FICHA RESUMEN- PROYECTO DE TRABAJO DE GRADO

TÍTULO: "Modelo de agrupamiento no supervisado para el análisis de la vulnerabilidad de la primera infancia en la ciudad de Cali"

- ÁREA DE TRABAJO: Ingeniería y Economía
- TIPO DE PROYECTO (Aplicado, Innovación, Investigación): Investigación
- ESTUDIANTE(S): Zuly M. Alfonso – Dario A. Medina Q - Diana P. Socha.
- CORREO ELECTRÓNICO: zmalfonso@javerianacali.edu.co, daryhinn0621@javerianacali.edu.co, dsocha12@javerianacali.edu.co,
- DIRECCIÓN Y TELEFONO: Calle 64 N 1 – 15 - 3053226005
- DIRECTOR: Mario Julian Mora Cardona
- VINCULACIÓN DEL DIRECTOR: Docente de la facultad de ingeniería
- CORREO ELECTRÓNICO DEL DIRECTOR: mariomora@javerianacali.edu.co
- CO-DIRECTOR (Si aplica): No aplica
- GRUPO O EMPRESA QUE LO AVALA (Si aplica): Pontifica universidad Javeriana
- OTROS GRUPOS O EMPRESAS: No aplica
- PALABRAS CLAVE (al menos 5): Primera infancia, Geoespacial, Microterritorio, Elementos de territorio, Modelo de agrupamiento, No supervisado, vulnerabilidad
- FECHA DE INICIO: Julio 2023
- DURACIÓN ESTIMADA (En meses): 8-11 meses
- RESUMEN: La niñez y la primera infancia en Colombia enfrentan desafíos y oportunidades, aunque se han logrado avances significativos, aún hay retos importantes para garantizar el bienestar de la niñez y la primera infancia en Colombia, es fundamental seguir trabajando para superar las barreras presentadas y garantizar que todos los niños puedan alcanzar su máximo bienestar. Mediante el análisis de datos de población de primera infancia (0 – 5 años) obtenidas de data SISBEN IV, en articulación con el sistema georreferenciado para la primera infancia desarrollado en el Centro de Investigación Aplicada Riqueza Completa de la Pontificia Universidad Javeriana de Cali (PUJC), se propuso un modelo de agrupamiento no supervisado para la observación de la vulnerabilidad en los microterritorios, con el fin de promover el buen desarrollo de la primera infancia de la ciudad de Cali enfocados en la ubicación estratégica y eficiente de CDI´S Centros de Desarrollo infantil, quienes promueven y protegen los derechos de la primera infancia.

Pontificia Universidad
JAVERIANA
Cali

“Modelo de agrupamiento no supervisado para el análisis de la vulnerabilidad de la primera infancia en microterritorios de la ciudad de Cali”

Zuly Mayerly Alfonso Mosquera
Darío Augusto Medina Quevedo
Diana Paola Socha Godoy

Proyecto de grado para optar al título de
Magister en Ciencia de Datos

Director
Ingeniero Mario Julián Mora Cardona, MSc.

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, DICIEMBRE 5 DE 2024

INTRODUCCIÓN **5**
2. DEFINICIÓN DEL PROBLEMA **7**
2.1 PLANTEAMIENTO DEL PROBLEMA 7
2.2 FORMULACIÓN DEL PROBLEMA 9
3. OBJETIVOS DEL PROYECTO **10**
3.1 OBJETIVO GENERAL 10
3.2 OBJETIVOS ESPECÍFICOS 10
4. JUSTIFICACIÓN **11**
5. MARCO TEORICO Y ANTECEDENTES **13**
5.1 MARCO TEÓRICO 13
5.1.1 Contexto Social 13
5.1.1.1 Primera infancia 13
5.1.1.2 Vulnerabilidad 14
5.1.1.3 Centros de Desarrollo Infantil 14
5.1.2 Contexto Geográfico 14
5.1.2.1 Microterritorios 14
5.1.2.2 Geodatos 15
5.1.3 Conceptos Técnicos 16
5.1.3.1 Analítica descriptiva y predictiva 16
5.1.3.2 Aprendizaje automático 16
5.1.3.3 Modelos supervisados 17
5.1.3.4 Modelos no supervisados 17
5.1.3.5 Modelos de agrupamiento 18
5.1.3.6 Métricas de evaluación de modelos 21
5.1.3.7 Ajuste de parámetros 21
5.1.3.8 Metodología Crisp-DM 22
5.2 ANTECEDENTES 22
6. METODOLOGÍA CRISP-DM PARA ANÁLISIS DE AGRUPAMIENTO NO SUPERVISADO 24
6.1 ESQUEMA DE TRABAJO 25
6.2 FASES DE DESARROLLO DEL PROYECTO. 25
7. PRESENTACION DE LA PROPUESTA **26**
7.1 Entendimiento del objetivo (Negocio) y datos. 26
7.2 Analitica descriptiva de un Microterritorio 45

7.3 Modelamiento y algoritmos de *clustering* 49
7.3.1 Modelo 1: K-MEANS. 50
- Visualización gráfica #1 Método del Codo para K- Means realizada en Python: 50
- Visualización gráfica #2 agrupamiento modelo K-Means realizada en python. 51
- Visualización gráfica #3 agrupamiento con PCA modelo K-Means realizada en python..... 52
7.3.2 Modelo 2: HDBSCAN 52
- Visualización grafica #5 3D en python del agrupamiento con PCA modelo HDSCAN. 56
7.3.3 Modelo 3: *OPTICS* 57
- Grafica #8 Visualización gráfica *clusters* modelo *OPTICS* sobre el Mapa de Comunas de Calí dada la salida del modelo final en Python:..... 60
- Modelo A: Pametros con refinamiento sobre un vecindario del 1% de la data, así: 61
8. VALIDACION DE LA PROPUESTA 65
8.1 Análisis de dimensiones por agrupamiento Modelo 1: KMEANS 66
8.2 Análisis de Modelo 2: *OPTICS* 68
7.3 Análisis de Modelo 3: HDBSCAN..... 74
9. IMPACTOS DEL PROYECTO 79
10. CONCLUSIONES 80
11. TRABAJOS FUTUROS 82
Bibliografía 83
Herramientas de programación: 87
ANEXO 1: Herramienta de visualización Objetivo 4: Gráfica en tableau de *clusters* generados por el modelo K-means..... 89
ANEXO 2: Herramienta de visualización Objetivo 4: Gráfica en tableau de *clusters* generados por el modelo HDBSCAN #33 Clusters. 89
ANEXO 3: Herramienta de visualización Objetivo 4: Gráfica en tableau de *clusters* generados por el modelo HDBSCAN #8 *Clusters.* 90
ANEXO 4: Herramienta de visualización Objetivo 4: Gráfica Visualización en tableau de *clusters* generados por el modelo *OPTICS Modelos A y B.* 90

INTRODUCCIÓN

El desarrollo durante la primera infancia (0-5 años) es una etapa fundamental en la vida de cualquier persona, ya que constituye el inicio de la adquisición de habilidades sociales, emocionales y cognitivas esenciales. Este periodo despierta un interés significativo en diversos campos, incluyendo la investigación científica, la formulación de políticas públicas, la educación y la salud, debido a que en estos primeros años se establecen las bases para el aprendizaje futuro, el bienestar integral y el desarrollo social [1] [2].

En la última década, la ciencia de datos se ha consolidado como una herramienta poderosa para transformar múltiples áreas del conocimiento, permitiendo identificar patrones y tendencias que antes eran difíciles de reconocer. En el ámbito del desarrollo infantil, esta disciplina ha facilitado nuevas perspectivas para abordar desafíos relacionados con la infancia, desde el diseño de intervenciones educativas hasta la evaluación de indicadores de bienestar infantil. Estas herramientas han ampliado nuestra capacidad para analizar grandes volúmenes de información, ofreciendo soluciones más precisas e innovadoras [3] [4].

El presente proyecto tiene como objetivo, mediante el uso de la ciencia de datos, identificar las características que definen un microterritorio, no desde un enfoque puramente territorial, sino centrándose en aquellos espacios geofísicos que suelen pasar desapercibidos. En este caso, el énfasis está en los microterritorios donde se manifiesta la vulnerabilidad de la primera infancia en la ciudad de Cali, específicamente en la Zona Cabecera. Actualmente, la existencia de esta vulnerabilidad no está claramente identificada. Por ello, se recurrió al análisis de conjuntos de datos provenientes de diversas fuentes para detectar patrones y relaciones en espacios georreferenciados, logrando así una comprensión más profunda de este fenómeno.

Utilizando modelos de aprendizaje no supervisado y algoritmos de agrupamiento, se identificaron características comunes entre los distintos microterritorios de la ciudad de

Cali. Entre los modelos empleados, se seleccionó aquel que mostró mayor eficiencia para alcanzar los objetivos del proyecto [5].

El uso de datos georreferenciados permitió visualizar patrones y relaciones en áreas de la ciudad previamente no identificadas, destacando las carencias en el desarrollo óptimo de la primera infancia. Los resultados evidencian la necesidad de diseñar estrategias basadas en este tipo de evidencia, con el fin de apoyar el crecimiento y desarrollo integral de los niños durante sus primeras etapas. Esto beneficiará no solo a los individuos, sino también a la sociedad en su conjunto [6], a través de iniciativas conjuntas entre entidades públicas y privadas.

2. DEFINICIÓN DEL PROBLEMA

2.1 PLANTEAMIENTO DEL PROBLEMA

La niñez y la primera infancia en Colombia enfrentan una serie de desafíos y oportunidades, y aunque se han logrado avances significativos, aún existen retos importantes para garantizar el bienestar de la niñez y la primera infancia en Colombia, por ello es fundamental trabajar para superar las barreras que se presentan y garantizar que todos los niños tengan la oportunidad de alcanzar su máximo bienestar [7].

A partir de estudios relevantes desde el Ministerio de Educación en compañía con el Instituto Colombiano de Bienestar Familiar en adelante ICBF y algunas empresas privadas se ha podido identificar que para brindar solución a muchos de los problemas que puede presentar una ciudad, es prioritario atender las necesidades de la primera infancia (0 – 5 años), por ello el ideal es centrar esfuerzos en la atención, el cuidado y la educación, en otras palabras, garantizar los derechos de los infantes, ya que el futuro de una sociedad o un territorio está en su niñez [8].

A continuación, algunos de estos estudios a resaltar respecto a bienestar de la primera infancia:

- “Lineamiento técnico para la atención a la primera Infancia” del ICBF, este documento suministra las directrices y procedimientos generales para el cuidado integral de los niños en esta etapa (Primera Infancia), adicionalmente enfatiza la importancia de salvaguardar los derechos de los niños y la promoción de su bienestar [9].

