y aplicación del algoritmo de clustering

Fayyad et al. (1996) hacía referencia a esta etapa como aquella que busca coincidir los objetivos planteados al inicio con un método de minería de datos. En este caso, un algoritmo de clustering.

Este último, se evalúa a través de una revisión de literatura en donde se listaron las diferentes técnicas que otros autores utilizaron en sus estudios, luego, en un espacio de discusión se plantea utilizar la técnica DBSCAN o K-means según la cantidad de ruido que se presente en un análisis posterior, debido a que estas técnicas se relacionan con las características (tamaño y tipos de variables) de la base de datos que se utilizará en el presente trabajo.

3.5 Interpretación de los resultados

En esta fase se busca convertir los datos arrojados por el algoritmo en información, y por lo tanto en conocimiento. Se busca clasificar y caracterizar los clusters de manera que sea posible extraer insights útiles para el cumplimiento del objetivo planteado.

4. Marco de referencia

4.1 Marco de antecedentes

En la Universidad Industrial de Santander se encuentran tres antecedentes que pueden ser útiles para el desarrollo de este trabajo en el ámbito técnico de la minería de datos, para ser más específicos, se desarrollaron en la escuela de estudios industriales y empresariales y se listan a continuación. También se encuentra un antecedente en la Universidad Santo Tomás que permite entender el contexto de los estudiantes en condición de desplazamiento, lo cual será clave para realizar posteriormente el análisis de los resultados.

Sanabria Ruiz et al. (2017) busca por medio de la red social X (antes twitter) entender el comportamiento de las personas que realizan los tweets mientras exista una tendencia vigente, de esta manera, generar información para poder predecir futuros patrones y/o tendencias. Para cumplir esto, los autores resaltan la diferencia entre clasificación y clustering de texto, en el que resaltan la importancia de la limpieza de los datos. Durante el análisis de datos, se utilizó el método Elbow para obtener el número de clusters óptimo para el desarrollo de la técnica k-means. Con esto, se logra la clasificación de los datos en cuatro clusters y la generación de información en función del cumplimiento de los objetivos. Aunque la diferencia de eficacia de la técnica no es significativa, la presencia de ruido en los tweets estudiados impidió trabajar con comodidad, por lo que es importante considerarlo en la presente investigación.

Muñoz Osorio et al. (2013) utilizó técnicas de minería de datos para realizar un estudio de seguimiento a los egresados de la Universidad Industrial de Santander. Para su puesta en marcha utilizó la metodología KDD, la cual consiste en un paso a paso para la extracción de conocimiento en bases de datos. Si bien el enfoque de este proyecto fue más general hacia la minería de datos,

tuvo un componente de análisis de clusters que sumado a su metodología hace que sirva como referencia para el trabajo actual.

Benavidez Robles et al. (2022) indagó en el problema de escasez de contratación de mano de obra calificada para el subsector de la industria de calzado, cuero y marroquinería. Para ello, por medio de técnicas de clustering como k-means, busca generar información que expliquen patrones o tendencias con el objetivo de mejorar la toma de decisiones en este. Tras la identificación de variables, el siguiente paso que los investigadores tomaron fue el preprocesamiento de datos, que implicó la correlación de las variables, y aquellas cuyos resultados fueron cercanos a uno, se consideran redundantes, puesto que explicaban un mismo comportamiento. En ese caso, fue eliminada una de cada par correlacionada para optimizar el procesamiento del algoritmo. Por medio de los resultados, las empresas se clasificaron en dos clusters, el primer grupo son aquellas más productivas y competitivas mientras que el segundo grupo es lo contrario. Los investigadores destacan que el segundo grupo se presenta debido a la falta de planeación en la demanda, que los lleva a costos elevados en materia prima.

Por último, Guerra Lozano et al. (2021) proporciona un contexto de la problemática existente en la educación de los estudiantes en condición de desplazamiento. En esta revisión documental se analiza la importancia de una educación inclusiva para intentar disminuir la brecha social y anímica que presentan estas víctimas en diferentes aspectos de su vida. Los autores separan la información encontrada en tres categorías, desplazamiento forzado, inclusión y desarrollo íntegro durante la infancia. Si bien su enfoque es cualitativo, es importante para conocer ciertas características clave para el estudio y entendimiento de esta población.

4.2 Marco teórico

El análisis de datos, en sus múltiples formas, busca responder una pregunta esencial: ¿qué patrones subyacentes se esconden en el caos de la información? En el contexto de grandes volúmenes de datos, esta tarea resulta cada vez más compleja. La minería de datos surge como una alternativa indispensable para abordar este desafío, definiéndose como el proceso de utilizar técnicas avanzadas, como machine learning y análisis estadístico, para descubrir información valiosa en conjuntos masivos de datos (Holdsworth, 2024). Entre las múltiples técnicas que ofrece, el clustering destaca como una metodología particularmente poderosa cuando no se dispone de información previa para clasificar los datos.

El análisis de datos moderno se basa en diversas metodologías de machine learning, que pueden agruparse en dos categorías principales: el aprendizaje supervisado y el no supervisado. En el aprendizaje supervisado, se trabajan con datos etiquetados, donde cada instancia del conjunto de datos está asociada a una salida conocida o deseada (IBM, 2020). Este enfoque es común en tareas como clasificación o regresión, donde el objetivo es construir modelos predictivos basados en ejemplos previamente observados.

Por otro lado, el aprendizaje no supervisado se ocupa de datos no etiquetados, donde no existe información previa sobre las relaciones o categorías de los datos. Aquí, el objetivo no es predecir un resultado conocido, sino descubrir patrones, estructuras o agrupamientos ocultos en los datos (IBM, 2020). Dentro de esta categoría, el clustering se posiciona como una técnica clave, capaz de segmentar datos en grupos con características similares, sin necesidad de supervisión previa.

También existen herramientas como PCA (o por sus siglas en español, análisis de componentes principales) útiles en contextos de machine learning. Es una técnica estadística utilizada para la reducción de la dimensionalidad en conjuntos de datos con gran número de variables. Su objetivo es transformar las variables originales en variables más pequeñas tratando siempre de explicar la mayor varianza posible (IBM, 2020). En investigaciones de clustering no solo se utiliza para reducir la estructura de los datos, sino también para graficar en dos o tres dimensiones los conjuntos que contienen más, y así poder tener una representación gráfica y preveer el posible agrupamiento de la población por analizar.

El clustering, también conocido como análisis de conglomerados, se centra en la agrupación de datos en subconjuntos homogéneos, denominados clusters. Estos clusters están diseñados para que las instancias dentro de un mismo grupo sean similares entre sí, pero significativamente diferentes de las pertenecientes a otros grupos. Este proceso resulta invaluable en aplicaciones prácticas, como la segmentación de clientes en marketing, la detección de anomalías en sistemas de seguridad, o incluso la identificación de comunidades en redes sociales. Sin embargo, su éxito depende del algoritmo utilizado, ya que la naturaleza de los datos puede variar drásticamente, desde conjuntos bien definidos hasta estructuras complejas y ruidosas.

Tradicionalmente, algoritmos como K-Means han dominado el panorama del clustering. Estos métodos particionales dividen los datos en un número predefinido de grupos basándose en su proximidad a un conjunto de centroides. Aunque efectivos en ciertas circunstancias, presentan limitaciones importantes: requieren que se conozca de antemano el número de clusters, asumen que estos tienen formas esféricas y son sensibles al ruido y a los valores atípicos. En respuesta a estas limitaciones, surgieron los métodos basados en densidad, diseñados para abordar conjuntos de datos con estructuras más complejas y resistencia al ruido.

En este contexto, Ester et al. (1996) formula un nuevo algoritmo: DBSCAN (o en español, agrupamiento espacial basado en densidad de aplicaciones con ruido). Este algoritmo redefine lo que significa un cluster. Para DBSCAN, un cluster no es un grupo de puntos cercanos entre sí; es una región densa en el espacio de datos que se expande progresivamente a partir de puntos núcleo, incorporando aquellos que cumplen criterios de densidad definidos por dos parámetros: eps (el radio de vecindad) y minPts (el número mínimo de puntos necesarios para formar un cluster). Los puntos que no pertenecen a ninguna región densa son etiquetados como ruido, una característica que lo distingue de métodos tradicionales incapaces de manejar datos ruidosos de manera efectiva.

El verdadero poder de DBSCAN radica en su flexibilidad. A diferencia de K-Means, no requiere que se especifique el número de clusters de antemano. Además, es capaz de identificar clusters de formas arbitrarias, lo que lo hace ideal para escenarios donde los datos no se ajustan a patrones geométricos simples. Por ejemplo, en análisis geoespacial, DBSCAN puede detectar regiones de alta densidad de eventos en mapas sin asumir que estas regiones tienen formas circulares o rectangulares. Este enfoque ha llevado al algoritmo a ser ampliamente utilizado en disciplinas que van desde la biología computacional hasta la detección de fraudes financieros. Al priorizar la densidad sobre las formas predeterminadas y al reconocer la importancia de identificar ruido y estructuras atípicas, este algoritmo refleja la naturaleza intrínsecamente desordenada de los datos reales.

5. Revisión de literatura

5.1 Análisis bibliométrico

Para garantizar una cobertura integral de la literatura relevante, se diseñó una ecuación de búsqueda basada en palabras clave estratégicas relacionadas con el tema de investigación. Estas

palabras clave se seleccionaron a partir de la consideración de tres ejes principales: clustering, poblaciones vulnerables y países en vía de desarrollo, utilizando operadores booleanos para homogenizar la búsqueda.

Para el primer eje, se consideró buscar palabras claves como “clustering” o “cluster analysis" debido a la relevancia metodológica para el presente proyecto, ya que esta técnica permitirá identificar patrones ocultos y segmentar datos en grupos homogéneos los estudiantes en condición de desplazamiento en Santander.

El siguiente eje son las poblaciones vulnerables, en el que se usaron palabras como “vulnerable population” o “educational data” unido por un operador de inclusión “and” a palabras como “child*” o “school*”. Debido a que este proyecto busca caracterizar una población vulnerable, se buscan poblaciones similares en la revisión de literatura puesto que suelen compartir variables, y de esta manera se podrá estudiar como en función de las variables cada autor utiliza una técnica u otra.

De forma similar el tercer eje propuesto, países en vía de desarrollo; utilizando palabras como “Colombia”, “latin America” o “developing country”. Se hizo importante añadir ese último eje en la ecuación de búsqueda ya que así se busca que las investigaciones resultantes compartan un contexto similar. Estos países enfrentan desafíos específicos, como desigualdades económicas y brechas sociales en general, permitiendo que los resultados reflejen condiciones similares al caso de estudio.

Además, se buscaba que todos los resultados estuvieran asociados con estudios en humanos, por lo que se añadió un operador AND para que esta condición se cumpla. Y, por último, en la ecuación de búsqueda se limita el tiempo de estudio para artículos publicados desde del año

2000, debido al periodo de tiempo en que la base de datos pública recogió datos y con el fin de obtener información actualizada.

La búsqueda se realizó en la base de datos SCOPUS debido a su alto impacto de sus resultados y calidad de sus autores, revistas e información en general. La ecuación de búsqueda quedó de la siguiente manera:

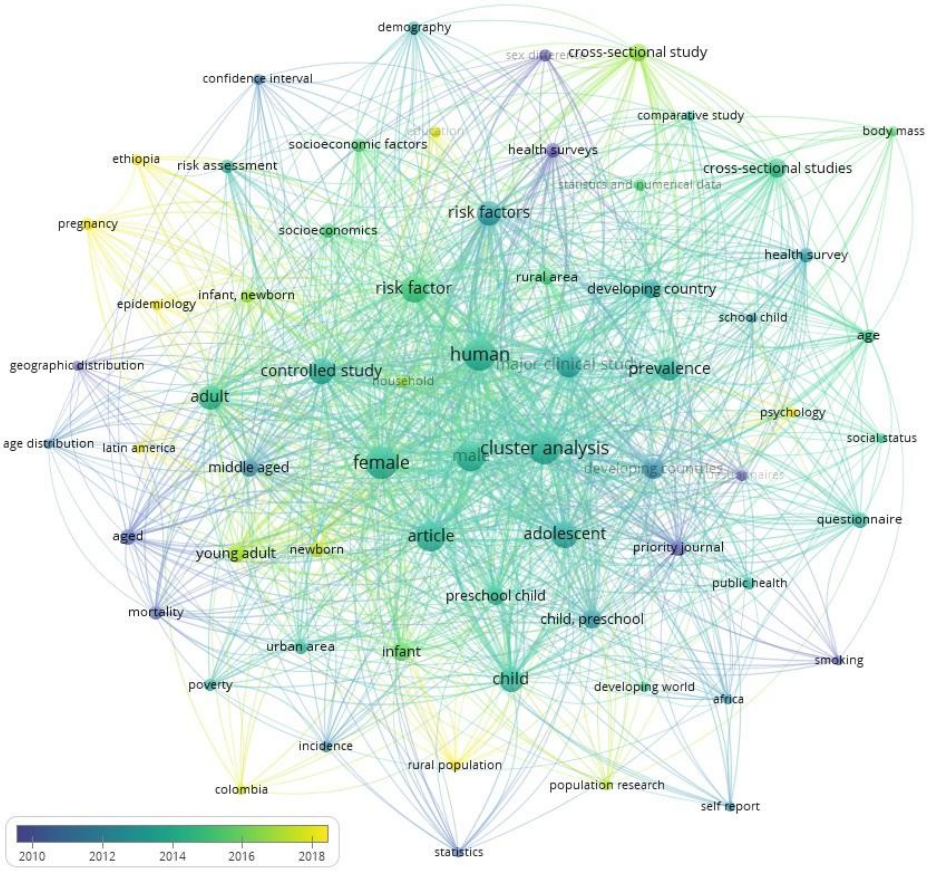
TITLE-ABS-KEY ((clustering OR "cluster analysis") AND ("vulnerable population" OR "educational data" OR ("risk factor" AND (child* OR school*))) AND (Colombia OR "latin america" OR "developing country") AND (human*)) AND PUBYEAR > 1999 AND PUBYEAR < 2025

Como resultados, arrojó 117 documentos, a partir de los cuales se aplicaron distintos factores de exclusión. El primero consistió en descartar aquellos que no contenían la palabra clave “Clustering”, reduciendo el número de fuentes a 69. A continuación, se eliminaron 19 documentos adicionales debido a que su enfoque era biológico o clínico, no mencionaba la técnica de clustering utilizada o se trataba de análisis de literatura. Finalmente, se excluyeron 13 artículos, ya que, en su técnica de análisis de datos, solo se mencionaban métodos de análisis con estadísticas descriptivas o análisis de varianza, los cuales no son relevantes para los objetivos de esta investigación. De este modo, quedaron finalmente 37 documentos para su revisión final.

En la Figura 2 Análisis de concurrencia., se presenta el análisis de concurrencia realizado en la herramienta VOSviewer.

Figura 2

Análisis de concurrencia.



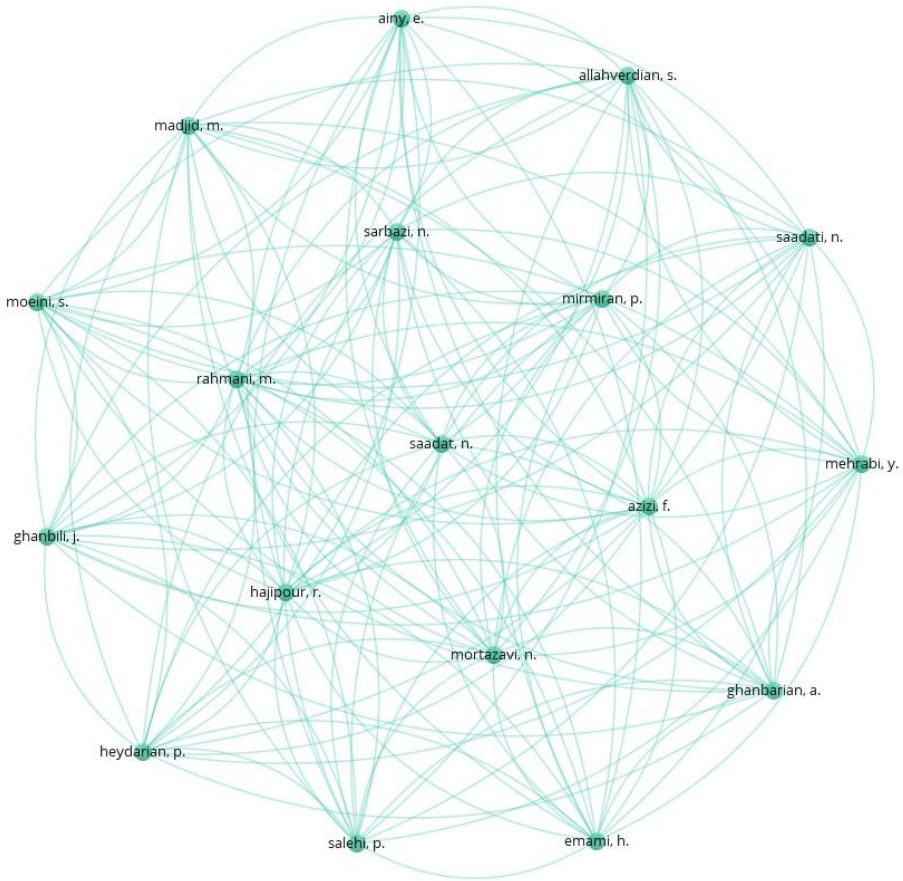
En los nodos centrales se encuentran palabras como “cluster analysis", “human”, “male”, “female”, “adolescent” y “developing countries”, lo que significa que el enfoque principal de los resultados se basa en análisis de conglomerados aplicados a poblaciones humanas en países en vía de desarrollo. Además, también aparecen nodos como “risk factor”, “socioeconomic”, “rural area", lo que indica que se están considerando factores de riesgo y socioeconómicos, los cuales se asocian directamente con situaciones de vulnerabilidad. La escala temporal muestra que temas asociados al análisis de cluster han cobrado relevancia desde 2018, nodos más recientes como “Colombia”, “latin america”, “psychology”, “education” y “rural population” muestran relación con los

anteriores, por lo que es correcto afirmar que viene incrementando un interés en el tema de estudio propuesto en la región de Latinoamérica.

Por otro lado, la Figura 3 Relación entre autores muestra la relación entre autores.