- “Actividades rectoras de la primera infancia y de la educación inicial”, se presenta un análisis del Ministerio de Educación de Colombia donde se destaca la relevancia del juego, el arte, la literatura y la exploración del entorno como actividades esenciales para el crecimiento y desarrollo de los niños en sus primeros años de vida (Primera Infancia) [10].

Actualmente, se encuentran estudios orientados a conocer el bien- estar de la niñez en la

ciudad de Cali y el Valle del Cauca, entendiendo estos estudios como la recolección de datos para calcular indicadores agrupados en cinco dimensiones tales como:

- Bienestar Material: pobreza extrema (SISBEN IV), Hacinamiento no mitigable, sin acceso a servicios públicos, sin acceso a agua potable y pisos y paredes inadecuadas.
- Salud: Bajo peso al nacer, vacunación pentavalente, vacunación triple viral, mortalidad < 1, mortalidad (0-4), mortalidad por desnutrición.
- Cuidado, educación y juego: Elegibles para Programas de Atención Integral a la Primera Infancia, atenciones priorizadas en la RIA, educación inicial, cobertura bruta en prejardín, cobertura bruta en jardín, cobertura en transición y déficit de beneficiarios en Atención Integral a la Primera Infancia.
- Bienestar materno: Fecundidad adolescente, control prenatal, partos atendidos por profesionales y razón de mortalidad materna.
- Seguridad y riesgo: Homicidios (0-5), mortalidad por causa externas (0-5), maltrato infantil (0-5), victimización por conflicto armado, maternidad infantil y violencia sexual (0-5).

En este proyecto mediante la ciencia de datos se evaluaron diferentes modelos de *clustering* o agrupamiento no supervisado, para obtener un modelo con el cual se puedan descubrir patrones de vulnerabilidad de la primera infancia (0-5 años) en microterritorios de la ciudad de Cali. Como consecuencia natural del descubrimiento de patrones de vulnerabilidad se pretende aportar en una visualización que facilite diseños e implementación de políticas públicas de optima cobertura, así como también desarrollar actividades que favorezcan el bienestar de este grupo poblacional en pro de garantizar sus derechos de manera eficiente en cuanto a calidad, cobertura y costos de los programas. Para lo anterior, poder tener criterios de similitud entre microterritorios y agruparlos por esos criterios, es fundamental porque servirán para dar un análisis sobre la vulnerabilidad de la primera infancia [11].

2.2 FORMULACIÓN DEL PROBLEMA

¿Cómo implementar un modelo de agrupamiento no supervisado que sea eficiente en descubrir microterritorios no visibles en donde la primera infancia se vulnere en su buen desarrollo en la ciudad de Cali?

¿Es posible la ubicación de CDIS (Centros de Desarrollo infantil del ICBF) en los microterritorios de la ciudad de Cali, dada la concentración poblacional de primera infancia?

¿Cuáles son las variables características propias de un microterritorio, para poder determinar la existencia de vulnerabilidad en la primera infancia?

¿Pueden encontrarse microterritorios con patrones similares de la ciudad de Cali, cuya concentración poblacional sea de primera infancia, a partir de algoritmos de *clustering* no supervisado?

¿Cómo visualizar y presentar los datos obtenidos del modelo de agrupamiento no supervisado seleccionado, sobre un mapa georreferenciado por Planeación Municipal (Distrital) de microterritorios de la ciudad de Cali, que permita evidenciar la afectación de los CDIS en relación con la vulnerabilidad de la primera infancia?

3. OBJETIVOS DEL PROYECTO

3.1 OBJETIVO GENERAL

Proponer un modelo de *clustering* para encontrar microterritorios similares en la ciudad de Cali en los que el desarrollo de la primera infancia sea vulnerable.

3.2 OBJETIVOS ESPECÍFICOS

- Realizar una analítica descriptiva de microterritorios, donde se concentre población de primera infancia utilizando datos de CNPV2018 y SISBEN IV que incluya ubicaciones de CDIS (Centros de Desarrollo Infantil del ICBF).
- Definir las variables que determinan las características de un microterritorio que presenten vulnerabilidad respecto a la primera infancia.
- Seleccionar un algoritmo de *clustering* adecuado, para encontrar microterritorios similares de la ciudad de Cali con concentración de población de primera infancia.
- Visualizar sobre un mapa de la ciudad de Cali, los microterritorios similares encontrados, donde se analice la afectación de los CDIS en relación con la vulnerabilidad de la primera infancia.

4. JUSTIFICACIÓN

El período de la primera infancia, comprendido entre la gestación y los 6 años de edad (menos un día), es un momento trascendental en el desarrollo de los niños, ya que representa una etapa crucial para la formación de las bases del aprendizaje en los próximos años de vida. En este aspecto, uno de los hitos más importantes es la educación. La UNESCO considera que una atención y educación de la primera infancia (AEPI) verdaderamente inclusiva significa mucho más que preparar a los niños para la escuela primaria; puede convertirse en la base del bienestar emocional y cognitivo a lo largo de la vida, además de ser una de las mejores inversiones que un país puede realizar, al promover el desarrollo holístico, la igualdad de género y la cohesión social [12].

El sitio web *El Mejor Lugar para CreSer* ofrece evidencia sobre el diseño, formulación y seguimiento de intervenciones en políticas públicas que impactan positivamente el bienestar de la primera infancia y garantizan el cumplimiento de sus derechos. Este enfoque permite identificar dimensiones que requieren intervención urgente, así como destacar buenas prácticas y logros en la protección de los derechos de la niñez. En este contexto, y como parte de la continuidad del estudio, se propone un modelo de *clustering* para identificar microterritorios similares en la ciudad de Cali donde el desarrollo de la primera infancia sea particularmente vulnerable.

Por otra parte, este proyecto busca dar visibilidad a los microterritorios que presentan patrones de vulnerabilidad similares pero que, a menudo, pasan desapercibidos dentro del territorio de la ciudad de Cali. Este enfoque es crucial para comprender de manera más precisa la distribución actual de los servicios ofrecidos por los Centros de Desarrollo Infantil (CDIS), fundamentales para el desarrollo de la primera infancia. Además, permitirá identificar áreas geográficas con menor acceso a estos servicios, facilitando la implementación de estrategias orientadas a mejorar tanto la cobertura como la equidad en el acceso a dichos recursos.

En contraste con lo mencionado anteriormente, también se busca identificar posibles áreas de superposición en la prestación de servicios dirigidos a la identificación de vulnerabilidades en la primera infancia. Además, esto facilitará la toma de decisiones informadas sobre la redistribución de recursos, evitando la concentración excesiva de instalaciones en ciertas zonas y, al mismo tiempo, abordando las deficiencias en otras, lo que contribuirá a una mejor optimización de los recursos.

Se espera lograr que la información sea adecuadamente difundida y conocida en toda la comunidad, con el fin de sensibilizar sobre la vulnerabilidad de la primera infancia en los diversos microterritorios de la ciudad de Cali. A través de este proceso de reconocimiento, se busca que los ciudadanos se empoderen y, en colaboración con las organizaciones público-privadas involucradas, tomen decisiones informadas sobre la atención a la primera infancia, esto permitirá la implementación de políticas públicas orientadas al desarrollo de la primera infancia, esencialmente adaptadas a las necesidades específicas de cada zona de la ciudad

Finalmente, este proyecto se desarrolla mediante la aplicación de técnicas aprendidas de ciencia de datos, las cuales se desarrollaron a lo largo de su ejecución, demostrando ser herramientas eficaces para abordar los desafíos asociados con la distribución y accesibilidad de recursos dirigidos a la primera infancia [13]. Es relevante destacar que los tres conjuntos de datos fundamentales para este estudio (CNPV 2018, SISBEN IV y CDIS) están disponibles, debidamente estructurados y curados en los servidores del 'Centro de Investigación Aplicada – Riqueza Completa', adscrito a la Facultad de Ciencias Económicas de la Pontificia Universidad Javeriana de Cali. Dichos datos son administrados por el director de este proyecto, conforme a las necesidades y objetivos institucionales. En particular, los datos de SISBEN IV se obtienen en tiempo real a través del acceso directo al Departamento Nacional de Planeación (DNP), con las actualizaciones correspondientes a su base de datos [15].

5. MARCO TEORICO Y ANTECEDENTES

5.1 MARCO TEÓRICO

En primer lugar, se buscar realizar una definición precisa para el proyecto en cuanto al alcance de los microterritorios, explorando la forma teórica y geoespacial, dadas las diferentes y múltiples interpretaciones que se puede tener dentro de un territorio, para poder establecer una definición, se debieron realizar análisis de datos, mapas y encontrar las técnicas y algoritmos de *clustering* más utilizados en temas de densidad y espacialidad que permitiera establecer la dimensión de microterritorio que fuese más eficiente para el proyecto.

5.1.1 Contexto Social

5.1.1.1 Primera infancia

El país cuenta con un nuevo marco jurídico (Código de la infancia y la adolescencia. Ley 1098 de 2006), el cual marca un hito para la defensa y garantía de los derechos humanos de los niños, las niñas y los adolescentes. En este marco se reconoce por primera vez y de manera legal el derecho al desarrollo integral en la primera infancia (Artículo 29): "la primera infancia es la etapa del ciclo vital en la que se establecen las bases para el desarrollo cognitivo, emocional y social del ser humano. Comprende la franja poblacional que va de los cero (0) a los seis (6) años. Son derechos impostergables de la primera infancia: la atención en salud y nutrición, el esquema completo de vacunación, la protección contra los peligros físicos y la educación inicial [14] ".

A partir de la constitución política se determina que aparte de la división general del territorio, habrá las que determine la ley para el cumplimiento de las funciones y servicios a cargo del Estado. Artículo 285 [15] , se definen las entidades territoriales como los departamentos, los distritos, los municipios y los territorios indígenas. Artículo 286 [15]. Los micro territorios se definen como el Espacio territorial y social conformado por un número de familias, que podrán ajustarse dependiendo de la concentración o dispersión poblacional [16].

5.1.1.2 Vulnerabilidad

Para el presente trabajo se tomó como definición de vulnerabilidad en la primera infancia toda situación que impide el ejercicio pleno de los derechos de los niños, niñas y adolescentes, esto acorde y en línea con el concepto de vulnerabilidad ICBF (Instituto Colombiano de Bienestar Familiar) "¿Qué es la vulneración de los derechos del niño, niña y adolescente? La vulneración de derechos es cualquier situación en la cual los niños, niñas y adolescentes queden expuestos al peligro o daño que pueda violar su integridad física y sicológica". Por lo anterior, se tomó como postulado importante para en adelante fuese seleccionadas algunas bases de datos relevantes, que la vulnerabilidad esta predicha al pertenecer al grupo de SISBEN sobre el cual están seleccionados los individuos que hacen parte de la data, en general al necesitar ayuda del estado para que se cumplan con sus derechos y satisfacer sus necesidades básicas.

5.1.1.3 Centros de Desarrollo Infantil

Los Centros de Desarrollo Infantil (CDI) son instituciones en los cuales se presta un servicio institucional que busca garantizar la educación inicial, cuidado y nutrición a niños y niñas menores de 5 años, en el marco de la Atención Integral y Diferencial, a través de acciones pedagógicas, de cuidado calificado y nutrición, así como la realización de gestiones para promover los derechos a la salud, protección y participación, que permitan favorecer su desarrollo integral [17].

5.1.2 Contexto Geográfico

5.1.2.1 Microterritorios

Espacio territorial y/o social conformado por individuos con características similares, barrios, comunas o un mismo número de familias, que podrán ajustarse dependiendo de la concentración o dispersión poblacional [18].

Dado el alcance del proyecto, se esperaba delimitar el concepto de microterritorios a partir de diversas opciones, tales como la distribución de niños por barrio o manzana y la ubicación de los Centros de Desarrollo Infantil (CDI) por barrio o manzana, sin embargo, esto presentó

desafíos para el análisis, particularmente, en la determinación del alcance sin sesgar la delimitación de territorial, finalmente, se definió que se establezca con los resultados de los modelos de clasificación y agrupamiento de estos, basados en los patrones similares encontrados.

5.1.2.2 Geodatos

Los geodatos, son datos que incluyen información relacionada con ubicaciones en la superficie de la Tierra. De esta manera es posible localizar en un mapa objetos, acontecimientos y otros fenómenos del mundo real en una zona geográfica específica identificada mediante coordenadas de latitud y longitud. Estos datos combinan información sobre la ubicación con características o atributos de otros conjuntos de datos durante un periodo determinado de tiempo [19]. Un sistema de geolocalización determina la ubicación de un elemento del territorio en un entorno físico (geoespacial) o virtual (Internet) [20].

Un Sistema de Información Geográfico (SIG) permite relacionar cualquier tipo de dato con una localización geográfica. Esto quiere decir que en un solo mapa el sistema muestra la distribución de recursos, edificios, poblaciones, entre otros datos de los municipios, departamentos, regiones o todo un país. Este es un conjunto que mezcla hardware, software y datos geográficos, y los muestra en una representación gráfica. Los SIG están diseñados para capturar, almacenar, manipular, analizar y desplegar la información de todas las formas posibles de manera lógica y coordinada [21].

Un shapefile almacena la ubicación geométrica y la información de atributos de las entidades geográficas. Las entidades geográficas de un shapefile se pueden representar por medio de puntos, líneas o polígonos (áreas). El espacio de trabajo que contiene shapefiles también puede contener tablas dBASE que, a su vez, pueden almacenar atributos adicionales que se pueden unir a las entidades de un shapefile [22].

El Marco Geoestadístico Nacional (MGN) permite referenciar la información estadística con los lugares geográficos correspondientes, dado que asocia cada dato estadístico al espacio de la superficie terrestre que lo está originando, lo cual contribuye al desarrollo del proceso

estadístico en cada una de sus fases. El MGN está constituido por áreas geoestadísticas (departamentos, municipios, cabeceras municipales, centros poblados, rural disperso, entre otras), delimitadas principalmente por accidentes naturales y culturales, y que son identificables en terreno [23].

5.1.3 Conceptos Técnicos

5.1.3.1 Analítica descriptiva y predictiva

La analítica descriptiva es una etapa preliminar del procesamiento de datos que crea un resumen de los datos históricos para proporcionar información útil, adicionalmente prepara los datos para su posterior análisis. La minería de datos organiza los datos y hace posible identificar patrones y relaciones en los mismos que de otra forma no sería visible, logrando que la consulta, información y visualización de los datos sea de forma más clara [24].
La analítica predictiva se usa para hacer predicciones sobre resultados futuros usando datos históricos combinados con modelos estadísticos, técnicas de minería de datos y aprendizaje automático. Se recurrirá al análisis predictivo para encontrar patrones en los datos con el fin de identificar riesgos y oportunidades [25].

5.1.3.2 Aprendizaje automático

El *machine learning* se entiende como aquel que, mediante el uso de datos y métodos estadísticos, realiza algoritmos que se entrenan para hacer clasificaciones o predicciones, y descubre información clave dentro de los datos. Un método del *machine learning* es el aprendizaje supervisado y se define por su uso de los conjuntos de datos etiquetados para entrenar los algoritmos y clasificar datos o predecir resultados con gran precisión. A medida que se introducen datos de entrada en el modelo, este adapta sus los nuevos pesos hasta que se haya ajustado y nivelado correctamente, esto se denomina proceso de validación cruzada, hito importante para asegurarse de que el modelo evite el sobreajuste o *overfitting* [26].

5.1.3.3 Modelos supervisados

Los modelos supervisados son un tipo de técnica de aprendizaje automático en la cual se entrena un algoritmo utilizando conjunto de datos etiquetados. Estos datos etiquetados consisten en pares de entrada y salida esperada, donde la salida esperada también se conoce como etiqueta o variable objetivo.

El objetivo de un modelo supervisado es aprender una función o relación entre las características de entrada y las salidas esperadas para poder hacer predicciones precisas sobre nuevas entradas sin etiquetar [27].

5.1.3.4 Modelos no supervisados

Los modelos de aprendizaje automático no supervisado son algoritmos que trabajan con datos sin etiquetas, a diferencia del aprendizaje supervisado, donde se entrenan modelos con datos etiquetados para predecir resultados específicos, el aprendizaje no supervisado busca identificar patrones, estructuras o relaciones en los datos por sí mismo, el objetivo es explorar los datos para encontrar estructuras subyacentes, patrones o agrupaciones; estos modelos no tienen información sobre las respuestas correctas durante el entrenamiento, por lo que el aprendizaje se basa en las características propias de los datos.

¿Que son etiquetas? Corresponden a los valores de salida o respuestas asociadas a los datos de entrada, en un conjunto de datos etiquetado, cada instancia de datos viene con una etiqueta correspondiente que indica el resultado deseado o la categoría a la que pertenece. Por ejemplo: si se está tratando con un ejercicio o problema de clasificación, las etiquetas podrían ser categorías como "spam" o "no spam" para correos electrónicos, o "canceroso" y "no canceroso" para muestras de tejido, por otro lado, si es un ejercicio o problema de regresión, las etiquetas podrían ser valores continuos, como el precio de una casa en función de sus características (superficie, ubicación, etc.)

Existen varias técnicas y algoritmos utilizados en modelos no supervisados, entre los cuales se incluyen: *Clustering* (Agrupamiento*), Anomaly detection* (Detección de anomalías), Reducción de dimensionalidad, Reglas de asociación [28].

5.1.3.5 Modelos de agrupamiento

En línea con los objetivos de esta investigación y el interés en el desarrollo del mismo, se ha decidido concentrar el análisis exclusivamente en modelos de *clustering* espacial y de densidad, esta selección se basa en la relevancia de estos modelos para satisfacer las necesidades específicas de la investigación y encontrar el que mejor se ajuste a los datos georreferenciados y a las expectativas previstas.

Los modelos de *clustering* espacial están diseñados para identificar patrones y agrupaciones en datos que tienen una componente espacial, es decir, que están relacionados con ubicaciones geográficas o posiciones en un espacio físico [29].

Mientras los modelos de *clustering* de densidad se basan en la idea de encontrar áreas de alta densidad de puntos en el espacio de características, independientemente del componente espacial. El *clustering* se realiza buscando regiones donde los puntos están más agrupados.

Para comprender la clasificación y que hacen los algoritmos, las fórmulas de distancia asociadas a los métodos de clústeres a utilizar son:

Distancia entre centroides (Euclidiana): La distancia entre los centroides de dos *clusters* es una medida comúnmente utilizada en *clustering*, especialmente en el algoritmo *K-means* [30].

$$d(C_1, C_2) = \sqrt{\sum_{i=1}^{n} (C_{1i} - C_{2i})^2}$$

donde C_1 y C_2 son centroides de los *cluster*, C_{1i} y C_{2i} son las coordenadas del $i - ésimo$ punto en los centroides, y n es el número de dimensiones.

El Modelo K-Means Espacial que toma en cuenta la ubicación geográfica de los datos. Un posible uso del modelo es en la segmentación de clientes basado en la ubicación, agrupando los clientes en *clusters* basados en su ubicación geográfica [31].

Distancia de enlace completo: También conocida como "*maximum linkage*," esta distancia es la máxima distancia entre cualquier par de puntos pertenecientes a los dos *clusters*. $d(C_1, C_2) = max\{d(x, y): x \in C_1, y \in C_2\}$ donde $d(x, y)$ es la distancia entre los puntos x e y de los *clusters* $C_1\ y\ C_2$, respectivamente. Distancia de enlace único: También conocida como "*minimum linkage*," esta distancia es la mínima distancia entre cualquier par de puntos pertenecientes a los dos *clusters*. $d(C_1, C_2) = min\{d(x, y): x \in C_1, y \in C_2\}$ donde $d(x, y)$ es la distancia entre los puntos x e y de los *clusters* $C_1\ y\ C_2$, respectivamente.

- Distancia promedio: Es la distancia promedio entre todos los puntos de un *cluster* a todos los puntos del otro *cluster.* $d(C_1, C_2) = \frac{1}{|C_1||C_2|}\sum_{x \in C_1}\sum_{y \in C_2} d(x.y)$ *d*onde $|C_1|\ y\ |C_2|$son las cantidades de puntos en los *clusters* $C_1\ y\ C_2$ respectivamente [32].
- Distancia de Mahalanobis: Esta distancia es una generalización de la distancia euclidiana que toma en cuenta las correlaciones entre variables. $d(C_1, C_2) = \sqrt{(C_1 - C_2)^T S^{-1}(C_1 - C_2)}$ donde $C_1\ y\ C_2$ son centroides, y S^{-1} es la matriz inversa de covarianza de los puntos [33].
- Distancia de Ward: Esta métrica, utilizada en el método de Ward para *clustering* jerárquico, minimiza la varianza dentro de cada *cluster*.

$$d(C_1, C_2) = \sqrt{\frac{|C_1||C_2|}{|C_1| + |C_2|}} . ||C_1 - C_2||$$

Donde $|C_1|\ y\ |C_2|$ son tamaños de los *clusters*, y $C_1\ y\ C_2$ son sus centroides [34].