Figura 3

Relación entre autores

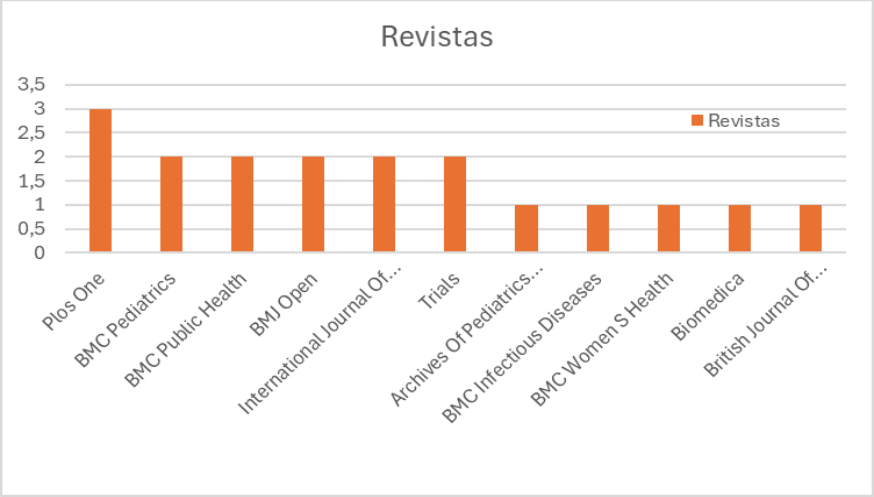


Esta figura generada por el programa VOSviewer muestra los autores que predominan en este tema, los links entre ellos pueden evidenciar una colaboración continua entre ellos, también, el tamaño de sus nodos refleja que este tema no está muy segmentado hacia un conjunto de investigadores en particular.

Además, la Figura 4 Revistas en donde fueron publicados los artículos. muestra el análisis de las revistas en donde fueron publicados los artículos.

Figura 4

Revistas en donde fueron publicados los artículos.

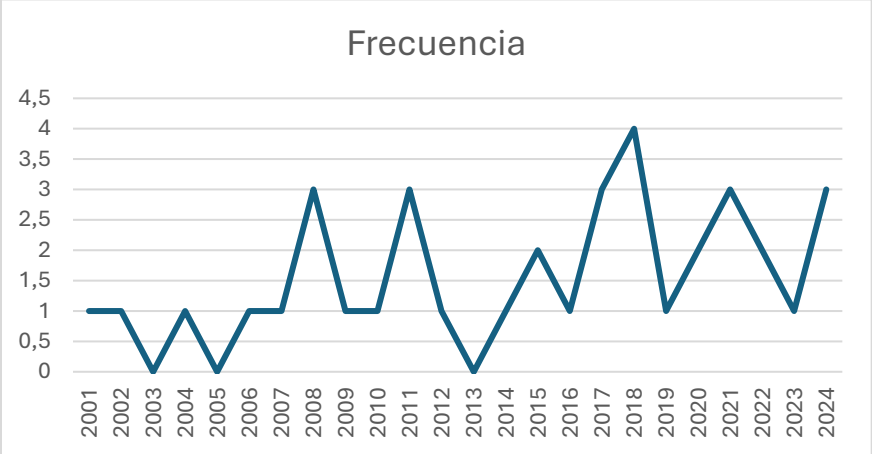


La figura muestra que no hay predominación en alguna revista, ya que la distribución de frecuencias es dispersa, esto refleja que se presenta una diversidad de enfoques y áreas de investigación que abordan el tema.

En adición, se presenta el análisis de frecuencia de publicaciones por año del conjunto analizado en la Figura 5 Análisis de frecuencia de publicaciones por año..

Figura 5

Análisis de frecuencia de publicaciones por año.

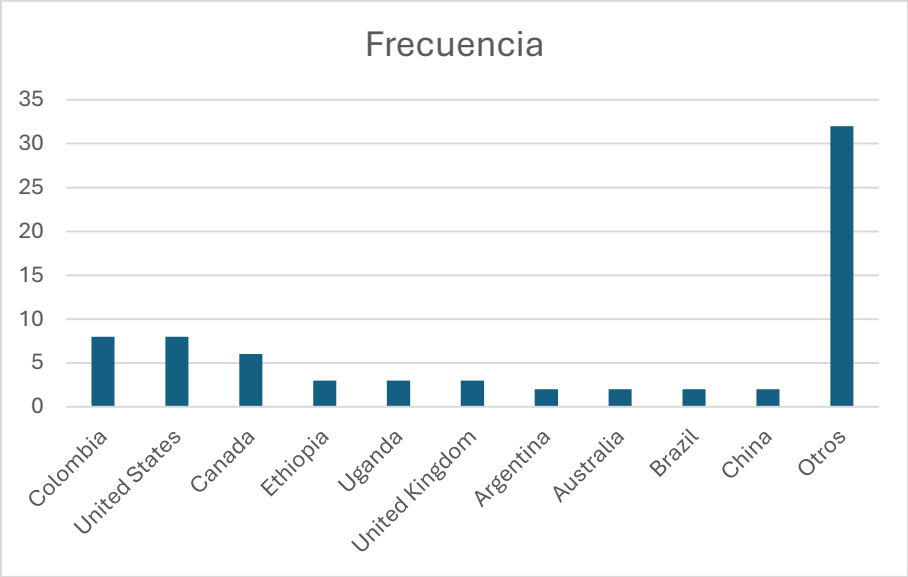


El análisis de frecuencia muestra que, si bien el tema no ha sido uno de análisis fuerte a lo largo del tiempo, se puede observar que ha venido incrementando el interés en el mismo en comparación con el inicio de década.

Por último, la Figura 6 Frecuencia por países muestra la frecuencia por país en donde se han presentado casos de estudio.

Figura 6

Frecuencia por países



Este análisis de frecuencia por países muestra aquellos en donde se han llevado a cabo las investigaciones, se puede observar que Colombia y Estados Unidos son los países en donde más estudios se han realizado, sin embargo, en la sección "otros" se encuentran muchos latinoamericanos y del continente africano, entendiendo que, en su mayoría, fueron analizados problemas encontrados en países en vía de desarrollo como se había planeado en la construcción de la ecuación de búsqueda.

5.2 Análisis de la literatura

En esta sección, se presentarán los resultados obtenidos de la ecuación de búsqueda descrita, presentando la problemática que llevó a la investigación de cada autor, resaltando las variables manejadas y el tamaño de la base de datos, según la información disponible. Además, se describirán la técnica utilizada y los resultados obtenidos, con el objetivo de tener un claro

panorama de los algoritmos utilizados en situaciones similares a la presente, y de esa manera determinar cuál puede ser la mejor opción para este estudio.

6.2.1 K-means

En primer lugar, Xu et al. (2024) analiza la vulnerabilidad social en áreas urbanas expuestas a desastres en el distrito de Hongshan en Wuhan, China. Los investigadores analizaron variables como el nivel educativo, el ingreso económico, la cobertura de seguridad social y factores relacionados con el entorno urbano, como la infraestructura comunitaria y las condiciones de vivienda. Estas variables permitieron medir tres dimensiones: exposición, sensibilidad y capacidad adaptativa.

En consecuencia, se empleó la técnica K-means, puesto que el investigador quería segmentar la población estudiada en tres clusters según su vulnerabilidad: alta, media y baja. Los resultados muestran que las comunidades con menos desarrollo corren un mayor riesgo, mientras que los habitantes con mejores condiciones de vivienda son menos vulnerables. Este método ha sido útil para agrupar la información según el nivel de vulnerabilidad en el que se encuentren, proporcionando así una base sólida para desarrollar políticas que permitan reducir los riesgos y satisfacer las necesidades de los grupos más desfavorecidos.

En contraste, Mebaraket al. (2024) analiza los hábitos de vida saludable en cuidadores de niños pertenecientes a comunidades vulnerables del Caribe colombiano, enfocándose en cómo estos cuidadores influyen en el desarrollo de hábitos saludables durante los primeros años de vida de los niños, particularmente en contextos de bajos recursos. El estudio utilizó una base de datos que incluyó a 544 cuidadores de niños entre 0 y 5 años.

Para abordar esta problemática, utilizó el algoritmo de K-means para identificar patrones en las actitudes de los cuidadores hacia los hábitos de vida saludable (HLH). Las variables se evaluaron mediante escalas de Likert de 5 puntos que midieron el nivel de acuerdo de los participantes con afirmaciones relacionadas con cada aspecto de los HLH. A partir de estas variables, el análisis permitió agrupar a los cuidadores en tres clusters principales, explicando el 73.53 % de la variabilidad de los datos. Este enfoque proporcionó una clasificación efectiva que puede orientar intervenciones específicas para promover mejores hábitos de vida en estas comunidades vulnerables.

En la Tabla 1 se pueden encontrar otros autores que también utilizaron k-means como técnica de agrupamiento para sus estudios.

Tabla 1
Otros autores que utilizaron k-means.

Autor	Artículo	Año	DOI
Strozzi et al.	Syndemic and syndemogenesis of low back pain in Latin-American population: a network and cluster analysis	2020	10.1007/s10067-020-05047-x
Dinya et al.	Profiles of suicidality and clusters of Hungarian adolescent outpatients suffering from suicidal behaviour	2009	10.1159/000228839
Ruger & Kim	Global health inequalities: An international comparison	2006	10.1136/jech.2005.041954

6.2.2 DBSCAN

Seguidamente Granados et al. (2023) examinan las inequidades de salud y la vulnerabilidad en poblaciones con enfermedades reumáticas y musculoesqueléticas (RMD) en América Latina, utilizando el enfoque syndemic. El objetivo es entender cómo factores epidemiológicos, biológicos, sociodemográficos y económicos interactúan para aumentar la vulnerabilidad en estas poblaciones, especialmente entre los grupos indígenas. El estudio incluyó a 44.560 individuos de cinco países latinoamericanos (Argentina, Colombia, Ecuador, México y Venezuela), de los cuales el 29.78% se identificaron como indígenas. Se recopilaron datos sobre variables como género, edad, escolaridad, tipo de sistema de salud, comorbilidades, discapacidad y estrés biomecánico articular, a través de cuestionarios.

El análisis se realizó en varias fases. Primero, se aplicaron análisis bivariados y regresión logística para explorar factores asociados con RMD. En la fase final, se utilizó DBSCAN (Agrupamiento espacial basado en densidad de aplicaciones con ruido) para realizar clustering y crear 20 clusters en la población indígena y 17 en la no indígena. Los resultados mostraron que las poblaciones indígenas enfrentan mayores niveles de dolor, discapacidad y comorbilidades, con acceso limitado a la atención médica. El análisis fue efectivo, ya que explicó el 73.53% de la variabilidad en los datos, identificando patrones de vulnerabilidad que permiten diseñar intervenciones más específicas para estos grupos.

6.2.3 Clustering jerárquico

A continuación, el estudio realizado por Chami et al. (2018) en Mayuge, Uganda, aborda la problemática de los desafíos en el diagnóstico y tratamiento de enfermedades en comunidades rurales y de bajos recursos, donde la presencia de comorbilidades complejas y sistemas de salud débiles dificultan la identificación y manejo adecuado de las enfermedades. La investigación se

centró en 16,357 individuos de 17 aldeas, que reportaron síntomas o enfermedades en los tres meses previos al estudio. Entre las variables analizadas se incluyen factores biosociales como el nivel socioeconómico, el acceso a agua potable, el saneamiento, la atención médica y la edad de los individuos. El estudio también consideró síntomas comunes, como fiebre y diarrea, que se asocian con altos índices de morbilidad y mortalidad en países de bajos ingresos.

Para el análisis de los datos, se emplearon técnicas de clustering jerárquico con el método de average linkage para agrupar los síntomas y enfermedades reportados, ponderando las relaciones entre los síntomas mediante riesgos relativos. Además, se aplicó modularidad y link clustering para identificar grupos de síntomas que podrían pertenecer a múltiples clusters. Los resultados mostraron que los individuos en hogares con menor nivel socioeconómico y peores condiciones de salud pública presentaron una mayor carga de comorbilidades, y que la fiebre fue un síntoma clave mal diagnosticado en muchos casos, especialmente como malaria. La técnica de clustering fue efectiva, ya que permitió identificar patrones importantes de comorbilidad y destacó la necesidad de estrategias de diagnóstico más precisas y personalizadas para las comunidades rurales, mejorando así las intervenciones de salud en contextos de bajos recursos.

En este contexto, Thörn et al.(2011) investiga las causas de la neumonía en niños de 1 a 35 meses en Goiania, Brasil, como parte del Latin America Epidemiological Assessment of Pneumococcus (LEAP study). Se analizaron datos de 11.521 niños que acudieron a emergencias con sospecha de neumonía entre mayo de 2007 y mayo de 2009, de los cuales 3.955 casos fueron confirmados radiológicamente. Las variables incluyeron indicadores socioeconómicos como ingresos familiares, nivel educativo de las madres, y acceso a servicios básicos. Los resultados mostraron que las tasas de incidencia eran significativamente más altas en áreas de bajos ingresos

en comparación con áreas de altos ingresos. Además, los hogares con menor nivel educativo y acceso limitado a servicios esenciales presentaron mayor incidencia de neumonía.

Para identificar patrones espaciales y agrupar áreas afectadas, se utilizó clustering espacial mediante el estadístico espacial de escaneo basado en Poisson. Este método detectó clusters significativos en las regiones occidental y sureste de la ciudad. El análisis incluyó simulaciones Monte Carlo para validar los clusters identificados y ajustar por población en riesgo. La técnica fue altamente efectiva al permitir identificar áreas de alta incidencia y sus factores asociados, proporcionando información clave para dirigir estrategias de vacunación y otras intervenciones en las zonas más desfavorecidas.

6.2.4 K-medoids

Por su parte, el estudio de Eyler et al. (2016) propone un modelo para evaluar el estatus económico en registros de trauma de países de ingresos bajos y medios, donde métricas tradicionales como ingresos o índices de riqueza son difíciles de recolectar. Utilizando datos de activos domésticos provenientes de las encuestas demográficas y de salud (DHS), desarrollaron un algoritmo basado en k-medoids clustering que identifica grupos poblacionales con perfiles económicos similares. Este método utiliza una métrica de disimilitud (Gower) para variables mixtas, lo que lo hace adecuado para datos categóricos como los empleados en este contexto. En una aplicación a los datos del DHS 2011 de Camerún, el algoritmo identificó 20 clusters económicos con un buen ajuste, lo que sugiere que los grupos están bien definidos. El modelo simplifica la evaluación de inequidades económicas en contextos con recursos limitados y puede ser replicado para mejorar el diseño de estrategias de prevención de lesiones y fortalecimiento de sistemas de salud en poblaciones vulnerables.

6.2.5 Anselin Local Moran's I

Por otro lado, Caicedo et al. (2019) en este estudio aborda la problemática de la exposición humana al virus de la rabia en Colombia durante un período de diez años (2007-2016), destacando la incidencia en diferentes escenarios epidemiológicos y poblaciones vulnerables. Los datos utilizados provienen de 666.411 casos notificados a través del sistema de vigilancia pública de Colombia (SIVIGILA). Las variables incluidas abarcaron características sociodemográficas, especies agresoras, tipos de agresión y áreas geográficas. La mayoría de los casos fueron provocados por perros (87.4 %), seguidos por gatos, murciélagos y animales de granja. Los grupos más vulnerables incluyeron niños, estudiantes y comunidades indígenas y afrodescendientes, especialmente en áreas rurales y de difícil acceso, con marcadas desigualdades en el acceso a la atención médica.

Para el análisis aplicó herramientas de análisis espacial y temporal, incluyendo el índice global de Moran para identificar patrones espaciales significativos y el Anselin Local Moran's I para detectar clusters y outliers. Se observaron cuatro escenarios principales: urbano, rural, amazónico y de desigualdad. En el contexto urbano, las agresiones de perros y gatos predominaban en ciudades densamente pobladas, mientras que, en el rural, las agresiones por animales de granja fueron más comunes. El análisis permitió identificar áreas críticas para orientar estrategias de vigilancia y prevención, destacando que la técnica empleada fue efectiva para mapear la incidencia y resaltar las desigualdades geográficas y sociodemográficas en la exposición al virus de la rabia.

6.2.6 Clustering en dos pasos

En cambio, el estudio de Petro et al. (2022) se enfocó en identificar clusters de factores de riesgo para presión arterial elevada (PAE) en adolescentes de 12 a 17 años en Montería, Colombia, analizando variables como sobrepeso, movimiento corporal, dieta de riesgo y aptitud física. La

muestra incluyó a 965 estudiantes, donde el 28,9% presentó PAE. Los datos sociodemográficos y de salud revelaron que factores como bajo movimiento corporal y baja aptitud física eran comunes. Además, el sobrepeso, fue el precursor con mayor asociación directa con PAE.

Para analizar los patrones de riesgo, se utilizó el método de clustering en dos pasos, aplicando el criterio de información bayesiano (BIC) para determinar el número óptimo de clusters y evaluando la cohesión y separación mediante la métrica de silhouette (valor = 0,6, indicando una solución adecuada). Se identificaron cinco clusters, que iban desde un perfil deseable sin factores de riesgo hasta un cluster con los cuatro precursores. Los adolescentes en el cluster con los cuatro factores de riesgo (sobrepeso, baja aptitud física, dieta de riesgo y bajo movimiento corporal) mostraron un riesgo significativamente mayor de PAE. Estos resultados subrayan la necesidad de intervenciones integrales que aborden múltiples comportamientos relacionados con la salud en esta población.

6.2.7 Regresión logística

De manera similar, Satty et al. (2024) investiga los factores asociados con la diarrea en niños menores de cinco años en Yemen, utilizando datos de la encuesta MICS 2022–2023. Con una muestra de 19.561 niños, el análisis incluyó variables como edad, género, nivel educativo de la madre, ingresos del hogar, acceso a agua potable y condiciones de saneamiento. Los resultados descriptivos mostraron que la prevalencia de diarrea fue del 37,3%, siendo mayor en áreas rurales (39,3%), en hogares con ingresos más bajos (44,7%) y entre niños de 6 a 23 meses (47,4%). Además, el nivel educativo limitado de las madres y el uso de sistemas de saneamiento no mejorados también se asociaron con un mayor riesgo de diarrea.

El análisis se realizó mediante regresión logística, empleando razones de probabilidades (OR) tanto no ajustadas como ajustadas. Los resultados mostraron que los niños de hogares más pobres tenían una probabilidad significativamente mayor de padecer diarrea. Además, los niños de 6 a 23 meses presentaron un riesgo elevado.