El modelo es DBSCAN (*Density-Based Spatial Clustering of Applications with Noise)*, es un algoritmo de *Clustering* Espacial Basado en Densidad con Ruido, en general se puede utilizar para datos espaciales, encontrando *clusters* basados en la densidad local de puntos espaciales [35]**,** existen dos formas para encontrar la distancia para el algoritmo DBSCAN. Una extensión del modelo de DBSCAN es ST-DBSCAN (Spatiotemporal DBSCAN) y se usa en datos que tienen una componente temporal además de la espacial [36]**.**

Condensación Jerárquica: Se construye un árbol jerárquico basado en la densidad, donde los *clusters* se forman por "condensación" de componentes densos conectados a diferentes niveles de densidad. No hay una fórmula explícita para "distancia" entre *clusters*; en su lugar, los *clusters* se definen por la persistencia de densidad (la estabilidad de un *cluster* en diferentes niveles de densidad) [37].

Puntaje de Persistencia: La estabilidad de un *cluster* se mide a través de su persistencia en el árbol de *clusters* jerárquico. Los *clusters* con mayor estabilidad son preferidos.

$$Stability(C) = \sum_{p \in C} (\lambda_{birt}\ (C) - \lambda_{death}(C))$$

Donde: $\lambda_birth\ (C)$ es el nivel de densidad donde el clúster C aparece, $\lambda_death\ (C)$ es el nivel de densidad donde el clúster C desaparece y p es un punto en el clúster C.

En el algoritmo HDBSCAN *(Hierarchical Density-Based Spatial Clustering of Applications with Noise)*, no se utiliza una fórmula matemática específica para calcular la "distancia" entre *clusters* de la manera tradicional en la que se hace en otros métodos de *clustering* como *K-means* o *clustering* jerárquico. En cambio, HDBSCAN se basa en la densidad de los puntos para construir una estructura de *cluster* jerárquica, y los *clusters* se forman mediante la condonación de componentes densos conectados. Es decir, en lugar de calcular una distancia entre *clusters*, HDBSCAN utiliza la jerarquía de densidad para identificar el punto donde los *clusters* son más significativos (donde tienen mayor estabilidad) y se realiza un corte para obtener los *clusters* finales [38].

Finalmente, el modelo *Optics*, basado en el ordenamiento de puntos para identificar la estructura de agrupamiento. Es un algoritmo basado en densidad similar a DBSCAN, pero este presenta una mejora porque puede encontrar agrupaciones significativas en datos que varían en densidad. Lo hace ordenando los puntos de datos de modo que los puntos más cercanos sean vecinos en el ordenamiento.

Esto facilita la detección de diferentes grupos de densidad. El algoritmo *OPTICS* solo procesa cada punto de datos una vez, similar a DBSCAN (aunque se ejecuta más lento que DBSCAN).

También hay una distancia especial almacenada para cada punto de datos que indica que un punto pertenece a un grupo específico. [39].

5.1.3.6 Métricas de evaluación de modelos

Las métricas de evaluación para un modelo no supervisado presentan una naturaleza diferente a las de los modelos supervisados, dado que, en el aprendizaje no supervisado, no hay etiquetas de verdad conocidas, las métricas suelen centrarse en la calidad de la agrupación de los *clusters* o la representación de los datos[40].

- **Coeficiente Silhouette**: Mide qué tan cerca están los puntos dentro de un mismo clúster en comparación con puntos de otros clústers [41].
- **Dunn Index**: Evalúa la separación entre clústeres y la cohesión dentro de ellos.
- **Calinski-Harabasz Index**: Calcula la relación entre la varianza entre *clusters* y la varianza dentro de los clústers.
- **Elbow Method**: Usado para determinar el número óptimo de *clusters* al observar la variación explicada en función del número de *clusters* [42].

5.1.3.7 Ajuste de parámetros

Al igual que las métricas de evaluación algunos modelos de aprendizaje automático no supervisado también pueden requerir un ajuste de hiperparámetros [43]. La literatura explorada muestra que en algoritmos de *clustering* como *K-means*, se debe elegir el número de clústers K, o en modelos como DBSCAN, se deben seleccionar parámetros como la distancia de épsilon y el número mínimo de puntos para formar un clúster, esto implica el uso de técnicas como:

- **Búsqueda en cuadrícula (Grid Search)**: Prueba diferentes combinaciones de hiperparámetros.
- **Búsqueda aleatoria (Random Search)**: Similar a la búsqueda en cuadrícula, pero explora un espacio de hiperparámetros más amplio de manera más eficiente.
- **Métodos bayesianos**: Usan modelos probabilísticos para encontrar combinaciones óptimas [44].

5.1.3.8 Metodología Crisp-DM

Cross-Industry Standard Process for Data Mining es un método estructurado y sistemático utilizado por científicos de datos para realizar proyectos de minería de datos. Esta metodología consta de seis fases principales, que son la comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y la implementación, metodología que se adoptó para desarrollar el presente proyecto.

- Primera fase, comprensión del negocio: Se debe identificar el problema a resolver y se define el objetivo del proyecto.
- Segunda fase: Comprensión de los datos: se realiza una exploración inicial de los datos para identificar su calidad, relevancia y limitaciones.
- Tercera fase, preparación de los datos: Se debe realizar el procesamiento de datos para limpiarlos, integrarlos y transformarlos para su posterior uso en la modelización.
- Cuarta fase, el modelamiento: Implica selección de técnicas de modelado adecuadas y la construcción de modelos predictivos a partir de datos procesados.
- Quinta fase, la evaluación: Se evalúan los modelos construidos para determinar su precisión y fiabilidad.
- Fase de implementación: Los modelos seleccionados se implementan en un entorno operativo y se monitorean de manera continua para garantizar que sigan cumpliendo con sus objetivos [45] [46] [47].

5.2 ANTECEDENTES

A continuación, algunos de los trabajos de investigación relacionados con el presente proyecto y que lo anteceden:

Desde el Centro de Investigación Aplicada de la Pontificia Universidad Javeriana Cali, “Riqueza Completa” el cual promueve agendas de interés colectivo sobre asuntos del desarrollo y de alto impacto en la creación de riqueza innovadora, incluyente y sostenible, que en el caso particular de este proyecto aporta un sin número de variables significativas

en el estudio del desarrollo de la primera infancia.

Desde el proyecto e iniciativa “El mejor lugar para Cre-ser” que busca consolidar la hoja de ruta más efectiva para asegurar el cumplimiento de los derechos de la primera infancia en el valle del cauca y el norte de del cauca. La iniciativa se concreta en el sistema de información georreferenciado de Bienestar en la primera infancia (SIG-PI), el cual mide 5 dimensiones que agrupan 28 indicadores para 55 municipios. Entre las dimensiones se encuentra Bienestar Material, medido a partir de Pobreza Extrema, hacinamiento no mitigable, sin acceso a servicios públicos, sin acceso a agua potable y pisos y paredes inadecuadas, la segunda dimensión se refiere a la salud y esta medida a partir de si el menor tuvo bajo peso al nacer, si tiene la vacuna pentavalente, si completo el esquema de vacunación de la triple viral, la tasa de mortalidad de niños menores a un año, mortalidad hasta los 4 años de edad y por último la mortalidad por desnutrición.

La tercera dimensión corresponde al Cuidado, educación y juego, medido a partir de si son elegibles para programas de atención Integral a la primera infancia, atenciones priorizadas en la RIA, educación inicial, cobertura bruta en prejardín, cobertura bruta en jardín, cobertura en transición y déficit de beneficiarios en atención integral a la primera infancia.

La cuarta dimensión corresponde al bienestar materno, a partir de la fecundidad adolescente, el control prenatal, partos atendidos por profesionales y la razón de mortalidad materna. Por último, está la dimensión de seguridad y riesgo, medido a partir de homicidios, mortalidad por causas externas, maltrato infantil en menores de 5 años, victimización por conflicto armado, maternidad infantil y violencia sexual (0-5) años. [48]

Desde el tablero de criminalidad de riqueza completa, se puede evidenciar los homicidios ocurridos entre o y 500 metros alrededor de diferentes sedes educativas del valle del cauca y del norte del cauca, a partir de variables como las causas del homicidio agrupados por homicidios asociados a pandillas, homicidios por riñas, tráfico de estupefacientes en menores cantidades, homicidios por intolerancia, homicidios por ajustes de cuentas delincuenciales, homicidios por violencia intrafamiliar y homicidios por balas perdidas.

Entre otras variables está el lugar de homicidio y el arma del homicidio, desde este trabajo se podrá tomar referencias para el desarrollo y entrenamiento del modelo final.

El estudio “Desarrollo sociocognitivo en la primera infancia: los retos por cumplir en salud pública en la zona Sabana Centro y Boyacá”, como una aproximación a la necesidad de construir sistemas de salud pública que garanticen el desarrollo infantil integral, de niñas y niños menores de seis años, para identificarlos como componentes necesarios a ser tenidos en cuenta en políticas públicas [49]**.**

Otros estudios, “Asignación del tiempo de niños y niñas en Colombia: factores para el desarrollo de la primera infancia”, trabajo donde se explica cómo influyen variables socioeconómicas, del cuidado y de uso del tiempo de un hogar, en el desarrollo de capital humano de los niños menores de cinco años, para esto, se utiliza la encuesta nacional de uso del tiempo (ENUT) y un modelo econométrico aplicando regresión ordenada [50].

Adicionalmente se tuvo acceso a tableros descriptivos de CNPV2018 y SISBEN IV. Cabe resaltar que no se encontraron trabajos previos realizados en donde se utilicen modelos de agrupamiento no supervisado para el análisis de la vulnerabilidad de la primera infancia en la ciudad de Cali.

6. METODOLOGÍA CRISP-DM PARA ANÁLISIS DE AGRUPAMIENTO NO SUPERVISADO

Para el desarrollo de este proyecto se empleó metodología CRISP DM, que fue considerada debido a que tiene un enfoque sistemático y estructurado para realizar proyectos de ciencias de datos, adicionalmente se cuenta con las bases teóricas, las cuales fueron compartidas en el curso Retos en Ciencia de Datos. Es importante resaltar que para el desarrollo de este trabajo no fue considerada la etapa de despliegue, dado que el alcance de la propuesta estuvo definido solo hasta el desarrollo del modelo y no la implementación a producción de este. Seguir esta metodología nos permitió como equipo integrar la teoría con la práctica de manera efectiva, permitiéndonos una comprensión más profunda de los

conceptos abordados en el proyecto, ya que fueron aplicados dentro de un contexto real y actual de nuestra sociedad.

6.1 ESQUEMA DE TRABAJO

Se acuerda con el equipo destinar para las diferentes tareas asignadas por el curso, director y el flujo propio del proyecto, mínimo 8 horas semanales por estudiante para elaboración y culminación de estas. Se realizaron reuniones semanales con el director para validar avances del proyecto y recibir las retroalimentaciones y aportes respectivos, además de reuniones periódicas con el docente del curso final proyecto aplicado III.

6.2 FASES DE DESARROLLO DEL PROYECTO.

A continuación, detallamos el proceso seguido por el equipo de trabajo para llevar a cabo este proyecto:

Tabla 1. Fases de CRISP-DM.

Tabla de actividades según metodología CRISP-DM		
Entendimiento del negocio	Definición del problema	Anteproyecto de grado.
	Objetivos del proyecto	
	Alcance	
	Justificación	
	Marco teórico de referencia y antecedentes	
	Antecedentes	
	Trabajos relacionados	
	Metodología	
	Recursos a emplear	
	Cronograma	
Entendimiento de los datos	Recolección de los datos	Análisis exploratorio de los datos.
	Explorar los datos	
	Descripción de Las variables	
Preparación de los datos	Limpieza de datos	Código con la preparación y limpieza
	Cálculo de nuevas variables	

Modelado	Encontrar modelo de *clustering* acorde a los datos y variables.	de datos
	Configurar modelo	
Evaluación	Evaluar el modelo	Documento con revisión de los resultados

Fuente: Elaboración propia.

Tabla 2. Tabla bajo esquema fase del modelo CRISP DM – objetivos específicos.

Fase CRISP-DM	**Objetivos específicos relacionados**
Entendimiento del negocio	Objetivo 1
Entendimiento de los datos	Objetivo 1
Preparación de los datos	Objetivo 2
Modelado	Objetivo 2
Evaluación	Objetivo 3
	Objetivo 4

Fuente: Elaboración propia

7. PRESENTACION DE LA PROPUESTA

7.1 Entendimiento del objetivo (Negocio) y datos.

Con el propósito de dar respuesta al primer objetivo, encontrar la definición de microterritorios y su alcance, como también de encontrar la primera infancia y su concentración en estos microterritorios, se llevó a cabo el análisis y exploración de ocho (08) bases de datos obtenidas, a continuación, serán enumeradas y citadas con su apostrofe:

- Base de datos (1): BD_Indicadores.csv.xlsx
- Base de datos (2): Caracterizaci_n_Madres_y_Padres_Comunitarios_ICBF_20240502.csv
- Base de datos (3): cnpv2018_personas_cali
- Base de datos (4): cnpv2018_mgn_cali

- Base de datos (5): sisben_cali
- Base de datos (6): TesisICBF_UnidadesdeServicio
- Base de datos (7): Unidades de Servicio (UDS) en Primera Infancia ICBF
- Base de datos (8): Barrio_Manzanas_Cali_Reprojected.shp

El análisis exploratorio de datos es una técnica utilizada y necesaria en ciencia de datos para examinar y comprender los datos antes de realizar análisis más detallados o para la realización de modelos. El objetivo principal del EDA (Análisis exploratorio de datos) es descubrir patrones, identificar relaciones, detectar valores atípicos y obtener una visión general-particular de los datos. Para comprender la estructura, el contenido y la calidad de los datos, se buscó información en las bases disponibles para el desarrollo del proyecto, a través de esta revisión, se identificaron las características y posibles limitaciones de los datos, lo que permitió establecer un punto de partida sólido para el desarrollo del mismo.

Se inicia al análisis detallado de cada una de las bases de datos, con el fin de determinar su pertinencia y significancia respecto al objetivo del proyecto, así:

- Se llevó a cabo el preprocesamiento de los datos, realizando ajustes a las características o variables para hacerlas más comprensibles, dado que las descripciones en los archivos originales eran abreviadas. Además, se convirtió el formato de los archivos de CSV a Excel, con el objetivo de facilitar el análisis, la exploración y los ajustes en esta misma herramienta.

- Se realizó una revisión detallada de cada una de las características contenidas en las bases de datos, con el propósito de evaluar su relevancia para el cumplimiento de los objetivos del proyecto. Como resultado, se llevó a cabo una preselección de las características de la siguiente manera:

Base de datos (6):

- Registros: 65.993
- Dimensiones:15

Tabla 3. Dimensiones base de datos (6): UDS Unidades de Servicio – ICBF.

UDS Tesis		
vigencia	int64	Año o período en el que los datos o el registro están vigentes.
codigo_regional	int64	Código único que identifica a una región administrativa o de operación.
nombre_regional	Object	Nombre de la región administrativa o de operación correspondiente.
nombre_centro_zonal	Object	Nombre del centro zonal al que pertenece la unidad de servicio
codigo_dane_departamento	int64	Código asignado por el DANE (Departamento Administrativo Nacional de Estadística) al departamento donde está ubicada la unidad de servicio
departamento	Object	Nombre del departamento donde está ubicada la unidad de servicio
codigo_dane_municipio	int64	Código asignado por el DANE al municipio donde está ubicada la unidad de servicio
municipio	Object	Nombre del municipio donde está ubicada la unidad de servicio
codigo_unidad_servicio	int64	Código único que identifica a la unidad de servicio dentro de la base de datos
nombre_unidad_servicio	Object	Nombre de la unidad de servicio
zona_ubicacion	Object	Zona donde está ubicada la unidad de servicio
direccion	Object	Dirección física donde se encuentra ubicada la unidad de servicio
estado	Object	Estado o condición actual de la unidad de servicio
fecha_de_corte	Object	Fecha en la que se realizó el corte o actualización de los datos
geocoded_column	Object	Columna que contiene información georreferenciada, como coordenadas (latitud y longitud) de la ubicación de la unidad de servicio

Fuente: Elaboración propia

Base de datos (5):

- Registros: 1.109.748
- Dimensiones: 116

Tabla 4. Dimensiones base de datos (5): Sisbén personas Cali.

Sisben_cali		
IDE_FICHA_ORIGEN	int64	Identificador único de la ficha de origen.
COD_DPTO	object	Código del departamento (Valle del Cauca).
COD_MPIO	int64	Código del municipio (Santiago de Cali).
D_COD_CLASE	object	Código de la clase de área (urbana o rural).
COD_BARRIO	int64	Código del barrio.
NOM_BARRIO	object	Nombre del barrio.
COD_VEREDA	int64	Código de la vereda.
NOM_VEREDA	object	Nombre de la vereda.
COD_CORREGIMIENTO	int64	Código del corregimiento.
NOM_CORREGIMIENTO	int64	Nombre del corregimiento.
COD_COMUNA	int64	Código de la comuna.
TOT_VIVIENDAS	object	Total de viviendas en el área censada.
TOT_HOGARES	object	Total de hogares en el área censada.
D_TIP_VIVIENDA	object	Tipo de vivienda.
D_TIP_MAT_PAREDES	object	Tipo de material de las paredes.
D_TIP_MAT_PISOS	int64	Tipo de material de los pisos.
D_IND_TIENE_ENERGIA	object	Indicador de si tiene energía eléctrica (sí/no).
TIP_ESTRATO_ENERGIA	object	Estrato socioeconómico del servicio de energía.
D_IND_TIENE_ALCANTARILLADO	object	Indicador de si tiene alcantarillado (sí/no).
D_IND_TIENE_GAS	object	Indicador de si tiene servicio de gas (sí/no).
D_IND_TIENE_RECOLECCION	int64	Indicador de si tiene servicio de recolección de basura (sí/no).
D_IND_TIENE_ACUEDUCTO	int64	Indicador de si tiene acueducto (sí/no).
TIP_ESTRATO_ACUEDUCTO	int64	Estrato socioeconómico del servicio de acueducto.
NUM_CUARTOS_VIVIENDA	object	Número de cuartos en la vivienda.
NUM_HOGARES_VIVIENDA	int64	Número de hogares en la vivienda.
D_TIP_OCUPA_VIVIENDA	int64	Tipo de ocupación de la vivienda (propia,

		arrendada, etc.).
NUM_CUARTOS_EXCLUSIVOS	int64	Número de cuartos exclusivos de la vivienda.
NUM_CUARTOS_DORMIR	object	Número de cuartos utilizados para dormir.
NUM_CUARTOS_UNICOS_DORMIR	object	Número de cuartos únicos utilizados para dormir.
D_TIP_SANITARIO	object	Tipo de sanitario (letrina, inodoro, etc.).
D_TIP_UBI_SANITARIO	object	Ubicación del sanitario (dentro/fuera de la vivienda).
D_TIP_USO_SANITARIO	object	Uso del sanitario (exclusivo/compartido).
D_TIP_ORIGEN_AGUA	object	Origen del agua para consumo (acueducto, pozo, etc.).
D_IND_AGUA_LLEGA_7DIAS	object	Indicador de si el agua llega 7 días a la semana (sí/no).
D_TIP_USO_AGUA_BEBER	object	Tipo de uso del agua para beber (filtrada, hervida, etc.).
D_TIP_ELIMINA_BASURA	object	Tipo de eliminación de basura (recolección, quema, etc.).
D_IND_TIENE_COCINA	object	Indicador de si tiene cocina (sí/no).
D_TIP_PREPARA_ALIMENTOS	object	Tipo de lugar donde se preparan los alimentos.
D_TIP_USO_COCINA	object	Uso de la cocina (exclusivo/compartido).
D_TIP_ENERGIA_COCINA	object	Tipo de energía utilizada para cocinar (gas, electricidad, etc.).
D_IND_TIENE_PC	object	Indicador de si tiene computadora (sí/no).
D_IND_TIENE_INTERNET	int64	Indicador de si tiene acceso a internet (sí/no).
D_IND_GASTO_ALIMENTO	object	Indicador de si se registra gasto en alimentos (sí/no).
VLR_GASTO_ALIMENTO	int64	Valor del gasto en alimentos.
D_IND_GASTO_EDUCACION	object	Indicador de si se registra gasto en educación (sí/no).
VLR_GASTO_EDUCACION	int64	Valor del gasto en educación.
D_IND_GASTO_SALUD	object	Indicador de si se registra gasto en salud (sí/no).
VLR_GASTO_SALUD	int64	Valor del gasto en salud.