Con respecto a Dagne et al. (2021), analiza la prevalencia y los factores asociados con la obesidad abdominal en una población adulta de Woldia, Etiopía, en 2020, mediante un diseño transversal basado en la comunidad. La muestra estuvo conformada por 823 adultos seleccionados por muestreo sistemático. Las variables evaluadas incluyeron datos sociodemográficos (edad, género, nivel educativo, ingresos), estilo de vida (actividad física, dieta) y antecedentes médicos (historial familiar de obesidad, enfermedades previas). Los resultados mostraron que la prevalencia de obesidad abdominal fue del 40,4%, siendo mayor en mujeres y en personas mayores de 40 años, así como en aquellos con niveles de ingresos más altos y menor actividad física.

Para identificar las asociaciones entre estas variables y la obesidad abdominal, se utilizó regresión logística. El análisis mostró que factores como la inactividad física, una dieta alta en calorías y antecedentes familiares de obesidad estaban significativamente asociados con un mayor riesgo de obesidad abdominal.

En contraste, Gabsteret al. (2021) realizó un análisis transversal que evalúa la prevalencia y los factores de riesgo asociados con Chlamydia trachomatis (CT) en adolescentes de 14 a 19 años en áreas urbanas y rurales indígenas de Panamá. Se recopilaron datos de 3166 estudiantes mediante un muestreo por conglomerados en dos etapas, con pruebas moleculares de PCR para detectar CT en muestras de orina. La prevalencia general fue del 15,8%, siendo más alta en mujeres (21,6%) que en hombres (9,1%), y no mostró diferencias significativas entre las áreas urbanas y rurales indígenas. Factores como múltiples parejas sexuales y embarazos previos fueron

determinantes en las mujeres, mientras que, en hombres, el incremento de la edad se asoció con una mayor prevalencia.

Para el análisis utilizó regresión logística de efectos aleatorios para identificar asociaciones entre los factores de riesgo y CT. Los modelos multivariables ajustados demostraron que las mujeres con tres o más parejas sexuales o antecedentes de embarazo presentaban mayor riesgo de CT. En hombres, la asociación con los factores estudiados no se mantuvo significativa tras los ajustes.

Adicionalmente, el estudio llevado a cabo por Coker et al. (2020) investiga los factores asociados con la utilización de servicios de salud por parte de madres de niños menores de cinco años con diarrea en Etiopía, utilizando datos de la Encuesta Demográfica y de Salud de 2011. Entre las 1.620 madres incluidas en el análisis, solo el 35% buscó atención médica en centros de salud para tratar la diarrea de sus hijos, mientras que el 60,2% no buscó ningún tratamiento y un pequeño porcentaje recurrió a farmacias o vendedores informales.

La regresión logística multivariable identificó factores clave asociados con una mayor probabilidad de utilizar servicios de salud: madres de áreas urbana, madres con información sobre sales de rehidratación oral, aquellas que asistieron al menos una vez a controles prenatales y aquellas que participaron en conversaciones comunitarias. El estudio concluye que la baja utilización de servicios de salud es un problema significativo, especialmente en áreas rurales.

Por último, se presentan otros artículos que utilizaron esta técnica y se listan en la Tabla 2.

Tabla 2
Otros autores que utilizaron regresión logística.

Cita	Artículo	Año	DOI
Debu Liga et al.	Magnitude and risk factors associated with adolescent pregnancy in Ethiopia	2023	10.69614/ejrh.v15i2.639
Zelka et al.	The effects of completion of continuum of care in maternal health services on adverse birth outcomes in Northwestern Ethiopia: a prospective follow-up study	2022	10.1186/s12978-022-01508-5
Mahumud et al.	Association of dietary intake, physical activity, and sedentary behaviours with overweight and obesity among 282,213 adolescents in 89 low and middle income to high-income countries	2021	10.1038/s41366-021-00908-0
Shaheen et al.	Factors Affecting Jordanian School Adolescents' Experience of Being Bullied	2018	10.1016/j.pedn.2017.09.003
Matias et al.	Prenatal and postnatal supplementation with lipid-based nutrient supplements reduces anemia and iron deficiency in 18-month-old bangladeshi children: A cluster-randomized effectiveness trial	2018	10.1093/jn/nxy078

Cita	Artículo	Año	DOI
Pirkle et al.	Model-based recursive partitioning to identify risk clusters for metabolic syndrome and its components: Findings from the International Mobility in Aging Study	2018	10.1136/bmjopen-2017-018680
El Kishawi et al.	Prevalence and associated factors influencing stunting in children aged 2-5years in the Gaza Strip-Palestine: A cross-sectional study	2017	10.1186/s12887-017-0957-y
Masaku et al.	High prevalence of helminths infection and associated risk factors among adults living in a rural setting, central Kenya: A cross-sectional study	2017	10.1186/s41182-017-0055-8
Lamichhane et al.	Factors associated with infant mortality in Nepal: A comparative analysis of Nepal demographic and health surveys (NDHS) 2006 and 2011	2017	10.1186/s12889-016-3922-z
Gray et al.	Prevalence and incidence of traumatic experiences among orphans in institutional and family-based settings in 5 low- and middle-income countries: A longitudinal study	2015	10.9745/GHSP-D-15-00093

Cita	Artículo	Año	DOI
Azage & Haile	Factors affecting healthcare service utilization of mothers who had children with diarrhea in Ethiopia: Evidence from a population based national survey	2015	10.22605/RRH3493
Schlick et al.	Occupational injuries among children and adolescents in Cusco Province: A cross-sectional study	2014	10.1186/1471-2458-14-766
Aristizábal et al.	Factors associated with fatal trauma in Medellín (Colombia) motorcyclists	2012	10.7705/biomedica.v32i1.603
Thörn et al.	Pneumonia and poverty: A prospective population-based study among children in Brazil	2011	10.1186/1471-2334-11-180
Akmatov	Child abuse in 28 developing and transitional countries-results from the multiple indicator cluster surveys	2011	10.1093/ije/dyq168
Wijesuriya et al.	DIABRISK - SL Prevention of cardio-metabolic disease with life style modification in young urban Sri Lankan's - study protocol for a randomized controlled trial	2011	10.1186/1745-6215-12-209

Cita	Artículo	Año	DOI
Abu Baker & Daradkeh	Prevalence of overweight and obesity among adolescents in irbid governorate, Jordan	2010	10.26719/2010.16.6.657
Semba et al.	Coverage of the national vitamin A supplementation program in Ethiopia	2008	10.1093/tropej/fmm095
Aydin et al.	Effects of sociodemographic factors on febrile convulsion prevalence	2008	10.1111/j.1442-200X.2008.02562.x
Rozi & Akhtar	Smoking among high school adolescents in Karachi, Pakistan [2]	2004	10.1093/ije/dyh128
Azizi et al.	Cardiovascular risk factors in an Iranian urban population: Tehran lipid and glucose study (phase 1)	2002	10.1007/s000380200008
al Arab et al.	The burden of trachoma in the rural Nile Delta of Egypt: A survey of Menofiya governorate	2001	10.1136/bjo.85.12.1406

6.2.8 Minería de datos educativa

Finalizando, Anuradha et al. (2015) afirma que existen técnicas y herramientas de minería de datos que se utilizan comúnmente en el análisis de minería de datos educativos. Entre ellos están los métodos de partición, jerárquicos y basados en la densidad. De modo que los algoritmos más utilizados son K-Means, algoritmo de clustering jerárquico, BIRCH, DBSCAN y clustering

in EDM. Resaltando que el primero resulta ser el más común, debido a su simplicidad computacional y en el desarrollo del algoritmo.

5.3 Discusión

Tras realizar la revisión de literatura sobre las técnicas de clustering más relevantes en el contexto de estudios de agrupamiento a poblaciones vulnerables, se hace evidente la necesidad de seleccionar una metodología que no solo garantice una adecuada segmentación de los datos, sino que también responda a las particularidades del conjunto de datos y los objetivos del presente estudio. Entre las técnicas de clustering analizadas, se mencionan las principales ventajas y desventajas de cada una.

K-Means es conocido por su facilidad en términos de entender y aplicar el algoritmo, también por su escalabilidad, ya que es bueno para bases de datos de gran tamaño, lo que lo convierte en una opción atractiva para este caso. Sin embargo, su principal desventaja es que requiere definir el número de clusters de antemano y no maneja bien los valores atípicos, lo que puede afectar la calidad de los resultados.

Por su parte, DBSCAN es especialmente útil para identificar estructuras no esféricas y para manejar ruido. No requiere especificar el número de clusters, lo que lo hace adecuado en situaciones donde este dato no está disponible. Sin embargo, su desempeño depende de una correcta elección de los parámetros, lo que puede ser un desafío en bases de datos grandes y complejas.

El clustering jerárquico, aunque eficaz en algunos contextos, presenta una alta complejidad computacional, especialmente cuando se trabaja con bases de datos grandes y variables mixtas, como en el caso de este estudio. Esto puede limitar su eficiencia y escalabilidad.

El método K-Medoids, aunque más robusto frente a los outliers que K-Means, no se destaca por su tiempo de ejecución ni por su capacidad de escalar eficientemente en bases de datos grandes. Este algoritmo utiliza puntos de datos reales como centros de los clusters, lo que lo hace más resistente a los outliers, pero su mayor complejidad computacional lo hace menos adecuado cuando se manejan grandes volúmenes de datos.

El clustering en 2 pasos es especialmente adecuado para manejar variables mixtas y puede procesar grandes volúmenes de datos de manera eficiente. Su enfoque de pre-clustering seguido de un clustering final proporciona una gran flexibilidad, permitiendo trabajar con diferentes tipos de datos. Sin embargo, requiere una cuidadosa selección de los parámetros y métricas de distancia para garantizar resultados óptimos.

En cuanto al índice de Moran's I, este es más útil en el contexto del clustering espacial y no se ajusta al enfoque de este proyecto, que no está centrado en la dimensión espacial. Finalmente, la regresión logística es adecuada para determinar las variables significativas que podrían influir en la condición de desplazamiento de los escolares, pero no es una técnica de clustering, ya que está orientada a predecir probabilidades en lugar de realizar segmentaciones directas de los datos.

Finalmente, las alternativas que más se adecúan al problema son K-means, DBSCAN y clustering en dos pasos. Debido a que la presencia de ruido puede afectar la viabilidad de la técnica que se aplique, se evaluará este factor en la base de datos y en caso de presentarse significativamente se elegirá DBSCAN, puesto que esta técnica lo maneja de una mejor forma. En caso contrario, se elegirá K-means debido a su simplicidad computacional. Clustering en dos pasos más que representar una técnica, es una metodología que implica combinar dos técnicas, por lo tanto, se mantendrá como una alternativa para tener en cuenta durante la ejecución del análisis de datos.

6. Limpieza, preprocesamiento y transformación de los datos

Antes de aplicar cualquier técnica de minería de datos, es necesario llevar a cabo un preprocesamiento y limpieza de la información que permita garantizar la calidad, consistencia y pertinencia de los datos que se emplearán en los modelos de clustering que sean seleccionados según los criterios mencionados en la sección anterior. Para llevar esto a cabo, se utilizó el lenguaje de programación Phyton por medio del entorno online Google colab, junto con librerías especializadas en análisis de datos como pandas, numpy, scikit-learn, matplotlib, entre otras. Todos los códigos, así como resultados podrán consultarse en el apéndice I adjunto a este documento.

El preprocesamiento incluye diferentes fases:

- Comprensión de la naturaleza de las variables a trabajar.
- Detección de valores faltantes, inconsistentes o irrelevantes.
- Eliminación de variables irrelevantes para el estudio.
- Transformación de las variables para su debida adecuación a los algoritmos de agrupamiento.

6.1 Descripción preliminar de las variables

La base de datos contiene variables de diferentes naturalezas. En este primer paso, se realiza una revisión exploratoria de los datos, para comprender su tipo, utilidad y posibles limitaciones en el contexto del análisis.

Mediante una línea de código: data.info(), siendo “data” el dataframe importado a phyton y por medio de la librería pandas, se logra visualizar en la Figura 7 Información de la base de datos las diferentes variables con las que cuenta la base de datos junto con su respectivo tipo, así como el tamaño del conjunto de datos.