D_IND_GASTO_SERV_PUBLICOS	object	Indicador de si se registra gasto en servicios públicos (sí/no).
VLR_GASTO_SERV_PUBLICOS	int64	Valor del gasto en servicios públicos.
D_IND_GASTO_ARRIENDO	object	Indicador de si se registra gasto en arriendo (sí/no).
VLR_GASTO_ARRIENDO	object	Valor del gasto en arriendo.
D_NUM_HABITA_VIVIENDA	int64	Número de personas que habitan la vivienda.
D_IND_EVENTO_INUNDACION	object	Indicador de si ha habido eventos de inundación (sí/no).
NUM_EVENTO_INUNDACION	int64	Número de eventos de inundación registrados.
D_IND_EVENTO_AVALANCHA	int64	Indicador de si ha habido eventos de avalancha (sí/no).
NUM_PERSONAS_HOGAR	int64	Número de personas en el hogar.
IDE_NACIONAL	object	Identificación nacional de la persona (número de cédula o NIT).
IDE_PERSONA	int64	Identificador único de la persona.
D_SEXO_PERSONA	object	Sexo de la persona (masculino/femenino).
EDAD_CALCULADA	object	Edad de la persona.
D_TIP_PARENTESCO	object	Tipo de parentesco con el jefe del hogar.
D_TIP_ESTADO_CIVIL	object	Estado civil de la persona.
D_IND_CONYUGE_VIVE_HOGAR	object	Indicador de si el cónyuge vive en el hogar (sí/no).
D_IND_PADRE_VIVE_HOGAR	object	Indicador de si el padre vive en el hogar (sí/no).
D_TIP_SEG_SOCIAL	object	Tipo de seguridad social de la persona.
D_IND_ENFERMO_30	object	Indicador de si la persona ha estado enferma en los últimos 30 días (sí/no).
D_IND_ACUDIO_SALUD	object	Indicador de si la persona ha acudido a servicios de salud en los últimos 30 días (sí/no).
D_IND_FUE_ATENDIDO_SALUD	object	Indicador de si la persona fue atendida en los servicios de salud (sí/no).
D_IND_ESTA_EMBARAZADA	object	Indicador de si la persona está embarazada

		(sí/no).
D_IND_TUVO_HIJOS	object	Indicador de si la persona ha tenido hijos (sí/no).
D_TIP_CUIDADO_NIÑOS	object	Tipo de cuidado de los niños (guardería, familiar, etc.).
D_IND_RECIBE_COMIDA	object	Indicador de si la persona recibe comida (sí/no).
D_IND_LEER_ESCRIBIR	object	Indicador de si la persona sabe leer y escribir (sí/no).
D_IND_ESTUDIA	int64	Indicador de si la persona estudia actualmente (sí/no).
D_NIV_EDUCATIVO	object	Nivel educativo alcanzado por la persona.
GRADO_ALCANZADO	object	Grado alcanzado en el nivel educativo.
D_IND_FONDO_PENSIONES	int64	Indicador de si la persona está afiliada a un fondo de pensiones (sí/no).
D_TIP_ACTIVIDAD_MES	object	Tipo de actividad realizada en el último mes (trabajando, buscando trabajo, etc.).
NUM_SEM_BUSCANDO	object	Número de semanas que la persona lleva buscando trabajo.
D_TIP_EMPLEADO	int64	Tipo de empleo de la persona (asalariado, independiente, etc.).
D_IND_INGR_SALARIO	object	Indicador de si la persona recibe ingresos por salario (sí/no).
VLR_INGR_SALARIO	object	Valor de los ingresos por salario.
D_IND_INGR_PENSION	object	Indicador de si la persona recibe ingresos por pensión (sí/no).
D_IND_INGR_REMESA_PAIS	object	Indicador de si la persona recibe remesas desde el país (sí/no).
D_IND_INGR_REMESA_EXTERIOR	object	Indicador de si la persona recibe remesas desde el exterior (sí/no).
D_IND_INGR_ARRIENDOS	object	Indicador de si la persona recibe ingresos por arriendos (sí/no).
D_IND_OTROS_INGRESOS	object	Indicador de si la persona tiene otros ingresos (sí/no).

D_IND_INGR_ESTADO	object	Indicador de si la persona recibe ingresos del estado (subsidios, etc.).
D_VLR_INGR_FAM_ACCION	object	Valor de los ingresos recibidos por programas como "Familias en Acción".
D_VLR_INGR_COL_MAYOR	int64	Valor de los ingresos recibidos por "Colombia Mayor" u otros programas similares.
D_VLR_INGR_OTRO_SUBSIDIO	object	Valor de otros subsidios recibidos.
IDE_HOGAR	object	Identificador único del hogar.
D_H_5	object	Indicador de si hay personas con discapacidades en el hogar.
D_I1	object	Indicador de si el hogar tiene acceso a servicios de educación.
D_I2	object	Indicador de si el hogar tiene acceso a servicios de salud.
D_I3	object	Indicador de si el hogar tiene acceso a programas de asistencia social.
D_I4	object	Indicador de si el hogar tiene acceso a servicios de empleo.
D_I5	object	Indicador de si el hogar tiene acceso a servicios de capacitación.
D_I6	object	Indicador de si el hogar tiene acceso a servicios de recreación.
D_I7	object	Indicador de si el hogar tiene acceso a servicios de cultura.
D_I8	object	Indicador de si el hogar tiene acceso a servicios de deportes.
D_I9	object	Indicador de si el hogar tiene acceso a servicios de transporte.
D_I10	object	Indicador de si el hogar tiene acceso a servicios de vivienda.
D_I11	object	Indicador de si el hogar tiene acceso a servicios de agua.
D_I12	object	Indicador de si el hogar tiene acceso a servicios de saneamiento.

D_I13	object	Indicador de si el hogar tiene acceso a servicios de energía.
D_I14	object	Indicador de si el hogar tiene acceso a servicios de telecomunicaciones.
D_I15	object	Indicador de si el hogar tiene acceso a servicios de seguridad.
C	float64	Código de clasificación del hogar según el SISBEN.
D_CLASIFICACION	object	Clasificación del hogar según el puntaje del SISBEN.
IDE_UNIGASTO	float64	Identificador único del gasto del hogar.
D_JEFE_UG	object	Indicador de si la persona es el jefe del hogar.
PERSUG	int64	Número de personas en la unidad de gasto del hogar.
FEC_ACTUALIZACION_CNS	object	Fecha de actualización de la información en el censo.
FEC_DIGITACION	int64	Fecha de digitación de la información.

Fuente: Elaboración propia

Base de datos (3):

- Registros: 1.822.869
- Dimensiones: 48

Tabla 5. Dimensiones base de datos (3): Censo Nacional de personas y vivienda Cali 2018

cnpv2018_personas_cali	
TIPO_REG	int64
U_DPTO	int64
U_MPIO	int64
D_UA_CLASE	object
COD_ENCUESTAS	int64
U_VIVIENDA	int64
P_NROHOG	float64

P_NRO_PER	int64
D_P_SEXO	object
D_P_EDADR	object
D_P_PARENTESCOR	object
D_PA1_GRP_ETNIC	object
D_PA11_COD_ETNIA	object
D_PA12_CLAN	object
D_PA21_COD_VITSA	object
D_PA22_COD_KUMPA	object
D_PA_HABLA_LENG	object
D_PA1_ENTIENDE	object
D_PB_OTRAS_LENG	object
D_PB1_QOTRAS_LENG	object
D_PA_LUG_NAC	object
D_PA_VIVIA_5ANOS	object
D_PA_VIVIA_1ANO	object
D_P_ENFERMO	object
D_P_QUEHIZO_PPAL	object
D_PA_LO_ATENDIERON	object
D_PA1_CALIDAD_SERV	object
D_CONDICION_FISICA	object
D_P_ALFABETA	object
D_PA_ASISTENCIA	object
D_P_NIVEL_ANOSR	object
D_P_TRABAJO	object
D_P_EST_CIVIL	object
D_PA_HNV	object
D_PA1_THNV	object
D_PA2_HNVH	object
D_PA3_HNVM	object
D_PA_HNVS	object
D_PA1_THSV	object
D_PA2_HSVH	object

D_PA3_HSVM	object
D_PA_HFC	object
D_PA1_THFC	object
D_PA2_HFCH	object
D_PA3_HFCM	object
D_PA_UHNV	object
D_PA1_MES_UHNV	object
PA2_ANO_UHNV	float64

Fuente: Elaboración propia

Base de datos (4):

- Registros: 690.604
- Dimensiones: 21

Tabla 6. Dimensiones base de datos (4): Censo Nacional de personas y Vivienda 2018 Marco Geoestadístico Nacional.

cnpv2018_mgn_cali		
U_DPTO	int64	Código del departamento. En el caso de Cali, este código corresponde al departamento del Valle del Cauca.
U_MPIO	int64	Código del municipio. Para Cali, este sería el código específico del municipio de Santiago de Cali.
UA_CLASE	int64	Clase de área de la unidad administrativa (urbana/rural).
UA1_LOCALIDAD	int64	Código de la localidad dentro de la unidad administrativa (en algunas ciudades, esto podría referirse a localidades o comunas).
U_SECT_RUR	int64	Código del sector rural.
U_SECC_RUR	int64	Código de la sección rural.

UA2_CPOB	int64	Código del centro poblado dentro de la unidad administrativa.
U_SECT_URB	int64	Código del sector urbano.
U_SECC_URB	int64	Código de la sección urbana.
U_MZA	int64	Código de la manzana. Se refiere a un área más pequeña dentro de una sección urbana o rural.
U_EDIFICA	int64	Identificador del edificio dentro de la manzana.
COD_ENCUESTAS	int64	Código de la encuesta realizada.
U_VIVIENDA	int64	Código único de la vivienda dentro del edificio o manzana.
COD_DANE_ANM	object	Código DANE del área no municipalizada (puede no aplicarse en todas las áreas urbanas).
DIV_DPTO	int64	División departamental.
DIV_MPIO	int64	División municipal.
DIV_SECT_RUR	int64	División del sector rural.
DIV_SECC_RUR	int64	División de la sección rural.
DIV_CENTRO_POBLADO	int64	División del centro poblado.
DIV_SECT_URB	int64	División del sector urbano.
DIV_SECC_URB	object	División de la sección urbana.