Figura 7
Información de la base de datos

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18062 entries, 0 to 18061
Data columns (total 22 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   d_ano        18062 non-null  int64 
 1   d_muni       18062 non-null  int64 
 2   d_nombmuni   18062 non-null  object
 3   d_provincia  18062 non-null  object
 4   d_nomsec     18062 non-null  object
 5   d_nomzon     18062 non-null  object
 6   dane_ant     18062 non-null  int64 
 7   d_nombinst   18062 non-null  object
 8   d_sede       18062 non-null  object
 9   d_nombsede   18062 non-null  object
 10  d_nomjor     18062 non-null  object
 11  d_grado      18062 non-null  int64 
 12  d_edad       18062 non-null  int64 
 13  d_genero     18062 non-null  object
 14  d_hombres    18062 non-null  int64 
 15  d_mujeres    18062 non-null  int64 
 16  d_tipo       18062 non-null  object
 17  metodo       18062 non-null  object
 18  sector       18062 non-null  int64 
 19  edad         18062 non-null  int64 
 20  etnia        18062 non-null  object
 21  discapa      18062 non-null  object
dtypes: int64(9), object(13)
memory usage: 3.0+ MB
```

Se puede observar que la base de datos se conforma por 22 campos y 18062 registros. Dentro de los tipos de variables, se pueden encontrar tres: categórica, binaria y enteras.

En primer lugar, las variables enteras son siete, aunque en la Figura 7 Información de la base de datoshaya nueve con la etiqueta “int64”, esto se debe a que dos de ellas son binarias. Acá se encuentran, por ejemplo, el año, el número asociado al municipio, la nomenclatura especializada del DANE el mismo también, el grado, la edad del encuestado (que se presenta dos veces) y el estrato. Algunas de estas, como lo es el número del municipio o la nomenclatura del DANE, no aportan información relevante para la caracterización mediante el clustering, dado que funcionan como etiquetas administrativas. Por otro lado, las demás variables,

en un principio, representan utilidad en el contexto al que se ha referido, pues representan dimensiones de interés para el análisis.

En segundo lugar, se presenta un par de variables binarias o dummies, una que hace referencia al género masculino y otra al femenino. Esta también se considera relevante, pues puede permitir identificar diferencias o similitudes entre los grupos de interés que surjan próximamente.

Por último, en las variables categóricas están el nombre de la provincia, del municipio, el sector del encuestado (colegio público o privado), nombre de la zona (rural o urbana), nombre, sede (y nombre de esta) de la institución educativa, el género (pero esta vez en palabras), tipo de desplazamiento, método de educación, etnia y discapacidad. Algunas de estas son relevantes pues describen características de las condiciones de los estudiantes, como, por ejemplo, el sector, la zona y el tipo de desplazamiento, entre otras; mientras tanto, otras como el nombre del colegio y sede se torna demasiado específico, por lo que se considera no incluir en el análisis.

6.2 Limpieza de datos

El segundo paso previo al aplicar el modelo de clustering a los datos, consiste en la limpieza de estos, con el fin de garantizar que la información a utilizar sea consistente y representativa.

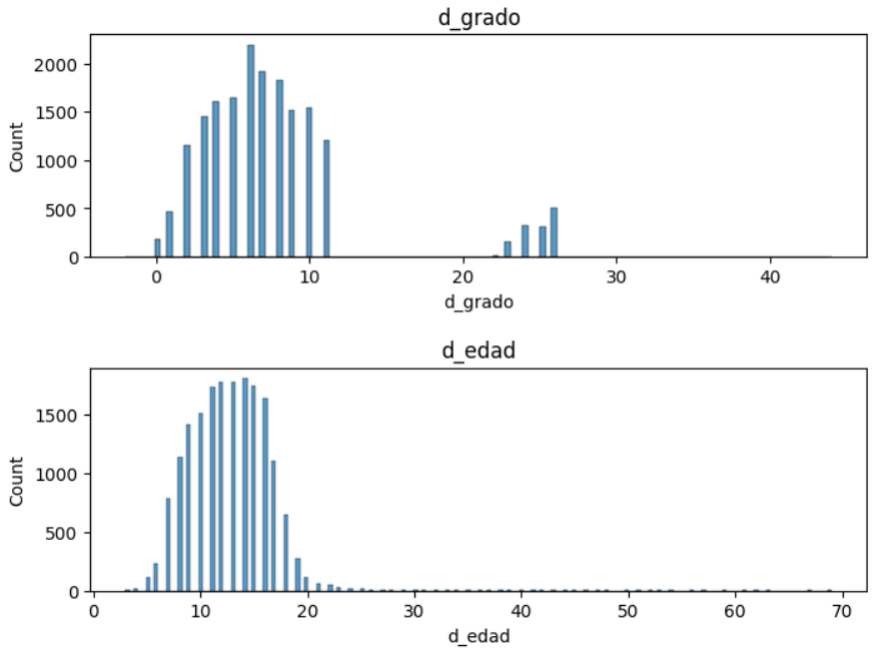
Primero, se procedió a la eliminación de registros duplicados. Inicialmente la base de datos contaba con 18.062 observaciones; sin embargo, tras poner en marcha lo explicado mediante la línea de código data.drop_duplicates(), el número final se redujo a 12.684 registros válidos. Esto se puede explicar de dos maneras, la primera es que haya existido errores durante la digitación y posterior carga de la información por parte del DANE a los sistemas de información; y la segunda es que dos estudiantes presenten las mismas características, además de estudiar en el mismo

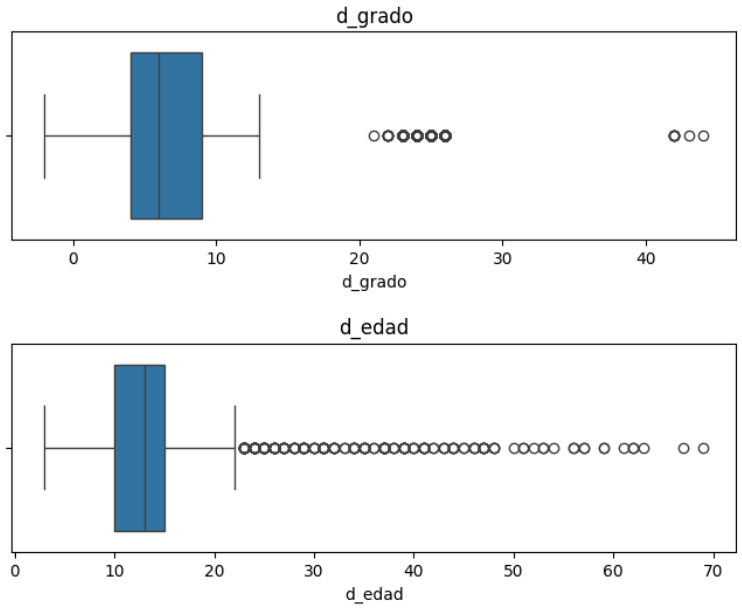
colegio, por lo que, a final de cuentas, representarán matemáticamente el mismo punto en la caracterización que se realizará.

Adicional a esto, se realiza un análisis exploratorio de los datos mediante representaciones gráficas. Para esto, los diagramas de barras serán útiles para las variables categóricas mientras que los diagramas de cajas y bigotes (boxplot) e histogramas lo son para las variables numéricas. Esto, para el primer caso, permite ver las distribuciones y frecuencias, de esta manera se podrá determinar la utilidad de una variable; en cambio, para el segundo, permite ver si se presentan casos atípicos u outliers que se deban tener en cuenta durante la ejecución del método de clustering a utilizar.

Figura 8

Distribución y boxplot de las variables grado y edad



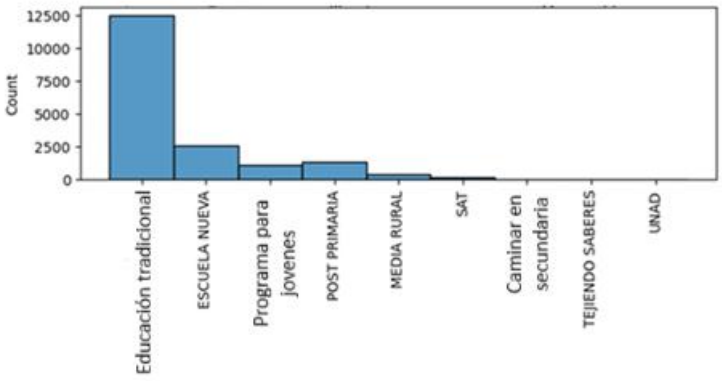


En la Figura 8

Distribución y boxplot de las variables grado y edad se puede observar, por un lado, que la variable grado, en su mayoría se distribuye como se esperaría con base en el sistema educativo tradicional, donde se distribuye de 1° a 11°, sin embargo, presentan valores negativos o 0, y por otro lado, valores entre 20 y 30; esto se explica ya que dentro de la base de datos se incluyen aquellos estudiantes que se encuentran en preescolar y en el otro extremo, aquellos que se han re insertado a la escolaridad, con métodos de educación que se describirán cuando se mencione dicha variable. Esto se refuerza mediante la distribución de la variable edad, donde se ve una mayoría de frecuencia en las edades relacionadas a grados de primaria y bachillerato, pero también se encuentran desde los 3 hasta los 69 años; sin embargo, no se consideran como valores atípicos, pues se tiene en cuenta que hay métodos de estudio que están diseñados para aquellas personas que se quieran volver al mundo académico sin importar su edad.

Figura 9

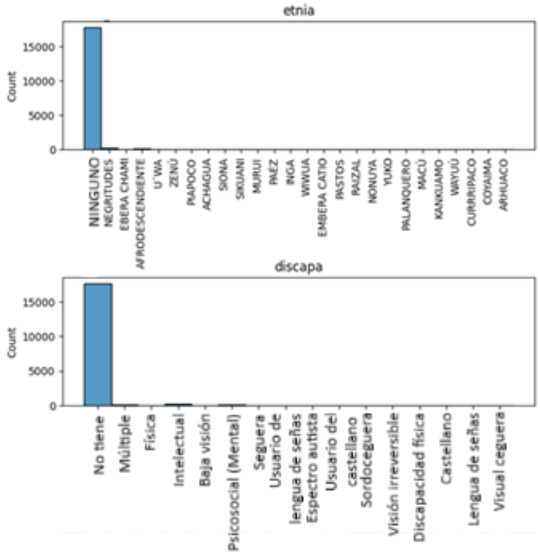
Distribución de la variable método de educación



En la Figura 9, se pueden observar los diferentes métodos de educación que tienen los encuestados, siendo la mayoría de educación tradicional, pero también considerando las otras opciones que, si bien pueden parecer una minoría, puede hacer la diferencia en la aplicación de la caracterización planteada.

Figura 10

Distribución de las variables etnia y discapacidad

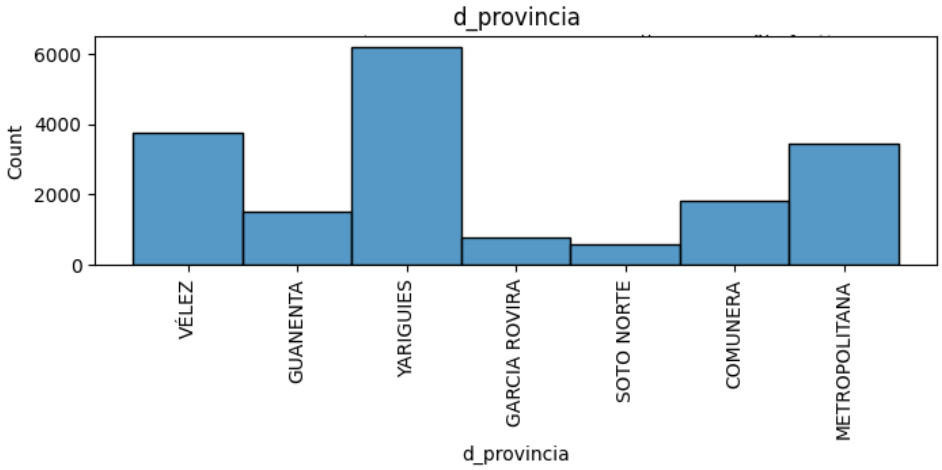


En la Figura 10 Distribución de las variables etnia y discapacidad, se observa que la mayoría de los estudiantes no

pertenecen a una etnia y no presentan discapacidad. Por este motivo, incluir estas dos variables en el análisis podrían generar ruido en los resultados sin aportar información relevante.

Figura 11

Distribución de la variable provincias de Santander



La Figura 11 Distribución de la variable provincias de Santander permite observar que queda mejor representar la distribución geográfica de los estudiantes a partir de las siete provincias en las que se divide Santander, en vez de los 86 municipios con los que cuenta el departamento.

Tras aplicar los procesos de depuración y análisis exploratorio, se define un conjunto de variables que serán consideradas en la siguiente etapa de transformación. Estas quedan de la siguiente manera:

- Edad. Variable numérica continua, clave para conocer el perfil etario de los estudiantes.
- Grado de escolaridad. Variable numérica discreta, permite conocer el nivel académico alcanzado al momento del registro.

- Provincia. Variable categórica, permite aportar un contexto geográfico a la base de datos.
- Género. Variable binaria, da contexto sobre el estudiante.
- Naturaleza de la institución. Variable binaria, aporta al contexto socioeconómico del estudiante.
- Zona de la institución. Variable binaria, también aporta al contexto socioeconómico del estudiante.
- Método de educación. Variable categórica, permite conocer las condiciones de estudio, además habla de las posibles barreras que tenga el registrado.
- Método de desplazamiento. Variable categórica, brinda un contexto sobre el motivo de la violencia que lo llevó a la condición mencionada.

De manera que, de los 22 campos, realmente quedan 9 que se consideraron relevantes para llevar a cabo la caracterización. Los campos que no fueron tenidos en cuenta son: año en el que se tomó el registro, nombre, número y nomenclatura del DANE del municipio, nombre, sede y nombre de la sede de la institución educativa, género (en palabras), al igual que el campo binario que resaltaba si el estudiante era o no del género femenino, sector socioeconómico, edad (pues estaba repetida), etnia y discapacidad.

6.3 Transformación de los datos

Por último, se llevó a cabo la transformación de variables categóricas mediante la técnica de codificación dummy. Esta decisión se adoptó para evitar que, al asignar valores numéricos arbitrarios a categorías (como 1,2,3…etc.) el modelo pueda interpretar que cierta característica tenga más importancia que otra, cuando realmente no se pretende esto. La codificación dummy transforma cada valor posible de la categoría en variables binarias, con valores de 0 o 1 según

corresponda. De tal manera que si un estudiante pertenece a la provincia Yarima, en ese campo saldrá 1, mientas que en las demás provincias resultarán 0.

Un aspecto importante de esta técnica es que genera n-1 categorías dummies para una categoría con n características, de tal manera que si los n-1 campos resultan 0, se entenderá que el campo al que le pertenece el valor 1 es aquel que no se haya incluido. Bajo esta lógica, en el caso de la variable binaria género, fue necesario eliminar una de las columnas ya existentes, ya que ambas representaban la misma información solo cambiando la perspectiva. Mantener ambas columnas habría ocasionado redundancia y por lo tanto ruido en la ejecución del algoritmo de clustering.

De esta manera, la base de datos resultante se estructuró de manera más limpia y adecuada, minimizando errores y asegurando que cada variable aporte información útil al posterior análisis.

Al realizar este paso, los 9 campos originales se convirtieron en 40 columnas con variables dummies y 2 numéricas. De tal forma que la base de datos resultante contiene 42 campos y 12.684 registros.

7. Aplicación de los métodos de clustering

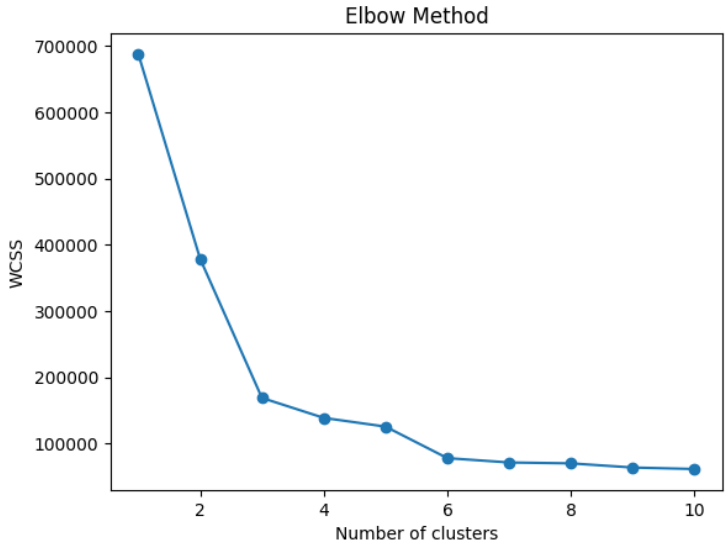
De acuerdo con el análisis de literatura realizado, se consideraron dos algoritmos de clustering como los más adecuados para la caracterización de los estudiantes en condición de desplazamiento: k-means y DBSCAN. Por un lado, el primero es un algoritmo altamente utilizado debido a su simplicidad y eficiencia, por otro lado, el segundo resulta útil para casos donde se presentan valores atípicos, pues permite identificarlos sin forzar su inclusión a algún grupo, además, permite no definir desde un principio el número de clusters resultantes.

Dado que el análisis exploratorio de los datos reveló una baja presencia de valores atípicos, se decide aplicar ambos métodos de manera comparativa. Esta estrategia permite no solo confirmar la robustez de los resultados de k-means, sino contrastarlos con la flexibilidad del DBSCAN, generando de esta manera una mayor calidad y profundidad en el análisis de los resultados.

7.1 K-means

Antes de la implementación del algoritmo k-means, es necesario hallar un método que permita el cálculo del número óptimo de clusters (k). Para ello, se utilizó el método del codo, el cual consiste en calcular la suma de los errores cuadráticos internos (inertia) para diferentes valores de k (Rodríguez, 2023). Al graficar dichos valores, se observa como progresivamente disminuye la inercia conforme va aumentando el número de clústeres; sin embargo, a partir de cierto punto, las mejoras disminuyen significativamente, generando así una curvatura o "codo" en la gráfica. Este punto se interpreta como el valor óptimo de k para el modelo que se está planteando.

Figura 12
Método del codo

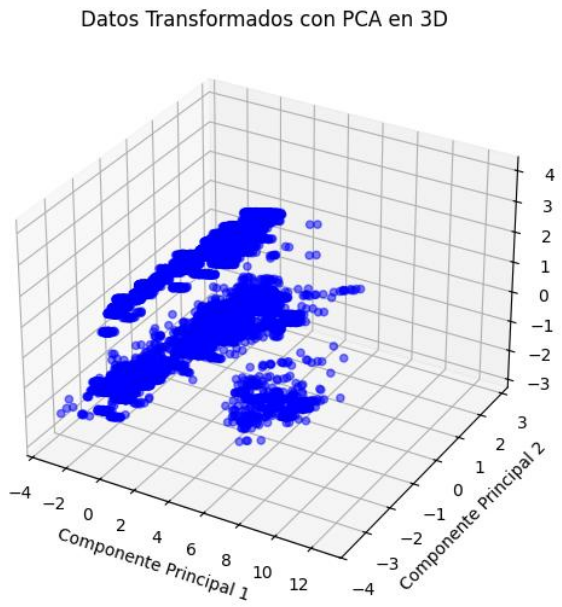


La Figura 12

Método del codo muestra la gráfica resultante de aplicar este método, la cual muestra que el punto que se está buscando es k=3, sin embargo, también es pertinente verificar los valores de k=4 y k=5 a la hora de aplicar el algoritmo de clustering mencionado.

Adicionalmente, en la búsqueda de la determinación del número de clústeres óptimo para este ejercicio, se plantea realizar un análisis de componentes principales (PCA), cuyo objetivo es reducir el número de dimensiones de grandes conjuntos de datos manteniendo la mayor parte del contenido original; para esto, busca aparentes correlaciones entre las diferentes variables hasta lograr un conjunto más pequeño denominado componentes principales (IBM, 2023).

Figura 13
Análisis de componentes principales



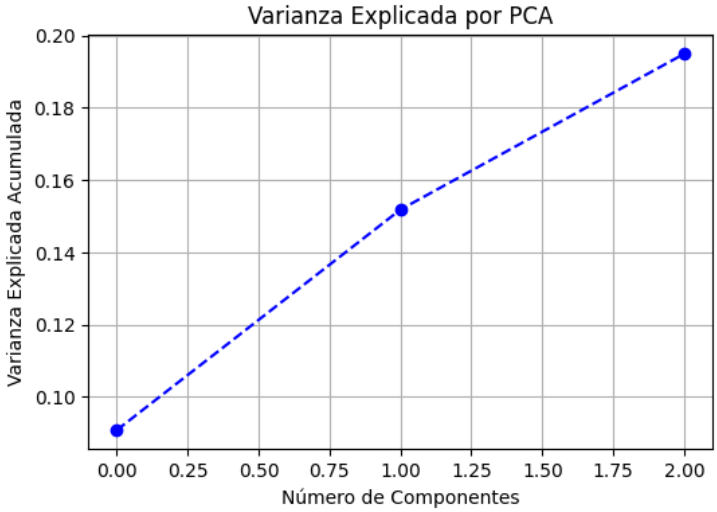
La Figura 13

Análisis de componentes principales permite observar en tres dimensiones como aparentemente los datos se distribuyen en tres grupos, por lo tanto, esto refuerza la idea del resultado proveniente

del método del codo, de tomar como valor k=3, fundamentándolo tanto como del punto de vista matemático, como en la evidencia que otorga el PCA, lo cual otorga robustez a esta decisión.

Cabe resaltar que, tras la transformación de variables categóricas a dummies, la base de datos quedó con 42 dimensiones, por lo que es entendible que al realizar el PCA, este solo logre explicar un poco menos del 20% de la varianza, como lo indica la Figura 14 Varianza explicada por PCA.

Figura 14
Varianza explicada por PCA



Una vez definido el parámetro que se necesitaba (k=3), se inicia con la ejecución del método k-means. Para esto, se utilizaron librerías altamente conocidas en el mundo de análisis de datos con Phyton, tales como pandas y scikit-learn, las cuales permiten realizar el cálculo de manera eficiente y confiable.

El algoritmo asignó cada uno de los registros a uno de los tres clústeres definidos, con base en la minimización de la distancia euclidiana respecto a los centroides de estos.

Además, con el propósito de evaluar la calidad del método, se procede a calcular el silhouette score, un indicador que mide la coherencia de la agrupación de cada uno de los estudiantes en los diferentes grupos. Este puede tomar valores entre -1 y 1, sugiriendo el valor mínimo asignaciones incorrectas, mientras que el valor máximo refleja clústeres bien separados y compactos. Se logra un silhouette score de 0,5023; lo cual se interpreta como que se logra una separación adecuada mas no perfecta, lo cual es consistente debido a la complejidad de los datos y la alta dimensionalidad de las variables.

De esta manera, la ejecución de k-means no solo logra generar una caracterización de los estudiantes estudiados en tres grupos, sino también validar la pertinencia del resultado mediante el cálculo de un indicador matemático objetivo, estableciendo así una base sólida para la futura comparación con el algoritmo DBSCAN.

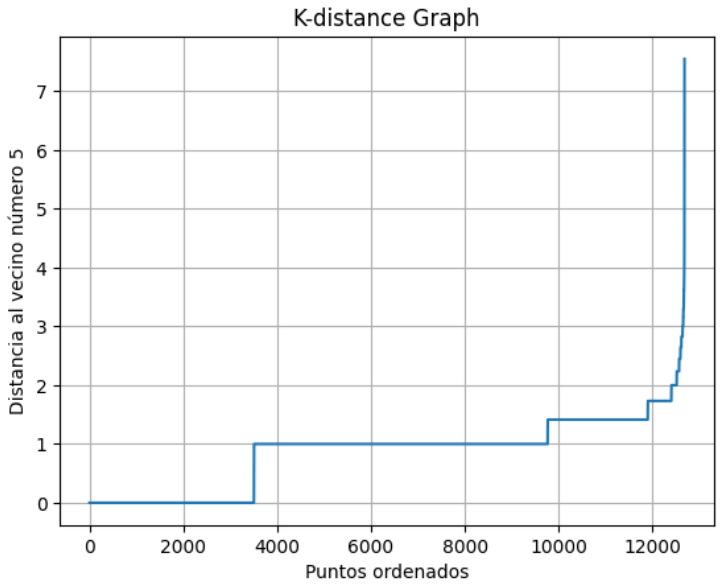
Cabe aclarar que, con el objetivo de evitar que variables como la edad o grado del estudiante tomara mayor importancia que las variables dummies debido a la naturaleza de cada uno, se decidió escalar los datos por medio de las librerías ya mencionadas antes, especialmente scikit-learn; sin embargo, el silhouette score muestra que, de esta manera, no es posible igualar la calidad del ajuste logrado anteriormente, pues resulta ser 0,1422. Esto se puede explicar debido a que al ser estudiantes el objeto de estudio en este caso, información como la edad y el grado constituyen atributos determinantes en la caracterización de estos, de manera que su peso no debe ser reducido como se pensaba al inicio. Por este motivo, se optó por mantener el modelo en su versión original, sin aplicar escalamiento.

7.2 DBSCAN

Para la implementación del algoritmo DBSCAN fue necesario, primero, calcular los parámetros de ajuste: épsilon (eps) y número mínimo de muestra (min_samples). El primero representa la distancia máxima entre dos puntos para que puedan considerarse vecinos. Para su cálculo, se utilizó el gráfico de k-distancias utilizando el quinto vecino más cercano. Este gráfico permite determinar la inflexión o la curva que se crea al ordenar las distancias entre cada punto y su vecino k-ésimo. Dicho punto permite diferenciar regiones densas de áreas dispersas o ruido.

Figura 15

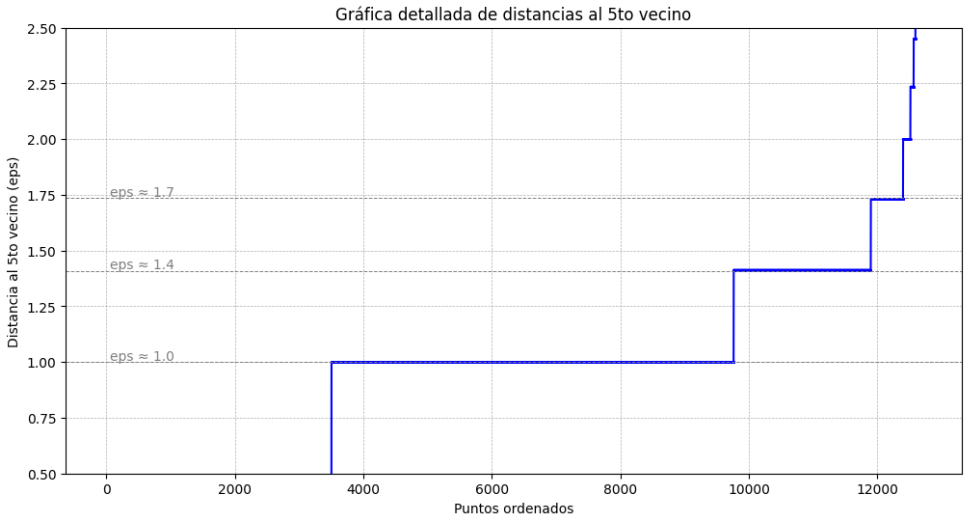
Gráfico k-distancias para el cálculo del parámetro epsilon



La Figura 15 Gráfico k-distancias para el cálculo del parámetro epsilon muestra el resultado del proceso descrito, en donde se puede observar que los valores donde se produce ese punto de inflexión o curva que se busca, está entre 1 y 2. Para esto, la Figura 16 Gráfico de k-distancias detallado permitirá ver más de cerca la gráfica y así decidir cuáles serán los resultados de épsilon a establecer en el algoritmo.

Figura 16

Gráfico de k-distancias detallado



De esta manera, épsilon tomará valores de 1, 1.4, 1.7 y, adicionalmente, se probarán valores de 1.8, 2 y 2.2 para observar como se comporta el silhouette score en función de aumentar este parámetro, aunque también se prestará atención a la interpretabilidad a la hora de analizar los resultados.

El segundo parámetro, el cual es el número mínimo de puntos necesarios para que una población se considere densa, o una muestra. Si una observación cuenta con al menos el número mínimo de muestra dentro del radio épsilon, se define como un punto núcleo y puede dar origen a un cluster. A diferencia del primer parámetro, este no cuenta con un método o gráfico que determine un valor óptimo; en la práctica, se selecciona mediante prueba y error o siguiendo convenciones. En este estudio, se utilizó en un principio el valor de 5 debido a que es el valor predeterminado que trae la librería Scikit-learn, al igual que como lo dice Yadav (2024).

Figura 17
Silhouette score del método DBSCAN

eps	Clusters	Silhouette Score	% Outliers
1.00	72	-0.3899	14.81
1.40	72	-0.3899	14.81
1.70	3	0.6501	2.78
1.80	4	0.6756	1.43
2.00	6	0.7029	0.91
2.20	6	0.7029	0.91

La Figura 17 muestra como varió el silhouette score, número de clústers y el porcentaje de outliers en función del valor de épsilon.

Posteriormente, en búsqueda de garantizar que las clasificaciones realizadas fueran estadísticamente sólidos, se realizó un pequeño cambio frente a esta aplicación del método DBSCAN. Esta vez, manteniendo los valores de épsilon que funcionaron, su buscó aumentar el parámetro número mínimo de muestra de 5 a 70, esto para ver cómo cambia el silhouette score, y, al mismo tiempo, asegurar que los clústeres identificados tengan un tamaño suficiente para resultar interpretables dentro del análisis social que se plantea. La Figura 18 muestra el resultado.

Figura 18
Silhouette score cambiando el número mínimo de muestra

eps	Clusters	Silhouette Score	% Outliers
1.70	1	nan	25.79
2.00	2	0.7047	3.52
2.20	2	0.7047	3.52

8. Resultados

Con el fin de obtener una caracterización clara de los estudiantes en condición de desplazamiento en Santander, se presentan los resultados de la aplicación de los algoritmos de clustering: k-means y DBSCAN. Estos se encuentran evaluados bajo diferentes indicadores de

calidad, como el silhouette score, número de clústeres, la proporción de estudiantes pertenecientes a cada uno de éstos, la interpretabilidad de los datos y el porcentaje de observaciones catalogadas como atípicas (outliers) para el caso del segundo.

Primero, se ejecutó el método de k-means evaluando diferentes valores de k con el objetivo de identificar la configuración más pertinente. En la Figura 19, Figura 20 e Figura 21 se muestran las proporciones de los grupos identificados con k-means; además, en la Figura 22 se puede observar el valor del silhouette para cada caso.

Figura 19

Proporción de clústeres con k-means (k=3)

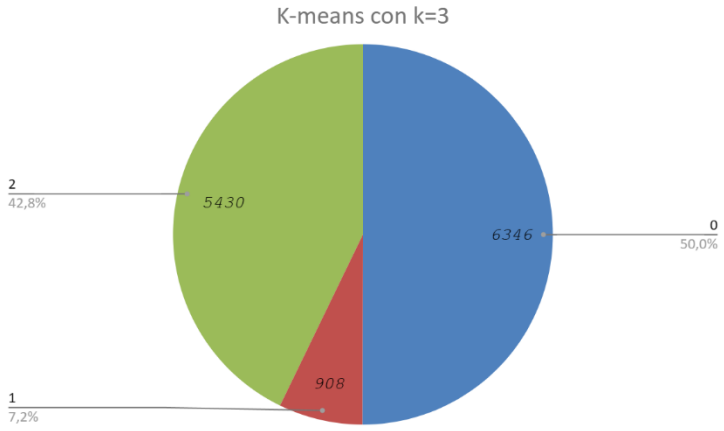


Figura 20

Proporción de clústeres con k-means (k=4)

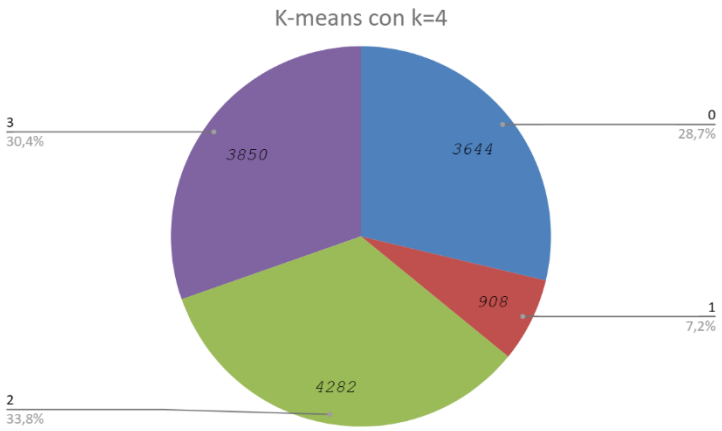


Figura 21

Proporción de clústeres con k-means (k=5)

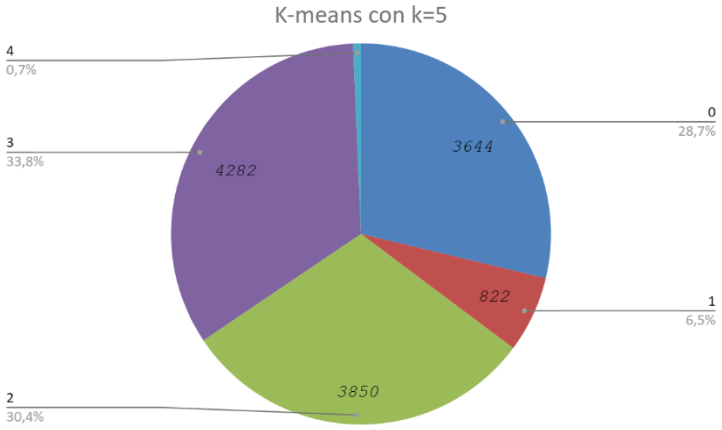
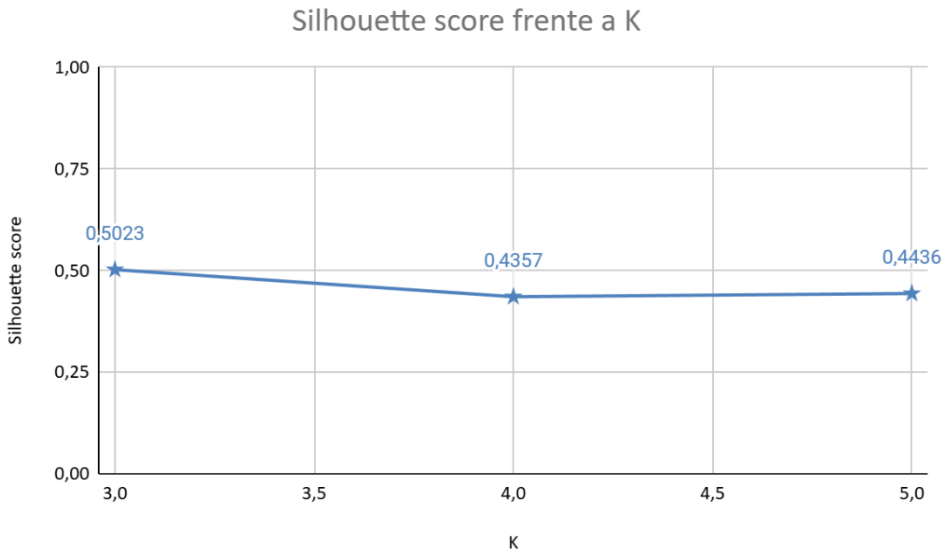


Figura 22
Silhouette score frente a k



De esta manera, los resultados obtenidos para k=5 pueden descartarse de manera preliminar, ya que uno de los clústeres agrupa únicamente a 86 estudiantes. Desde el punto de vista de la relevancia analítica, este tamaño resulta poco representativo, pues corresponde aproximadamente al 1% del total de la muestra. En consecuencia, no sería viable generar información que sirva, entre otras cosas, para diseñar estrategias diferenciadas para un subgrupo tan reducido, existiendo alternativas con otros valores de k que generan particiones más equilibradas y útiles para el análisis.

Por otro lado, entre k=3 y k=4, si bien existe una proporción más justa a lo que se busca, el primero muestra un mejor ajuste. Esto se debe a que, al superar el umbral de 0,5 del silhouette score, se puede afirmar que los clústeres formados presentan una mayor coherencia interna y una separación más definida entre los grupos, es decir, muestra que los estudiantes se agrupan de manera más natural y diferenciada.

A diferencia de k-means, que obliga a dividir la muestra en un número fijo de clústeres previamente definidos, DBSCAN ofrece la ventaja de identificarlos, y al mismo tiempo, determinar los estudiantes que no encajan en ninguno, tratándolos como outliers. Esto permite observar una dinámica distinta, pues no se divide con base en el número de grupos en los que se debe dividir, sino la densidad de los datos en el espacio de características. Sin embargo, la selección de los parámetros como épsilon y número mínimo de muestra puede modificar de manera significativa los resultados, por lo que resulta necesario explorar varias configuraciones antes de seleccionar la más adecuada.

Al aplicar esto con un valor fijo de min_samples=5, y unos valores variables de épsilon que se puede observar en la Figura 17 (1.0, 1.4, 1.7, 2.0 y 2.2), es posible no considerar los primeros dos valores debido a que presentan un silhouette score negativo, lo que indica que para estos casos, se presenta una baja coherencia en la conformación de los grupos, además que se generan demasiados (72 clústeres para estos casos); además, el último presenta los mismos resultados que el penúltimo. Así, el análisis se centra en los tres escenarios restantes, donde se identifican particiones con 3, 4 y 6 clústeres, cada una con proporciones distintas tanto en su composición interna como en su proporción de outliers.

Tabla 3
Resultado con DBSCAN (eps=1.7)

Cluster	Cantidad
0	11669
1	590
2	72

Figura 23
Proporción de clústeres con DBSCAN (eps=1.7)

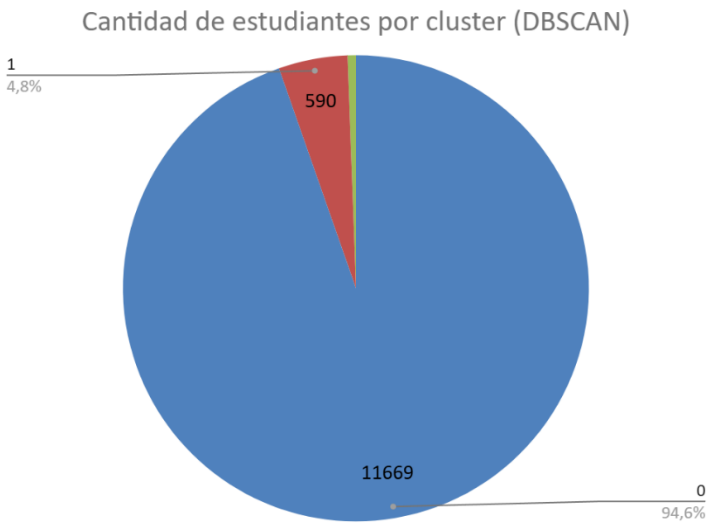


Tabla 4
Resultado con DBSCAN (eps=1.8)

Cluster	Cantidad
0	11760
1	733
2	5
3	4

Figura 24
Proporción de clústeres con DBSCAN (eps=1.8)

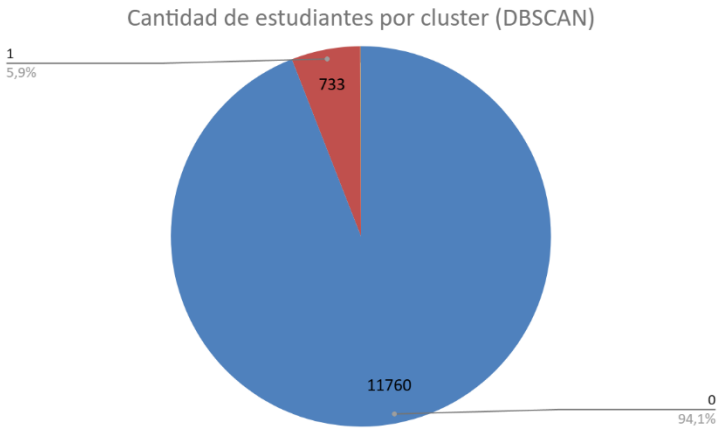
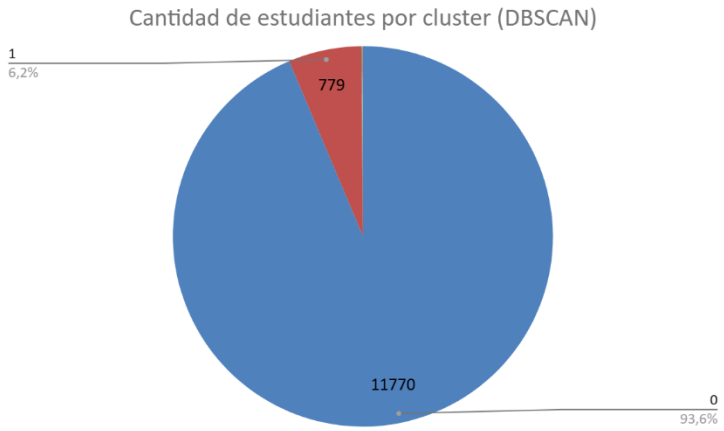


Tabla 5
Resultado con DBSCAN (esp=2.0)

Cluster	Cantidad
0	11770
1	779
2	5
3	5
4	5
5	5

Figura 25

Proporción de clústeres con DBSCAN (eps=2.0)



En la información ilustrada entre las ilustraciones 23 y 25, y entre las tablas 3 y 5, se puede determinar que, si bien, no hay un caso en el que la proporción sea la más apropiada, pues en el caso más equilibrado, hay un cluster con 70 estudiantes, se opta por elegir dicho resultado, el cual se logró con un valor de épsilon=1.7 y un valor de min_samples=5. Esta decisión se toma con el fin de tener una buena base de comparación frente a los resultados obtenidos mediante k-means.

Posteriormente, se cambió el valor de min_samples a 70 con el mismo objetivo de comparar diferentes resultados. Se utilizaron los mismos valores de épsilon, y presentaron los mismos resultados: dos clústeres, uno con 11741 y otro con 497 estudiantes. Entonces, se decide no tener en cuenta esta variación, ya que muestra una caracterización muy general y no será de mayor utilidad al compararse con el otro método ya usado.

En consecuencia, se analiza el resultado de k-means con k=3 y el mismo de DBSCAN con eps=1.7 y min samples=5.

8.1 Análisis descriptivo de los resultados

Con el fin de comprender las características de los grupos generados, se realiza un análisis estadístico-descriptivo de los estudiantes pertenecientes a cada cluster, tanto para el modelo k-means como DBSCAN.

8.1.1 K-means

La Tabla 6 muestra los valores de edad y grado, junto con algunos estadísticos descriptivos principales. Se incluyen la media, desviación estándar y los cuartiles 1,2 y 3 (que representan el 25%, 50% y 75% de los datos), dado que estos permiten mostrar la tendencia central, la dispersión y amplitud de los datos.

Tabla 6
Estadísticos descriptivos de edad y grado por clúster (K-means)

Variable	Estadístico	Clúster 1	Clúster 2	Clúster 3
Edad	Media (años)	14,6	21,07	9,23
	Desviación estándar	1,85	8,3	1,73
	Q1-Q2-Q3	13-15-16	17-18-21	8-9-11
Grado	Media	8,29	24,8	3,45
	Desviación estándar	1,72	1,89	1,54
	Q1-Q2-Q3	7-8-10	24-25-26	2-4-5

En términos de edad y grado escolar, los clústeres muestran diferencias claras. Se puede ver que en el tercero se encuentran los estudiantes más jóvenes, que, como se esperaría, cursan los

grados menores, y luego van progresivamente agrupando a los de mayor edad y nivel educativo en los clústeres 1 y 2. Los valores de los cuartiles refuerzan esta progresión, respaldando también la diferencia estadística entre los grupos.

Tabla 7
Distribución de género por clúster (%)

Género	Clúster 1	Clúster 2	Clúster 3
Masculino	51,67	51,98	52,95
Femenino	48,33	48,02	47,05

En términos generales, no se presenta una diferencia representativa entre el género masculino o femenino, sin embargo, se observa que el primero es mayoritario para los tres clústeres. Ahora bien, pasando a las características geográficas plasmadas en la Tabla 8 se pueden observar contrastes más notorios.

Tabla 8
Distribución de la provincia de residencia por clúster (%)

Provincia	Clúster 1	Clúster 2	Clúster 3
García Rovira	5,34%	6,50%	4,73%
Guanenta	10,09%	**20,04%**	8,97%
Metropolitana	**18,20%**	14,87%	**18,73%**
Soto Norte	4,10%	0,88%	4,36%
Vélez	**23,35%**	12,44%	**20,63%**

Provincia	Clúster 1	Clúster 2	Clúster 3
Yariguies	**27,99%**	**21,37%**	**33,70%**
Comuneros	10,94%	**23,90%**	8,88%

En la Tabla 8 se puede observar que la única provincia del departamento de Santander que representa una participación relevante en cada uno de los clústeres es Yariguies, además, para el primero y tercero también los conforman, en mayoría, Vélez y Metropolitana, mientras que, para el segundo, Guanenta y Comuneros.

Por otro lado, algo importante en esta caracterización es la descripción del carácter de la institución educativa, así como la zona en donde esta se encuentra ubicada.

Tabla 9
Distribución por carácter y zona de la institución (%)

Variable	Clúster 1	Clúster 2	Clúster 3
Institución pública	97,68%	35,13%	95,73%
Institución privada	2,32%	64,87%	4,27%
Zona urbana	51,13%	69,16%	43,02%
Zona rural	48,87%	30,84%	56,98%

En la Tabla 9 se observa que para los clústeres 1 y 3, predomina la institución pública, mientras que no hay una clara distinción entre la zona donde está ubicada la institución. En contraste, el clúster 2 presenta una mayor inclinación por la institución privada y la zona urbana.

Ahora bien, resulta relevante observar cómo se organizan las jornadas en las que los estudiantes reciben clases, dado que esta variable refleja la forma en la que las instituciones organizan los recursos educativos, así como los estudiantes como adaptaban este contexto a sus rutinas.

Tabla 10
Distribución por jornada académica (%)

Jornada	Clúster 1	Clúster 2	Clúster 3
Fin de semana	0,00%	**42,07%**	0,00%
Mañana	**67,24%**	16,85%	**72,93%**
Nocturna	0,00%	**28,08%**	0,00%
Tarde	7,11%	1,10%	**11,93%**
Única	**22,46%**	0,22%	**11,90%**
Completa	3,20%	11,67%	3,24%

En la Tabla 10 se muestra la importancia de las jornadas de la mañana y única para los clústeres 1 y 3, añadiendo la de la tarde para este último también, a diferencia del segundo clúster, donde predomina las jornadas del fin de semana y nocturna por sobre las otras opciones.

Como parte final de la descripción del entorno educativo de los estudiantes, se incluyen las distribuciones de los métodos de enseñanza que se pueden ofertar en el país a través del aval del ministerio de educación, lo que permitirá tener una visión más clara y completa de sus condiciones de formación.

Tabla 11
Distribución por método de enseñanza (%)

Método de enseñanza	Clúster 1	Clúster 2	Clúster 3
Educación tradicional	**77,23%**	0,77%	**55,19%**
Escuela nueva	1,83%	0,00%	**42,91%**
Media rural	5,26%	0,00%	0,00%
Post primaria	**15,49%**	0,00%	1,90%
Programa para jóvenes	0,00%	**79,52%**	0,00%
SAT	0,00%	**17,51%**	0,00%
Tejiendo saberes	0,00%	0,11%	0,00%
UNAD	0,00%	2,09%	0,00%
Caminar en secundaria	0,19%	0,00%	0,00%

En la Tabla 11 se puede ver que, si bien hay una gran variedad de métodos de enseñanza disponibles, cada cluster cuenta con solo dos que conforman más del 90% para cada uno de los casos. La educación tradicional es mayoritaria para los clústeres 1 y 3, mientras que post-primaria también forma una proporción importante para el primero, así como escuela nueva para el tercero. En cambio, el segundo clúster muestra mayor proporción en programas para jóvenes y SAT.

Por último, es importante para cerrar la caracterización, presentar la distribución de la variable asociada al tipo de desplazamiento que sufrieron los estudiantes, dado que ésta aporta el

contexto social necesario para comprender las condiciones particulares de la población objeto de estudio del presente trabajo de grado.

Tabla 12
Distribución por tipo de desplazamiento (%)

Tipo de desplazamiento	Clúster 1	Clúster 2	Clúster 3
Actos terroristas	0,03%	0,00%	0,06%
Amenazado	0,11%	0,22%	0,13%
Confinamiento	0,00%	0,00%	0,02%
Delitos contra la libertad	0,02%	0,00%	0,09%
Desaparición forzada	0,13%	0,44%	0,64%
Desplazamiento forzado	5,36%	6,39%	9,98%
Desvinculados de grupos armados	0,08%	2,97%	0,06%
En situación de desplazamiento	**26,08%**	**39,32%**	**19,48%**
Hijos de adultos desmovilizados	0,61%	1,76%	0,66%
Homicidio	0,06%	0,00%	0,06%
Lesiones personales físicas	0,03%	0,00%	0,02%
Pérdida de bienes e inmuebles	0,00%	0,00%	0,04%
Secuestro	0,02%	0,00%	0,00%
Víctima	**66,47%**	**46,04%**	**64,73%**

Tipo de desplazamiento	Clúster 1	Clúster 2	Clúster 3
Víctima de minas	0,35%	0,22%	0,70%
Vinculación de niños	0,00%	0,00%	0,02%
Abandono o despojo	0,66%	2,64%	3,31%
Otros	0,30%	1,76%	1,12%
Sin información	0,28%	0,88%	2,03%

Como la Tabla 12 lo puede evidenciar, es una constante para todos los clústeres que se distribuyan en dos categorías mayoritariamente: víctima y en situación de desplazamiento. Por esta razón, resulta pertinente presentar una tabla que sintetice la información de la siguiente manera.

Tabla 13
Resumen de distribución por tipo de desplazamiento

Tipo de desplazamiento	Clúster 1	Clúster 2	Clúster 3
En situación de desplazamiento	26,08%	39,32%	19,48%
Víctima	66,47%	46,04%	64,73%
Otros	7,45%	14,65%	15,78%

De esta manera, se puede observar que la condición de víctima es mayoritaria para todos los clústeres, aunque para el segundo en menor proporción.