Fuente: Elaboración propia

Base de datos (1):

- Registros: 174
- Dimensiones: 13

Tabla 7. Dimensiones base de datos (1): Bienestar Material primera infancia - Cali

Otras dimensiones
Código Municipio

Nombre Municipio
Número de niños (0-5) registrados en Sisbén IV
% de niños (0-5) que habitan en hogares en situación hacinamiento no mitigable
Número de niños (0-5) que habitan en hogares en situación hacinamiento no mitigable
% de niños (0-5) que habitan en viviendas con pisos y paredes
Número de niños (0-5) que habitan en viviendas con pisos y paredes
% de niños (0-5) que habitan viviendas sin acceso a agua potable
Número de niños (0-5) que habitan viviendas sin acceso a agua potable
% Niños en pobreza extrema
Número Niños en pobreza extrema
% de Niños que habitan en vivienda sin conexión a servicios públicos
Número de Niños que habitan en vivienda sin conexión a servicios públicos
Bienestar material
Código Municipio
Nombre Municipio
Número de niños (0-5) registrados en Sisbén IV
% de niños (0-5) que habitan en hogares en situación hacinamiento no mitigable
Número de niños (0-5) que habitan en hogares en situación hacinamiento no mitigable
% de niños (0-5) que habitan en viviendas con pisos y paredes
Número de niños (0-5) que habitan en viviendas con pisos y paredes
% de niños (0-5) que habitan viviendas sin acceso a agua potable
Número de niños (0-5) que habitan viviendas sin acceso a agua potable
% Niños en pobreza extrema
Número Niños en pobreza extrema
% de Niños que habitan en vivienda sin conexión a servicios públicos
Número de Niños que habitan en vivienda sin conexión a servicios públicos

Fuente: Elaboración propia

Base de datos (6):

- Registros: 65.993
- Dimensiones: 16

Tabla 8. **Dimensiones base de datos (6): Unidades de servicio para la primera infancia.**

UDS Primera Infancia	
Vigencia	int64
Código Regional	int64
Nombre Regional	Object
Nombre Centro Zonal	Object
Código DANE Departamento	int64
Departamento	Object
Código DANE Municipio	int64
Municipio	Object
Número de Contrato	Float64
Código Unidad Servicio	int64
Nombre Unidad Servicio	Object
Zona Ubicación	Object
Dirección	Object
Estado	Object
Fecha de corte	Object
Georreferenciación	Object

Fuente: Elaboración propia

A continuación, se procedió a la depuración y selección de las bases de datos con las que se ejecutarían los modelos, así como de las columnas y datos relevantes que podían ser esenciales y fundamentales para responder a los objetivos establecidos del proyecto. Este proceso de depuración se realizó varias veces, contando también con la guía del director del proyecto, buscando asegurar que solo se incluyeran los datos que realmente aportan valor al análisis y que se alinean con las metas del estudio, con el fin de lograr un modelo solido

al final del proyecto.

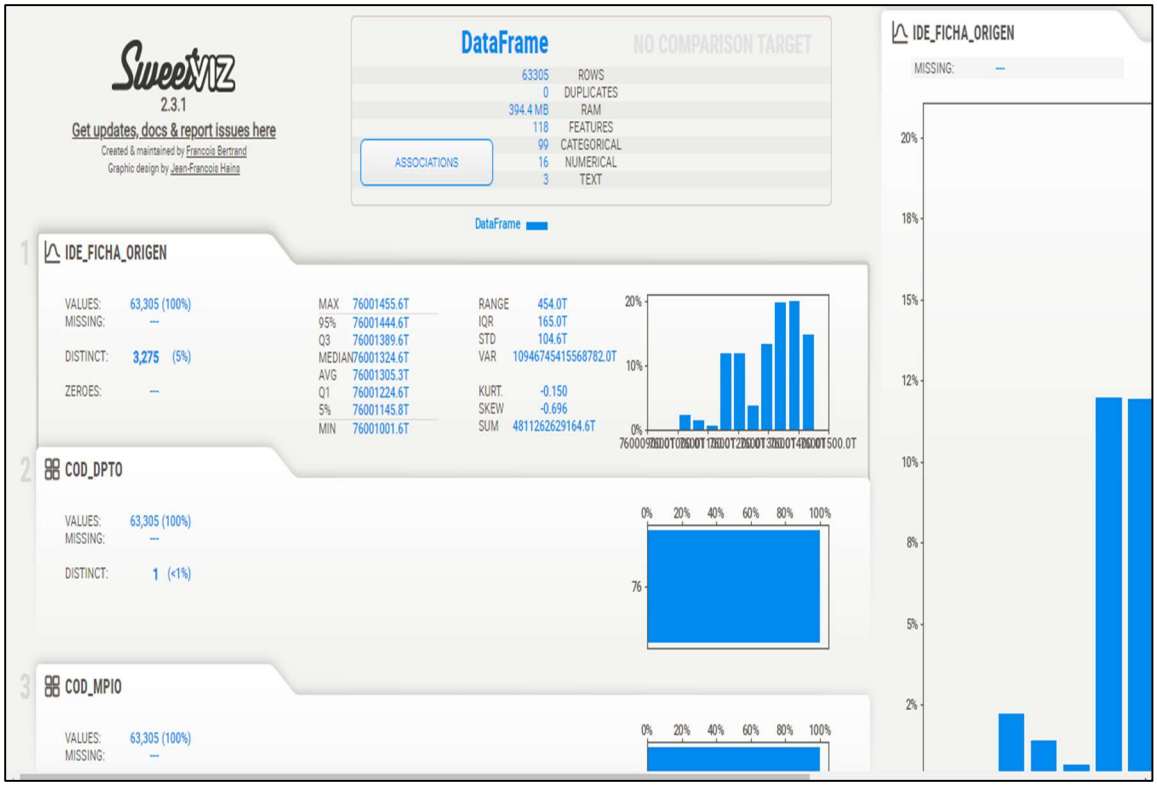
Las bases que finalmente no fueron relevantes y no se tuvieron en cuenta para avanzar en los siguientes pasos fueron:

- Base de datos (1): Esta data contenía información e indicadores de bienestar de la primera infancia, sin embargo, por la complejidad de esta y al encontrarse información no relevante para el proyecto fue reemplazada por la base de datos N.5, dado que contenía la información practica y mayoritaria para el objetivo del estudio.
- Base de datos (2): Esta data contenía información sobre características físicas y sociales de padres y madres comunitarios no relevantes para el proyecto, dado que no son las variables objeto de estudio.
- Base de datos (3): Esta data contenía información sobre características físicas y sociales de las personas y sus viviendas, no relevantes para el proyecto.
- Base de datos (4): Esta data contenía información sobre características físicas y sociales de viviendas, no relevantes para el proyecto.
- Base de datos (6): Datos similares a los existentes en la base de datos N.7 la cual contenía información actualizada constantemente en línea por el ICBF.

Finalmente, se llevó a cabo un análisis estadístico descriptivo de la información recopilada y tratada, con este análisis tuvimos una visión general de los patrones y tendencias presentes en los datos, también pudimos tener las bases sólidas para el futuro análisis y la estructuración de el o los modelos de aprendizaje. Para el análisis estadístico descriptivo se usaron algunas librerías y código abierto Python.

- Usando librería AutoViz y librería SweetViz:

En el desarrollo de la analítica descriptiva de la base de datos seleccionada (Sisben_cali), se determinó el uso de la librería Autoviz, que es una librería de python que genera de forma automática el análisis exploratorio de los datos de la fuente de escogida. Para tener una comparativa y mayor validez en el análisis exploratorio de los datos se hace uso de las librerías SweetViz, que permitió de un modo más gráfico y entendible como el que se muestra a continuación:



Fuente: Elaboración propia - Usando python y la librería SweetViz.

- Usando la librería **DTale**:

Posterior al uso de AutoViz y SweetViz, se decidió el uso de otra librería de python, DTale, que muestra de una forma sencilla y en formato HTML, la matriz de Correlación de Pearson para determinar las correlaciones de las variables de la fuente de datos seleccionada, y de

esta manera poder ir realizando un filtrado o la selección de variables para el modelo de agrupamiento y a su vez descartando las variables que no generarían valor al proyecto.

Fuente: Elaboración propia - Usando python y la librería DTale - Matriz de Correlación de Pearson.

Finalmente se integró la base de trabajo con datos de SISBEN IV y los CDI´s encontrados por puntos de georreferenciación, en un solo libro de excel, con los siguientes resultados:

- 118 columnas o dimensiones.
- 63.305 filas (registros).
- Ningún campo con valores nulos.

Tabla 9. **Clasificación de las variables en el data set:**

Classifying variables in data set	
Number of Numeric Columns	2
Number de Integer-Categorical Columns	22
Number of String-Categorical Columns	30
Number de Factor-Categorical Columns	0
Number of String-Boolean Columns	36
Number de Integer-Boolean Columns	0
Number of Driscrete String Columns	1
Number of NLP String Columns	0
Number of Date Time Columns	2
Number of ID Columns	1
Number of Columns to Delete	24
Predictors classified	118

Fuente: Elaboración propia

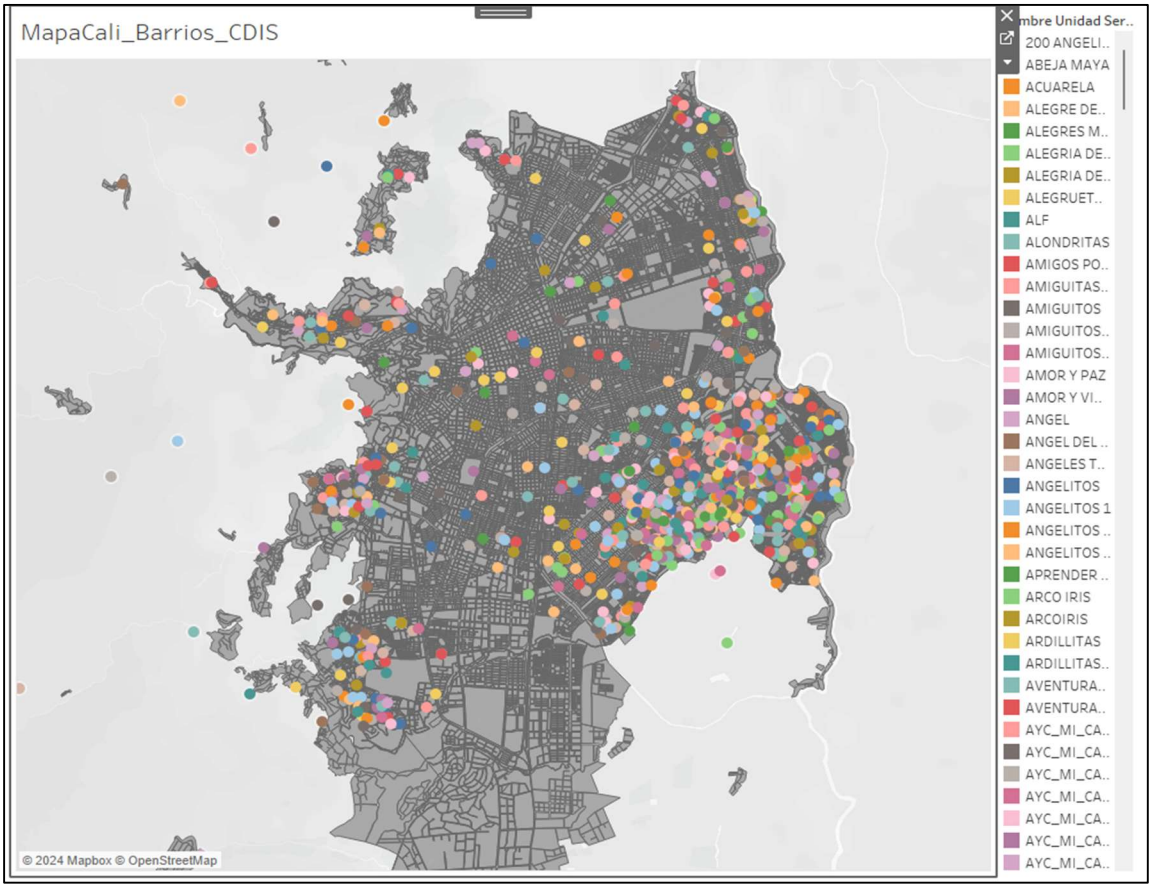
En conjunto con esta etapa estadística, se inició la configuración de información en la herramienta de visualización tableau, elaborando un mapa detallado de la ciudad de Cali, el cual fue segmentado por barrios, apoyados en la base de datos N.8 donde encontramos datos georreferenciados de los barrios de la ciudad de Cali. Paralelamente se adiciona a la historia en tableau la información georreferenciada que se tenía de CDI´s en la base de datos N. 7. Esta historia-mapa permitió identificar, marcar y realizar un trazado de los Centros de Desarrollo Infantil (CDIs) al insertar en un segundo plano la información georreferenciada de CDI´s disponible en la base de datos N. 7.