Con lo anterior, se culmina la caracterización descriptiva de los clústeres resultantes del algoritmo k-means, resaltando las particularidades que resalta en cada grupo en términos sociodemográficos y educativos. No obstante, dado que este método parte de un grupo predefinido de clústeres, resulta importante analizar los resultados mediante el otro método, DBSCAN, que los define según la forma y densidad de la naturaleza de los datos, lo cuál permite comparar los hallazgos y darle mayor fuerza y robustez al posterior análisis. Por ello, a continuación, se presenta el análisis descriptivo de lo hablado, siguiendo el mismo esquema utilizado para facilitar la comparación entre ambos.

8.1.2 DBSCAN

Para iniciar con la caracterización de los clústeres identificados, en primera instancia se debe conocer datos básicos como el promedio de edad y grado, así como otros estadísticos importantes. Esto permite ubicar a los estudiantes dentro de su trayectoria académica, así como el estado de maduración educativa en la que se encuentran.

Tabla 14
Estadísticos descriptivos de edad y grado por clúster (DBSCAN)

Variable	Estadístico	Clúster 1	Clúster 2	Clúster 3
Edad	Media (años)	12,12	18,09	17,44
	Desviación estándar	3,2	2,17	1,39
	Q1-Q2-Q3	10-12-15	17-18-19	16-17-19
Grado	Media	6,07	24,79	23,9
	Desviación estándar	2,9	1,04	0,75

Variable	Estadístico	Clúster 1	Clúster 2	Clúster 3
	Q1-Q2-Q3	4-6-8	24-25-26	23-24-24

En la Tabla 14 se presentan los resultados descriptivos de las variables edad y grado para los clústeres que se identificaron mediante el método DBSCAN. El primer grupo está conformado por estudiantes con una edad media de 12,12 años y un grado promedio de 6, por lo que es correcto afirmar que se encuentran en la transición de básica primaria a bachillerato. En contraste, los clústeres dos y tres muestran un promedio de 18 y 17,4 años respectivamente, y dado que su grado promedio es 23,9 son estudiantes que posiblemente ya pasaron por la media vocacional (grado 11).

Otra variable importante de analizar es la proporción de género dentro de los clústeres. Esta variable permite identificar si hay diferencia significativa en la representación de hombres y mujeres dentro de los grupos, lo cual fortalece el perfil sociodemográfico de los estudiantes.

Tabla 15
Distribución por género (DBSCAN)

Género	Clúster 1	Clúster 2	Clúster 3
Masculino	0,5228	0,5339	0,5417
Femenino	0,4772	0,4661	0,4583

De esta manera se puede reconocer que, no se presenta diferencia significativa en la proporción de género de los estudiados, aunque se presenta una mayoría masculina en cada uno de los clústeres.

Continuando con la caracterización se presenta la dimensión geográfica, que muestra la distribución según las provincias del departamento de Santander. Este aspecto es relevante para identificar zonas con mayor representación y, al mismo tiempo, reconocer si existen patrones territoriales en las víctimas del conflicto, específicamente los desplazados por la violencia para este caso.

Tabla 16
Distribuciones por provincia (DBSCAN)

Provincia	Clúster 1	Clúster 2	Clúster 3
García rovira	5,00%	1,19%	**29,17%**
Guanenta	9,49%	**23,39%**	16,67%
Metropolitana	**18,46%**	15,25%	2,78%
Soto norte	4,22%	0,17%	0,00%
Vélez	**22,07%**	13,90%	6,94%
Yariguies	**30,82%**	**20,68%**	**23,61%**
Comuneros	9,94%	**25,42%**	**20,83%**

La Tabla 16 muestra que una provincia predominante para todos los clústeres es la Yariguies, así como la comuneros para los clústeres 2 y 3. Además, se destacan tanto Vélez como la metropolitana también para el clúster 1, la Guanentá para el segundo y la García Rovira para el tercero.

Por otro lado, las variables carácter y zona de ubicación de la institución educativa forman un aspecto importante en la caracterización. Estas permiten contextualizar de qué manera los estudiantes reciben las clases y las condiciones que se presentan para que ellos puedan acceder sin problema, ofreciendo una visión más clara de las diferencias que pueden existir entre los grupos.

Tabla 17
Distribución por carácter y zona (%) (DBSCAN)

Variable	Clúster 1	Clúster 2	Clúster 3
Institución pública	96,86%	35,25%	0,00%
Institución privada	3,14%	64,75%	100,00%
Zona urbana	47,46%	86,10%	0,00%
Zona rural	52,54%	13,90%	100,00%

Los resultados muestran contrastes significativos entre las variables presentadas en la Tabla 17. El clúster 1 casi concentra su totalidad en la institución pública, mientras se ve equilibrado entre la zona urbana y rural. Por su parte, el clúster 2 y 3 son más que todo de carácter privado, aunque presentan diferencias en la zona; el segundo es más en el casco urbano y el tercero en su totalidad pertenece al rural.

Finalizando lo relacionado al entorno educativo es importante tener en cuenta las proporciones del método de enseñanza que se presentan en cada uno de los grupos.

Tabla 18
Distribución por método de enseñanza (%) (DBSCAN)

Método de enseñanza	Clúster 1	Clúster 2	Clúster 3
Educación tradicional	67,14%	0,00%	0,00%
Escuela nueva	20,74%	0,00%	0,00%
Media rural	2,83%	0,00%	0,00%
Post primaria	9,21%	0,00%	0,00%
Programa para jóvenes	0,00%	100,00%	0,00%
SAT	0,00%	0,00%	100,00%
Tejiendo saberes	0,00%	0,00%	0,00%
UNAD	0,00%	0,00%	0,00%
Caminar en secundaria	0,08%	0,00%	0,00%

Esta información muestra que, para el primer cluster predominan los métodos: educación tradicional, escuela nueva y post primaria, mientras que programas para jóvenes totaliza la proporción del segundo cluster así como SAT para el tercero.

Por último, acabando la caracterización por parte del método DBSCAN, es necesario presentar las distribuciones de la variable tipo de desplazamiento, pues es lo que dará sentido a la parte del análisis de carácter social que se busca en el presente trabajo.

Tabla 19
Distribución por tipo de desplazamiento (%) (DBSCAN)

Tipo de desplazamiento	Clúster 1	Clúster 2	Clúster 3
Actos terroristas	0,04%	0,00%	0,00%
Amenazado	0,12%	0,00%	0,00%
Comfinamiento	0,00%	0,00%	0,00%
Delitos contra la libertad	0,05%	0,00%	0,00%
Desaparición forzada	0,33%	0,34%	0,00%
Desplazamiento forzado	7,40%	4,58%	2,78%
Desvinculados de grupos armados	0,07%	0,17%	0,00%
En situación de desplazamiento	23,06%	44,07%	40,28%
Hijos de adultos desmovilizados	0,61%	1,53%	0,00%
Homicidio	0,06%	0,00%	0,00%
Lesiones personales físicas	0,02%	0,00%	0,00%
Pérdida de bienes e inmuebles	0,02%	0,00%	0,00%
Secuestro	0,01%	0,00%	0,00%
Víctima	66,15%	49,32%	56,94%
Víctima de minas	0,39%	0,00%	0,00%
Vinculación de niños	0,01%	0,00%	0,00%

Tipo de desplazamiento	Clúster 1	Clúster 2	Clúster 3
Abandono o despojo	0,12%	0,00%	0,00%
Otros	0,57%	0,00%	0,00%
Sin información	0,99%	0,00%	0,00%

O bien, dado que mayoritariamente para todos los clústeres están las condiciones de víctima y situación de desplazamiento, se sintetiza la información de la Tabla 19 en la

Tabla 20
Resumen de distribución por tipo de desplazamiento (%) (DBSCAN)

Tipo de desplazamiento	Clúster 1	Clúster 2	Clúster 3
En situación de desplazamiento	23,06%	44,07%	40,28%
Víctima	66,15%	49,32%	56,94%
Otros	10,79%	6,61%	2,78%

La caracterización presentada hasta el momento para ambos resultados conforma un acercamiento al entendimiento de los estudiantes en condición de desplazamiento. No obstante, esta descripción por sí sola no es suficiente, puesto que se requiere un análisis interpretativo que permita dotar de sentido los hallazgos y conectarlos con la realidad social y educativa del contexto estudiado. Por lo tanto, en la siguiente sección se realizará un análisis comparativo y reflexivo de los resultados.

8.2 Análisis de los resultados

Una vez descritas las principales características sociodemográficas y educativas de la población estudiantil en condición de desplazamiento en Santander, el siguiente paso es interpretar los resultados obtenidos mediante los métodos de agrupamiento. En esta sección, el análisis se orienta a comprender lo que representa cada cluster en términos de perfiles estudiantiles, resaltando sus particularidades, vulnerabilidades y contrastes internos. Se inicia con k-means para luego terminar con DBSCAN.

8.2.1 K-means

El primer cluster identifica principalmente a estudiantes adolescentes en básica secundaria, con una edad promedio de 14,6 años y ubicados en promedio en octavo grado (8,29). Se trata de un grupo relativamente homogéneo en términos de edad y grado, lo que indica trayectorias más estables que otros clústeres. Estos se encuentran ubicados en su mayoría en tres provincias: Yariguies (27,99%), Vélez (23,35%) y Metropolitana (18,20%).

Un rasgo central es su marcada pertenencia al sistema público de educación, puesto que el 97,68% de los observados estudian en instituciones oficiales, y casi la mitad proviene de escuelas rurales (48,87%). Esto sugiere estudiantes que dependen fuertemente de la oferta estatal, probablemente en municipios con gran parte rural, donde la cobertura educativa privada es reducida.

Sobre la jornada académica, predomina la de la mañana (67,24%) y la única (22,46%), reflejando así las jornadas tradicionales del sistema educativo colombiano. Esto se refuerza al ver que el método de enseñanza predominante para este cluster es el tradicional (77,23%), aunque también la presencia de la post primaria (15,49%) muestra una proporción significativa. Este

último método de enseñanza se caracteriza por ser flexible y multigrado (presencia de múltiples grados en un mismo salón) dirigido a estudiantes rurales entre 6° y 9° en veredas o corregimientos con baja matrícula, lo que confirma la relevancia de este clúster en contextos semiurbanos.

Respecto al tipo de desplazamiento, la mayoría de los estudiantes son catalogados como víctimas (66,47%) mientras que un 26,08% se encuentra en situación de desplazamiento. Aquí, es necesario hacer una precisión conceptual: aunque es un hecho que todos los estudiantes observados han sido desplazados seguramente por distintos motivos, la base de datos distingue únicamente dos categorías (las ya mencionadas), a pesar de que existen diecinueve formas diferentes de clasificar esta condición. Esta simplificación, probablemente generada durante el proceso de recolección de los datos por parte de la entidad gubernamental pertinente, genera cierta ambigüedad y limita la riqueza de los hallazgos.

Aun así, la diferencia entre las dos resulta significativa. Según la Corte Constitucional de Colombia (2013) ser clasificado como víctima implica que el daño sufrido está relacionado directamente en actos del conflicto armado y necesita probarse formalmente, mientras que la categoría “situación de desplazamiento” se asocia a cualquier manera de violencia que haya generado desarraigo y esta no necesariamente debe pertenecer explícitamente a estos hechos. Esto traduce a que, aproximadamente, por cada 10 personas pertenecientes a este clúster, 6 fueron víctimas del conflicto armado, 3 fueron desplazados por hechos violentos ajenos a este (como, por ejemplo, acciones de bandas criminales) y 1 fue catalogado por otro tipo de desplazamiento.

En resumen, este clúster representa un perfil típico de estudiante en condición de desplazamiento en secundaria pública, que, si bien cuenta con cierta estabilidad educativa, depende arraigadamente de la oferta oficial. Su relevancia radica en que concentra a una parte importante de la población (puesto que este grupo alberga aproximadamente el 50% de los estudiados), que

pese a estar insertada en el sistema educativo, requiere apoyos adicionales para evitar la deserción, especialmente en territorios rurales donde las opciones son limitadas.

Por su parte, el segundo cluster está compuesto principalmente por estudiantes de mayor edad, en promedio, 21 años, mientras que el grado promedio es de 24,8. Cabe resaltar que, aunque el sistema educativo tradicional colombiano contempla hasta grado 11, en la base de datos que se utilizó aparecen valores superiores. Esto se interpreta como una representación de estudiantes que, por su edad y método de aprendizaje, ya se encuentran vinculados a procesos de educación superior, bien sea en programas técnicos, tecnológicos y/o profesionales. Este perfil contrasta con los otros dos grupos que arrojó k-means, dado que refleja jóvenes que probablemente han tenido trayectorias más largas prolongadas y estables, pero también muestra algunos que retomaron sus estudios en edades no convencionales (pues se encuentran personas de hasta 69 años, aunque no represente al grueso de la muestra). Adicionalmente, respecto a las provincias donde mayormente se distribuyen los estudiantes, están las provincias de Comuneros (23,90%), Guanentá (20,04%) y Yariguies (21,37%), siendo la última la única similitud frente al anterior grupo.

En cuanto a las instituciones educativas a las que asisten estas personas, predominan las de carácter privado (64,78%) que se ubican en zonas urbanas (69,16%); para explicar esto, hay que mencionar que los métodos de aprendizaje más representativos son no tradicionales, como programa para jóvenes (79,52%) y sistema de aprendizaje tutorial o SAT (17,51%). El primero suele estar asociado a estrategias de acceso y permanencia en la educación superior, en forma de becas, subsidios o convenios con universidades privadas o instituciones que ofrecen programas técnicos/tecnológicos. Por otra parte, el segundo método (SAT) es un modelo flexible que busca ampliar cobertura en comunidades rurales, aunque se atribuyen más a la secundaria y media vocacional. Esto refuerza la idea de que en el clúster 2 concentra a estudiantes que enfrentaron

dificultades o incluso interrupciones en su trayectoria escolar, pero que encontraron alternativas institucionales para continuar con su formación.

Respecto a la jornada, se observa una fuerte representación en modalidades también alternativas como fines de semana (42,07%) y nocturna (28,08%), lo que indica que una gran parte de estos posiblemente combina sus estudios con actividades laborales u otras responsabilidades, reforzando así el argumento anteriormente mencionado.

Por último, en lo referente al tipo de desplazamiento, se observa que, aunque la mayoría se clasifica como víctima (46,04%), también hay una proporción considerable en situación de desplazamiento (39,32%) y un porcentaje no menor en “otros” (14,65%). Esta distribución es más heterogénea que en los demás clústeres, por lo que se interpreta como una mayor diversidad en las experiencias de movilidad forzada.

Para terminar con los grupos generados por K-means, está el clúster 3. Este agrupa a los de menor edad, con una media de 9,2 años y una distribución muy concentrada entre los 8 y 11 años, además de un grado promedio de 3,45, ubicándolos así en básica primaria. La baja desviación estándar (1,73 años), en este caso muestra que se trata de un grupo bastante homogéneo y en pleno inicio de su etapa escolar. En términos de ubicación geográfica, este grupo se distribuye mayoritariamente en provincias como Yariguies (33,70%), que termina siendo un factor común en todos los clústeres, Vélez (20,63%) y Metropolitana (18,73%), al igual que el primero.

En términos de acceso educativo, la gran mayoría asiste a instituciones públicas, lo que al igual que el primero, demuestra una gran dependencia de la oferta estatal. Además, más de la mitad reside en zonas rurales (59,68%), lo que contrasta al cluster 2 que era más orientado a lo urbano y

también coincide con las condiciones zonales del primero. Esto los posiciona como una población en contextos de mayor vulnerabilidad estructural.

Respecto a la jornada, predomina la mañana (72,93%), seguida por tarde (11,93%) y única (11,90%). La ausencia de jornadas alternativas es consistente con su edad, ya que son estudiantes que aún dependen de esquemas escolares tradicionales.

En la dimensión de método de aprendizaje se observa la educación tradicional siendo mayoría (55,19%) junto con nueva escuela (42,19%), este último siendo similar a post-primaria (flexibilizan la educación mediante guías para aprendizaje autónomo y salones multigrados, utilizado en zonas rurales donde la demanda no es suficiente como para métodos más tradicionales) con la diferencia que es enfocado 100% en la básica primaria. Esta distribución refleja un equilibrio entre prácticas pedagógicas tradicionales con alternativas diseñadas para responder a diferentes realidades educativas.

Finalmente, en cuanto al tipo de desplazamiento sobresale víctima (64,73%), seguida por la situación de desplazamiento (19,48%) y “otros” (15,78%). Aunque las proporciones son similares con el primer cluster, esta vez llaman la atención por tratarse de niños en etapas tempranas de escolaridad, lo que sugiere que la problemática social tratada impacta incluso a la población infantil, condicionando sus trayectorias desde un principio.

8.2.2 DBSCAN

En esta sección de presentan los resultados obtenidos mediante el método DBSCAN, el cual consta en la separación de 3 clústeres, aunque con la diferencia frente a k-means de que un pequeño grupo de observaciones (2,78% del total) son consideradas valores atípicos. Estos hallazgos ofrecen una visión complementaria de la división hecha por el anterior método y aportan

matices relevantes en pro de caracterizar a los estudiantes en condición de desplazamiento en Santander.

El primer cluster concentra a estudiantes que se encuentran cursando su educación básica, con una edad media de 12 años y un grado promedio de 6°. Esto refleja que el grupo avanza en el sistema educativo respecto al rango esperado para su edad. Acá se ven representados tanto estudiantes de primaria como de bachillerato.

De la misma manera, en su mayoría se trata de estudiantes pertenecientes a instituciones públicas (96,86%), con una distribución equilibrada entre zona rural y urbana. Y si bien, el 67,14% de los observados pertenecen a la escuela tradicional, hay una proporción entre la escuela nueva y la post primaria que no puede ser ignorada, siendo del 20,74% y 9,21% respectivamente; pues sugiere que una buena parte de este grupo cursa sus estudios en territorios donde la cobertura estructural y tradicional no se puede garantizar por falta de densidad poblacional o territorios de baja demanda estudiantil.

En cuanto a la organización del tiempo escolar, la distribución la conforma tres jornadas diferentes pero comunes en las condiciones descritas: única (17,53%), mañana (70,07%) y tarde (9,29%).

Respecto al tipo de desplazamiento, más de la mitad se encuentra catalogada como víctima (66,15%), mientras que el 23,06% se ubica en la categoría condición de desplazamiento. Esto muestra que, aunque todos comparten la situación de movilidad forzada, la mayoría presenta causas más cercanas al conflicto armado.

Este cluster es representativo para los estudiantes observados de las provincias Yariguies, Vélez y Metropolitana, y en conjunto, revela un perfil en el que aparentemente se hace una correcta

transición entre la básica primaria y secundaria, con predominio en instituciones públicas rurales y bajo metodologías adaptadas a contextos de dispersión, pero sin perder el foco en la educación tradicional.

Para el caso del clúster 2 se encuentran contrastes pronunciados con respecto al anterior. Para empezar, los estudiantes reunidos aquí cuentan, en promedio, con 18 años; además, se muestran con un gran avance en el sistema educativo, con un grado promedio de 24,79, que como se había explicado antes, refleja grados de permanencia en la educación superior, bien sea técnicas, tecnológicas o pregrados.

A nivel institucional, predomina la matrícula en establecimientos privados (64,75%) que, lejos de mostrar una mayor ventaja socioeconómica, parece estar relacionado con la participación de programas de inclusión como el programa para jóvenes, puesto que concentra al 100% de los casos en este cluster conformado por 590 estudiantes. Se trata de iniciativas, que, aunque se desarrollan mediante las instituciones privadas y urbanas (86,1%), funcionan como políticas de compensación para quienes fueron víctimas del conflicto armado, que, para este caso, representa al 49.32% de la muestra.

Ahora bien, no se puede afirmar que se trata de un grupo tradicional de educación superior, pues las jornadas en su mayoría son durante los fines de semana (44,92%) y nocturnas (32,37%), lo cuál se puede interpretar como una aparente vulnerabilidad económica, que en un futuro puede desenlazar en deserción estudiantil o simplemente que, en general, con jóvenes que no están en la misma dinámica que quienes transitan la escolaridad tradicional.

En síntesis, este cluster refleja un grupo de jóvenes, mayoritariamente de las provincias Comuneros (25,42%), Guanentá (23,39%) y Yariguies (20,68%) que han enfrentado una doble

vulnerabilidad: el desplazamiento forzado bien sea en condición de víctima (49,32%) o en situación (44,07%) y rezago educativo en sus etapas regulares. Su presencia en programas compensatorios es una muestra de resiliencia y de las oportunidades que el sistema ofrece, pero también un recordatorio de las consecuencias que provoque no mantener este tipo de políticas de inclusión durante el tiempo.

El clúster 3, por su parte, reúne a un grupo de 72 estudiantes, lo cuál se hace muy reducido, pero también muy específico en cuanto a sus características. Su edad media es de 17 años y su grado promedio también demuestra su transición hacia niveles avanzados. Sin embargo, lo más llamativo en este clúster es la combinación de instituciones 100% rurales pero privadas, una combinación poco común en el sistema educativo colombiano, donde la ruralidad suele asociarse con la oferta pública. A esto se suma la distribución de jornadas, pues mientras la mayoría de los clústeres se distribuyen en jornadas tradicionales como en la mañana, en este prevalece la jornada completa (73,61%) y fin de semana (26,39%). Sin embargo, esto se aclara al ver que el 100% de los estudiantes se educan mediante el sistema de aprendizaje tutorial (SAT). Este modelo, fue creado por la fundación FUNDAEC con el aval del ministerio de educación, diseñado específicamente para comunidades rurales dispersas. Aunque se gestiona como carácter privado, funciona como un mecanismo oficial para garantizar el acceso a la educación donde el acceso a instituciones convencionales es limitado.

Esto se refuerza con la presencia de provincias con alta ruralidad, como García Rovira (29,17%), Comuneros (20,83%) y Yariguies (23,61%), además de la casi nula representación de provincias mayormente urbanas como la Metropolitana (2,78%).

En cuanto al tipo de desplazamiento, la mayoría de los estudiantes fueron reconocidos como víctimas (56,94%) aunque la proporción con respecto a aquellos en situación de desplazamiento no es muy alejada (40,28%).

Las vulnerabilidades de este grupo no se limitan únicamente al desplazamiento que sufrieron, sino también enfrentan condiciones de aislamiento geográfico, dependencia de un único modelo pedagógico (SAT) y posibles restricciones económicas asociadas al carácter privado de la oferta, lo cual los expone a riesgos de continuidad educativa si se reducen apoyos institucionales o si el modelo enfrenta limitaciones en la cobertura.

En síntesis, el clúster 3 representa a estudiantes que logran avanzar en su trayectoria escolar gracias a un sistema alternativo que responde a su contexto rural, pero que acumula un conjunto de vulnerabilidades que hacen que su situación sea delicada.

8.3 Discusión

La principal diferencia entre ambos algoritmos es, desde un punto de vista metodológico, el balance de los grupos obtenidos. Con el uso de DBSCAN, fue posible identificar perfiles muy particulares y bien determinados, como estudiantes vinculados al método de aprendizaje SAT, lo que indica la gran sensibilidad del método para identificar ciertas concentraciones de datos. Pero con detalles tan cercanos resultan clústeres altamente desequilibrados: uno enfocado en más de 11,000 estudiantes y el otro con un poco más de 72, lo que no ofrece una buena comparación objetiva. En contraste, k-means generó agrupaciones con tamaños más proporcionales, resultando en una correspondencia más cercana entre clústeres y, por lo tanto, un mejor análisis comparativo.

Por lo tanto, se eligen los resultados de k-means como base del análisis principal. Esta decisión está respaldada por el hecho de que el método logró generar clústeres con proporciones

más equilibradas y perfiles diferenciados que contribuyen positivamente desde una descripción estadística como desde la comprensión de las vulnerabilidades educativas de la población estudiada. DBSCAN fue interesante porque añadió matices al señalar los extremos, pero no tiene el mismo nivel de claridad, interpretación ni de una implementación mayor en cuanto a representatividad para determinar conclusiones generales. Por lo tanto, se decide que k-means es la mejor estrategia para justificar los hallazgos y recomendaciones del estudio.

De esta manera, se concluye el análisis de los resultados hablando respecto a las vulnerabilidades que presenta cada clúster.

Para el primero, la vulnerabilidad más marcada proviene de la presión socioeconómica que pueden llegar a sentir. Muchos adolescentes, aún sin tener la edad legal para empezar a trabajar, se ven obligados a insertarse en oficios informales o labores que demandan fuerza física para ayudar a sostener a sus familias económicamente. Esta situación los expone a situaciones de riesgo, como la delincuencia o economías ilegales que aparecen como alternativas de rápidos ingresos. De este modo, el abandono escolar en este grupo no responde solo a factores sociales como el desplazamiento sino a la tensión de continuar estudiando y generar ingresos inmediatos para sobrevivir.

Para el caso del segundo se presentan diferentes matices. Por un lado, sigue estando presente las necesidades económicas que se traduce en lograr un equilibrio entre estudiar y subsistir con urgencia, combinando trabajo con sus responsabilidades académicas. A esa condición, se suma la dificultad de acceder a las oportunidades (recordando que este clúster depende en su mayoría del carácter privado, mediante auxilios o ayudas gubernamentales), pues la demanda por cupos y becas suele superar a la oferta disponible, sumado a procesos donde no siempre priorizan a quienes más necesitan y facilita estos recursos a personas con mayor influencia. Por último, ciertas barreras

mentales que poseen los estudiantes que pierden continuidad escolar por falta de ritmo académico o bajos rendimiento, que en algunos casos deriva a percepciones desmotivadoras que se traduce en deserción escolar. En conjunto, esto demuestra que, para este grupo, se debe combinar acceso con acompañamiento académico y socioeconómico para evitar que la entrada a oportunidades se convierta en un punto de fuga más que en una motivación para que puedan superarse a nivel profesional.

Para finalizar, la vulnerabilidad principal del tercer clúster está en la fragilidad de la oferta educativa en zonas apartadas. Aunque el método de nueva escuela está dirigido para remendar y adaptarse a estas situaciones, la falta de recursos, distancias extensas (sobre todo si se piensa en que sobre todo son niños de temprana edad quienes deben recorrerlas) y la dispersión poblacional dificultan la continuidad. Además, si bien en estas edades, la responsabilidad económica no recae sobre ellos, sí lo hace sobre sus familias, quienes pueden priorizar la subsistencia inmediata frente a la educación de sus hijos.

9. Conclusiones

El presente trabajo tuvo como propósito caracterizar a los estudiantes en condición de desplazamiento en Santander mediante técnicas de clustering, con el fin de identificar características comunes y diferencias entre los grupos de esta población. Tras una revisión de literatura para definir los métodos con los que se llevaran a cabo esta segmentación, se seleccionaron K-means y DBSCAN como algoritmos principales para su aplicación.

Metodológicamente, la comparación mostró que ambos métodos coinciden en que el mejor agrupamiento consta de tres grupos, aunque muestra también diferencias sustanciales en cómo los separa, por un lado, DBSCAN permitió identificar perfiles muy específicos, pero con resultados altamente desbalanceados que limitaron su utilidad práctica. En contraste, K-means generó clústeres más equilibrados aún generando perfiles diferenciados, lo que permitió una interpretación más clara de las trayectorias educativas. Por ello, los resultados de este último se tomaron como la base principal del análisis, utilizando los de DBSCAN solamente como complemento.

Los clústeres obtenidos mediante K-means, no solo reflejaron diferencias demográficas y educativas sino también vulnerabilidades específicas que atraviesan la experiencia escolar de los estudiantes en condición de desplazamiento. El primer grupo lo conforman adolescentes que enfrentan presiones socioeconómicas que los pueden enfrentar al trabajo informal, lo que aumenta su riesgo de abandono escolar. El segundo agrupa a los que tienen un mayor recorrido en el ámbito académico pero cuya permanencia se puede ver comprometida por la dificultad de encontrar un equilibrio entre estudio y subsistencia, la falta de oportunidades o injusta competencia en acceso a becas o apoyos y la desmotivación derivada de la falta de ritmo educativo. Finalmente, el tercero corresponde a niños en gran parte que habitan en zonas rurales y apartadas, cuya fragilidad radica

en la limitada oferta educativa, la dispersión geográfica y las condiciones que sus familias puedan presentar. En conjunto, estos hallazgos muestran que las vulnerabilidades que presentan desde el enfoque académico responden a la interacción de factores socioeconómicos, institucionales y territoriales que requieren de estrategias diferenciadas para promover la permanencia educativa.

Finalmente, este trabajo demuestra que el uso de técnicas de clustering aporta una mirada novedosa para la caracterización de este tipo de poblaciones vulnerables, al permitir segmentar perfiles y reconocer sus necesidades diferenciadas. De esta manera, los resultados aquí expuestos pueden servir como punto de partida para orientar políticas públicas y programas educativos que fortalezcan la permanencia escolar de los estudiantes en condición de desplazamiento en Santander.

10. Recomendaciones

En primer lugar, es importante señalar que en ese trabajo no se pudo diferenciar de una manera óptima de la variable: tipo de desplazamiento. Si bien se presentaban diecinueve diferentes categorías que describían esta variable, durante la toma de los datos parece que solo se consideraron dos, víctima y condición de desplazamiento. Para futuros estudios, se recomienda que esta variable pueda describir todos los matices disponibles, con el fin de tener un análisis más detallado desde el punto de vista social.

De igual manera, se sugiere complementar la base de datos agregando aquellas personas que desertaron y no regresaron al sistema educativo. Su caracterización puede aportar una visión más clara del problema, permitiendo identificar factores determinantes de la deserción y así poder diseñar estrategias de prevención.

Otro aspecto por considerar en futuras investigaciones es el uso de modelos predictivos que, a partir de diferentes variables sociodemográficas, económicas, académicas, entre otras, permitan anticipar aquellos estudiantes que tengan una mayor probabilidad de abandonar sus estudios. Este enfoque fortalecería la utilidad práctica de los hallazgos, pero también facilitaría la creación de sistemas de alerta temprana.

Finalmente, sería pertinente evaluar la aplicación de otras técnicas de clustering con el fin de contrastar los resultados y validar la robustez de las segmentaciones obtenidas en este trabajo.

Referencias bibliográficas

Abu Baker, N. N., & Daradkeh, S. M. (2010). Prevalence of overweight and obesity among adolescents in Irbid governorate, Jordan. *Eastern Mediterranean Health Journal = La Revue de Sante de La Mediterranee Orientale = Al-Majallah al-Sihhiyah Li-Sharq al-Mutawassit*, *16*(6), 657–662. http://europepmc.org/abstract/MED/20799595

Akmatov, M. K. (2011). Child abuse in 28 developing and transitional countries—results from the Multiple Indicator Cluster Surveys. *International Journal of Epidemiology*, *40*(1), 219–227. https://doi.org/10.1093/ije/dyq168

al Arab, G. E., Tawfik, N., El Gendy, R., Anwar, W., & Courtright, P. (2001). The burden of trachoma in the rural Nile Delta of Egypt: a survey of Menofiya governorate. *British Journal of Ophthalmology*, *85*(12), 1406. https://doi.org/10.1136/bjo.85.12.1406

Anuradha, C., Velmurugan, T., Anandavally, R., & Professor, A. (2015). Clustering Algorithms in Educational Data Mining: A Review…C.Anuradha et al., CLUSTERING ALGORITHMS IN EDUCATIONAL DATA MINING: A REVIEW. In *International Journal of Power Control and Computation(IJPCSC)* (Vol. 7, Issue 1). www.ijcns.com

Aristizábal, D., González, G., Fredy Suárez, J., & Roldán, P. (2012). Factores asociados al trauma fatal en motociclistas en Medellín, 2005-2008. In *Biomédica* (Vol. 32).

Aydin, A., Ergor, A., & Ozkan, H. (2008). Effects of sociodemographic factors on febrile convulsion prevalence. *Pediatrics International*, *50*(2), 216–220. https://doi.org/https://doi.org/10.1111/j.1442-200X.2008.02562.x

Azage, M., & Haile, D. (2015). Factors affecting healthcare service utilization of mothers who had children with diarrhea in Ethiopia: evidence from a population based national survey. In *Rural and Remote Health* (Vol. 15).

Azizi, F., Rahmani, M., Emami, H., Mirmiran, P., Hajipour, R., Madjid, M., Ghanbili, J., Ghanbarian, A., Mehrabi, J., Saadat, N., Salehi, P., Mortazavi, N., Heydarian, P., Sarbazi, N., Allahverdian, S., Saadati, N., Ainy, E., & Mœini, S. (2002). Cardiovascular risk factors in an Iranian urban population: Tehran Lipid and Glucose Study (Phase 1). *Sozial- Und Präventivmedizin*, *47*(6), 408–426. https://doi.org/10.1007/s000380200008

Benavidez Robles, K. J., Jiménez Wandurraga, L. B. T., Lamos Díaz, H., & Puentes Garzón, D. E. (2022). *Un algoritmo de clustering difuso para el análisis de las causas que afectan el desarrollo, productividad y competitividad de la industria de calzado, cuero y marroquinería con relación a su producción y mano de obra* [Universidad Industrial de Santander]. https://noesis.uis.edu.co/server/api/core/bitstreams/1f470576-d1fd-4140-837a-23489dbe94d1/content

Caicedo, M. R. A., De Arruda Xavier, D., Caicedo, C. A. A., Andrade, E., & Abel, I. (2019). Epidemiological scenarios for human rabies exposure notified in Colombia during ten years: A challenge to implement surveillance actions with a differential approach on vulnerable populations. *PLoS ONE*, *14*(12). https://doi.org/10.1371/journal.pone.0213120

Chami, G. F., Kabatereine, N. B., Tukahebwa, E. M., & Dunne, D. W. (2018). Precision global health and comorbidity: A population-based study of 16 357 people in rural Uganda. *Journal of the Royal Society Interface*, *15*(147). https://doi.org/10.1098/rsif.2018.0248

Charte, F. (2020, November 17). *Cómo es el proceso de extraer conocimiento a partir de bases de datos*. Campus MVP. https://www.campusmvp.es/recursos/post/el-proceso-de-extraccion-de-conocimiento-a-partir-de-bases-de-datos.aspx

Coker, E., Katamba, A., Kizito, S., Eskenazi, B., & Davis, J. L. (2020). Household air pollution profiles associated with persistent childhood cough in urban Uganda. *Environment International*, *136*. https://doi.org/10.1016/j.envint.2020.105471

Corte Constitucional de Colombia. (2013, June 24). *Auto 119 de 2013*. https://www.corteconstitucional.gov.co/relatoria/autos/2013/a119-13.htm

Dagne, S., Menber, Y., Petrucka, P., & Wassihun, Y. (2021). Prevalence and associated factors of abdominal obesity among the adult population in Woldia town, Northeast Ethiopia, 2020: Community-based cross-sectional study. *PLoS ONE*, *16*(3 March). https://doi.org/10.1371/journal.pone.0247960

Debu Liga, A., Erango Boyamo, A., & Negash Jabir, Y. (2023). Magnitude and Risk Factors Associated with Adolescent Pregnancy in Ethiopia: Risk factors associated with early childbearing among teenage girls. *Ethiopian Journal of Reproductive Health*, *15*(2). https://doi.org/10.69614/ejrh.v15i2.639

Dinya, E., Csorba, J., Sörfozo, Z., Steiner, P., Ficsor, B., & Horvath, A. (2009). Profiles of Suicidality and Clusters of Hungarian Adolescent Outpatients Suffering from Suicidal Behaviour. *Psychopathology*, *42*(5), 299–310. http://dx.doi.org/10.1159/000228839

El Kishawi, R. R., Soo, K. L., Abed, Y. A., & Muda, W. A. M. W. (2017). Prevalence and associated factors influencing stunting in children aged 2-5years in the Gaza Strip-Palestine: A cross-sectional study. *BMC Pediatrics*, *17*(1). https://doi.org/10.1186/s12887-017-0957-y

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. www.aaai.org

Eyler, L., Hubbard, A., & Juillard, C. (2016). Assessment of economic status in trauma registries: A new algorithm for generating population-specific clustering-based models of economic status for time-constrained low-resource settings. *International Journal of Medical Informatics*, *94*, 49–58. https://doi.org/10.1016/j.ijmedinf.2016.05.004

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From Data Mining to Knowledge Discovery in Databases; From Data Mining to Knowledge Discovery in Databases*. *17*. https://doi.org/10.1609/aimag.v17i3.1230

Gabster, A., Mayaud, P., Ortiz, A., Castillo, J., Castillero, O., Martínez, A., López, A., Aizprúa, B., Pitano, S., Murillo, A., & Pascale, J. M. (2021). Prevalence and determinants of genital Chlamydia trachomatis among school-going, sexually experienced adolescents in urban and rural Indigenous regions of Panama. *Sexually Transmitted Infections*, *97*(4), 304–311. https://doi.org/10.1136/sextrans-2019-054395

Gobernación de Santander. (2022, August 31). *ESTUDIANTES EN SITUACIÓN DE DESPLAZAMIENTO EN SANTANDER*. Datos Abiertos. https://www.datos.gov.co/Educaci-n/ESTUDIANTES-EN-SITUACI-N-DE-DESPLAZAMIENTO-EN-SANT/wemp-b8gd/about_data

Granados, Y., Gastelum Strozzi, A., Alvarez-Nemegyei, J., Quintana, R., Julian-Santiago, F., Santos, A. M., Guevara-Pacheco, S., Loyola-Sanchez, A., Goycochea-Robles, M. V., Juarez, V., Garza-Elizondo, M. A., Rueda, J. C., Burgos-Vargas, R., Londoño, J., Pons-Estel, B. A., & Pelaez-Ballestas, I. (2023). Inequity and vulnerability in Latin American Indigenous and

non-Indigenous populations with rheumatic diseases: a syndemic approach. *BMJ Open*, *13*(3). https://doi.org/10.1136/bmjopen-2022-069246

Gray, C. L., Pence, B. W., Ostermann, J., Whetten, R. A., O'donnell, K., Thielman, N. M., & Whetten, K. (2015). *Prevalence and Incidence of Traumatic Experiences Among Orphans in Institutional and Family-Based Settings in 5 Low-and Middle-Income Countries: A Longitudinal Study*. www.ghspjournal.org

Guerra Lozano, A. V., Díaz Mejía, E. L., González González, J. A., & León Rodríguez, A. M. (2021). *INCLUSIÓN DE LOS NIÑOS Y NIÑAS VICTIMAS DEL DESPLAZAMIENTO FORZADO EN EL CONTEXTO ESCOLAR* [Universidad Santo Tomás]. http://hdl.handle.net/11634/35053

Hernández Burgos, F. A., & Murillo Estepa, P. (2015). *LOS ESCENARIOS EDUCATIVOS INFORMALES COMO ESPACIOS DE INCLUSIÓN Y CALIDAD DE VIDA DE MENORES EN SITUACIÓN DE MARGINACIÓN Y DESPLAZAMIENTO* [Universidad de Sevilla]. http://hdl.handle.net/11441/34782

Holdsworth, J. (2024, June 28). *¿Qué es la minería de datos?* IBM. https://www.ibm.com/es-es/topics/data-mining

IBM. (2020, November 8). *¿Qué es el machine learning (ML)?* IBM. https://www.ibm.com/es-es/topics/machine-learning

IBM. (2023, December 8). *¿Qué es el análisis de componentes principales (PCA)?* IBM. https://www.ibm.com/es-es/think/topics/principal-component-analysis

Lamichhane, R., Zhao, Y., Paudel, S., & Adewuyi, E. O. (2017). Factors associated with infant mortality in Nepal: A comparative analysis of Nepal demographic and health surveys (NDHS) 2006 and 2011. *BMC Public Health*, *17*(1). https://doi.org/10.1186/s12889-016-3922-z

Mahumud, R. A., Sahle, B. W., Owusu-Addo, E., Chen, W., Morton, R. L., & Renzaho, A. M. N. (2021). Association of dietary intake, physical activity, and sedentary behaviours with overweight and obesity among 282,213 adolescents in 89 low and middle income to high-income countries. *International Journal of Obesity*, *45*(11), 2404–2418. https://doi.org/10.1038/s41366-021-00908-0

Masaku, J., Mutungi, F., Gichuki, P. M., Okoyo, C., Njomo, D. W., & Njenga, S. M. (2017). High prevalence of helminths infection and associated risk factors among adults living in a rural setting, central Kenya: A cross-sectional study. *Tropical Medicine and Health*, *45*(1). https://doi.org/10.1186/s41182-017-0055-8

Matias, S. L., Mridha, M. K., Young, R. T., Khan, M. S. A., Siddiqui, Z., Ullah, M. B., Vosti, S. A., & Dewey, K. G. (2018). Prenatal and postnatal supplementation with lipid-based nutrient supplements reduces anemia and iron deficiency in 18-month-old bangladeshi children: A cluster-randomized effectiveness trial. *Journal of Nutrition*, *148*(7), 1167–1176. https://doi.org/10.1093/jn/nxy078

Mebarak, M., Mendoza, J., Romero, D., & Amar, J. (2024). Healthy Life Habits in Caregivers of Children in Vulnerable Populations: A Cluster Analysis. *International Journal of Environmental Research and Public Health*, *21*(5). https://doi.org/10.3390/ijerph21050537

Muñoz Osorio, M. M., Pintón Mateus, M. S., & Lamos Diaz, H. (2013). *ESTUDIO DE SEGUIMIENTO A EGRESADOS POR MEDIO DE TÉCNICAS DE MINERÍA DE DATOS*

PARA LA UNIVERSIDAD INDUSTRIAL DE SANTANDER [Universidad Industrial de Santander]. https://noesis.uis.edu.co/server/api/core/bitstreams/116a36ef-ca5e-45a7-a154-c599d08f1d2b/content

Petro, J., Arango-Paternina, C. M., Lema-Gomez, L., Eusse-Lopez, C., Petro-Petro, J., Lopez-Sanchez, M., Watts-Fernandez, W., & Perea-Velasquez, F. (2022). Fitness, fatness, body movement, and diet in adolescents: clustering and associations with elevated blood pressure. *Journal of Sports Medicine and Physical Fitness*, *62*(7), 856–866. https://doi.org/10.23736/S0022-4707.21.12266-2

Pirkle, C. M., Wu, Y. Y., Zunzunegui, M. V., & Gómez, J. F. (2018). Model-based recursive partitioning to identify risk clusters for metabolic syndrome and its components: Findings from the International Mobility in Aging Study. *BMJ Open*, *8*(3). https://doi.org/10.1136/bmjopen-2017-018680

Rodríguez, M. (2023, June 22). *Método Elbow, para elegir el número ideal de clusters*. Platzi. https://platzi.com/tutoriales/2127-intro-algebra/11447-metodo-elbow-para-elegir-el-numero-ideal-de-clusters/

Rozi, S., & Akhtar, S. (2004). Smoking among high school adolescents in Karachi, Pakistan. *International Journal of Epidemiology*, *33*(3), 613–614. https://doi.org/10.1093/ije/dyh128

Ruger, J. P., & Kim, H.-J. (2006). Global health inequalities: an international comparison. *Journal of Epidemiology and Community Health*, *60*(11), 928. https://doi.org/10.1136/jech.2005.041954

Sanabria Ruiz, V. A., Lamos Díaz, H., & Martínez Quezada, D. O. (2017). *Aplicación de técnicas de agrupamiento (clustering) para el análisis estadístico de tendencias en twitter basado en*

el lenguaje de programación R. [Universidad Industrial de Santander]. https://noesis.uis.edu.co/server/api/core/bitstreams/993389cd-a69a-4116-923a-8843c12e7d98/content

Satty, A., Salih, M., Abdalla, F. A., Ashraf, A. F., Gumma, E. A. E., Saad Mohamed Khamis, G., Adam, A. M. A., Hassaballa, A. A., Hamed, O. M. A., & M. S. Mohammed, Z. (2024). Statistical Analysis of Factors Associated with Diarrhea in Yemeni Children under Five: Insights from the 2022–2023 Multiple Indicator Cluster Survey. *Journal of Epidemiology and Global Health*. https://doi.org/10.1007/s44197-024-00253-1

Schlick, C., Joachin, M., Briceño, L., Moraga, D., & Radon, K. (2014). Occupational injuries among children and adolescents in Cusco Province: A cross-sectional study. *BMC Public Health*, *14*(1). https://doi.org/10.1186/1471-2458-14-766

Scikit-learn. (2025). *DBSCAN*. https://scikit-learn.org/stable/modules/generated/dbscan-function.html

Semba, R. D., de Pee, S., Sun, K., Bloem, M. W., & Raju, V. K. (2008). Coverage of the National Vitamin A Supplementation Program in Ethiopia. *Journal of Tropical Pediatrics*, *54*(2), 141–144. https://doi.org/10.1093/tropej/fmm095

Shaheen, A. M., Hammad, S., Haourani, E. M., & Nassar, O. S. (2018). Factors Affecting Jordanian School Adolescents' Experience of Being Bullied. *Journal of Pediatric Nursing*, *38*, e66–e71. https://doi.org/10.1016/j.pedn.2017.09.003

Strozzi, A. G., Peláez-Ballestas, I., Granados, Y., Burgos-Vargas, R., Quintana, R., Londoño, J., Guevara, S., Vega-Hinojosa, O., Alvarez-Nemegyei, J., Juarez, V., Pacheco-Tena, C., Cedeño, L., Garza-Elizondo, M., Santos, A. M., Goycochea-Robles, M. V., Feicán, A., García, H.,

Julian-Santiago, F., Crespo, M. E., … Originarios), O. behalf of G. (Grupo L. A. D. E. de P. (2020). Syndemic and syndemogenesis of low back pain in Latin-American population: a network and cluster analysis. *Clinical Rheumatology*, *39*(9), 2715–2726. https://doi.org/10.1007/s10067-020-05047-x

Thörn, L. K. A. M., Minamisava, R., Nouer, S. S., Ribeiro, L. H., & Andrade, A. L. (2011a). Pneumonia and poverty: A prospective population-based study among children in Brazil. *BMC Infectious Diseases*, *11*. https://doi.org/10.1186/1471-2334-11-180

Thörn, L. K. A. M., Minamisava, R., Nouer, S. S., Ribeiro, L. H., & Andrade, A. L. (2011b). Pneumonia and poverty: A prospective population-based study among children in Brazil. *BMC Infectious Diseases*, *11*. https://doi.org/10.1186/1471-2334-11-180

Wijesuriya, M., Gulliford, M., Vasantharajah, L., Viberti, G., Gnudi, L., & Karalliedde, J. (2011). DIABRISK - SL Prevention of cardio-metabolic disease with life style modification in young urban Sri Lankan's - study protocol for a randomized controlled trial. *Trials*, *12*. https://doi.org/10.1186/1745-6215-12-209

Xu, J., Takahashi, M., & Li, W. (2024). Identifying vulnerable populations in urban society: a case study in a flood-prone district of Wuhan, China. *Natural Hazards and Earth System Sciences*, *24*(1), 179–197. https://doi.org/10.5194/nhess-24-179-2024

Yadav, A. (2024, August 26). *DBSCAN Algorithm Explained*. Medium. https://medium.com/%40amit25173/dbscan-algorithm-explained-853c0bac9bda

Zelka, M. A., Yalew, A. W., & Debelew, G. T. (2022). The effects of completion of continuum of care in maternal health services on adverse birth outcomes in Northwestern Ethiopia: a

prospective follow-up study. *Reproductive Health*, *19*(1). https://doi.org/10.1186/s12978-022-01508-5