Con la visualización geográfica de los CDI´s, se logró una mayor comprensión de la densidad y ubicación de estos en la ciudad, permitiendo tener un contexto crítico para analizar visualmente el acceso a los servicios de estos centros de servicios en relación con la ubicación o concentración de la población (Primera infancia) que los requiere.

Visualización georreferenciada de ubicación de barrios y CDI'S sobre mapa |de la ciudad de Cali:

- **Barrios y manzanas de la ciudad de Cali:**

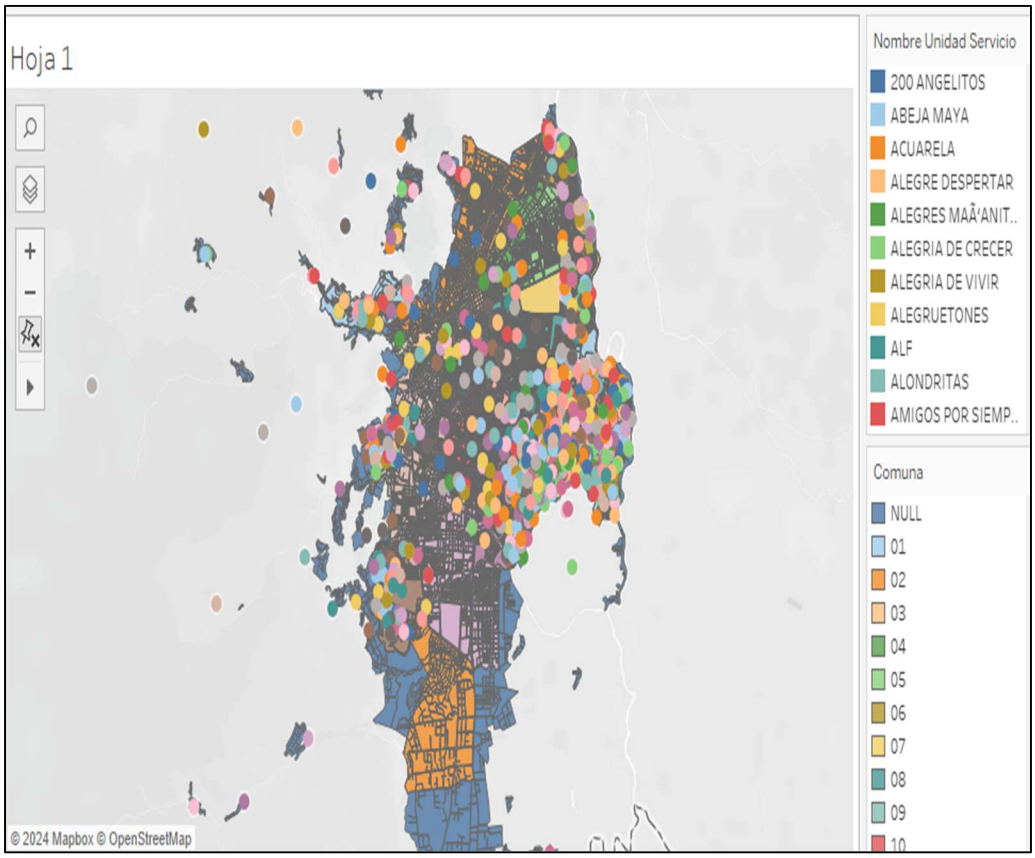
En la visualización a continuación (realizada en Tableau), se observa el mapa de Cali con sus 355 barrios y zonas veredales cercanas, donde se incluyen los 1299 CDIs (Centros de Desarrollo Infantil), esto como una primera aproximación para poder determinar la distribución a priori de los mismos. Esta visualización se origina de la concatenación de las fuentes de datos: 1.- Barrio_Manzanas_Cali_Reprojected.shp y 2.- TesisICBF_UnidadesdeServicio (Filtradas por la ciudad de Cali), posterior a ello se usa una marca para la geometría del mapa de la ciudad de Cali y otra marca para la ubicación por puntos de colores de cada uno de los CDIS georeferenciados por latitud y longitud.



Fuente: Elaboración propia

- **CDI´S por Barrios en la ciudad de Cali:**

Ahora a la visualización en precedencia se le realiza el ajuste de incluir “Comuna”, para conocer la distribución de CDIS en cada una de las 22 comunas de la ciudad de Cali. Se realiza con la inclusión del filtro de “Comuna” en la marca correspondiente a la geometría del mapa de Cali.



Fuente: Elaboración propia

7.2 Analitica descriptiva de un Microterritorio.

De la base de datos depurada, correspondiente a 118 columnas y 63.305 filas (registros), se requirió nuevamente realizar un proceso de limpieza, esto con la finalidad de eliminar ruido en la data para la ejecución de los modelos no supervisados, es así como se logra una fuente de datos de 12 columnas y 53.389 filas.

Tabla 10. Dimensiones seleccionadas para el modelo:

Columna	Tipo de dato
COD_COMUNA	int64
COD_BARRIO	int64
TIP_ESTRATO_ENERGIA	object
D_SEXO_PERSONA	int64
D_TIP_OCUPA_VIVIENDA	object
D_CLASIFICACION	object
D_TIP_CUIDADO_NIÑOS	object
NUM_HOGARES_VIVIENDA	int64
NUM_PERSONAS_HOGAR	int64
PERSUG	int64
PORCENTAJE_PARTICIPACION_NIÑOS_CALI	float64
COEFICIENTE_ATENCION_CDI	float64

Fuente: Elaboración propia

Para poder tener una data robusta para la ejecución de los modelos, esta última limpieza se detalla así:

- Los registros de los barrios en donde no existían niños ni CDI asignados se eliminaron, debido a que podría sesgar los resultados de agrupamiento y por el contrario al dejarlos en nuestra base de datos, se convertirían en datos que ocasionarían ruido en nuestro análisis del modelo.
- Se evidenciaron 9915 niños en 158 barrios diferentes que no tenían CDI asignado directamente en su barrio, es decir el 15, 66% del total de niños pudieron no tener el servicio de CDI o se tuvieron que desplazar hacia el CDI más cercano, estos registros se tuvieron en cuenta puesto que correspondían a nuestro principal objeto de estudio y nos brindaron un nivel de información importante.
- 51 CDI en 17 barrios diferentes de la ciudad de Cali no tienen niños asignados, lo cual puede indicar que se está perdiendo la oportunidad de un mejor servicio para

algunas zonas diferentes donde se identifica que existe un número de niños mucho mayor que el número de CDI asignados a ellos y por tanto también tendría lugar a una fuga de recursos, sin embargo estos datos no son relevantes en nuestro estudio y se eliminarán teniendo en cuenta que no influyen en el nivel de vulnerabilidad de un niño, puede llegar a ser un nuevo objetivo en estudios futuros a fin de optimizar recursos del estado.

Se conformó la base de data definitiva compuesta por 12 variables y un conjunto de datos de 63.201 registros, asi:

1. COD_COMUNA: Comuna 1 a 22 respecto a la distribución espacial de la ciudad de Cali.
2. NOM_BARRIO: Barrio de la ciudad de Cali.
3. TIP_ESTRATO_ENERGIA: Estrato de energía de la vivienda donde habita el infante.
4. D_SEXO_PERSONA: Sexo del infante.
5. D_TIP_OCUPA_VIVIENDA: Tipo de ocupación de la vivienda donde habita el infante.
    1. En arriendo o subarriendo
    2. Propia pagando
    3. Propia pagada
    4. Con permiso del propietario
    5. Posesión sin título, ocupante de hecho
6. D_CLASIFICACION: Clasificación dada por el SISBEN.

    - A01: Población pobreza extrema
    - A02: Población pobreza extrema
    - A03: Población pobreza extrema
    - A04: Población pobreza extrema
    - A05: Población pobreza extrema
    - B01: Población pobreza moderada
    - B02: Población pobreza moderada
    - B03: Población pobreza moderada
    - B04: Población pobreza moderada
    - B05: Población pobreza moderada

- B06: Población pobreza moderada
- B07: Población pobreza moderada

7. D_TIPO_CUIDADO_NIÑOS: Quien se dedica al cuidado del infante.
   0. No informa.
   1. Asiste a un lugar comunitario, jardín o centro de desarrollo infantil o colegio.
   2. Con su padre o madre en la casa.
   3. Con su padre o madre en el trabajo.
   4. Con empleada o niñera en la casa.
   5. Al cuidado de un pariente de 18 años o más.
   6. Al cuidado de un pariente menor de 18 años.
   7. En casa solo.
   8. Otro
   9. No aplica por flujo.
8. NUM_HOGARES_VIVIENDA: Número de hogares que habitan en la vivienda del infante.

- 1 – 8

9. NUM_PERSONAS_HOGAR: Cuantas personas hacen parte del hogar del infante.

- 1 – 23

10. PERSUG: Personas que participan en el gasto del hogar donde se encuentra el infante.

- 2 – 23

11. COCIENTE_ATENCION: Cociente encontrado entre el número de CDI en cada barrio entre el número de niños reportados por el Sisbén en cada barrio.
12. PORCENTAJE_PARTICIPACIÓN: Porcentaje encontrado a partir del número de niños totales hallados en cada barrio de la ciudad de Cali entre el total de niños en Cali, este porcentaje permite identificar la participación de niños por barrio en la ciudad de Cali.

- Se hizo necesario calcular nuevas variables, se relaciona la variable CDI por barrio versus número de niños menores a cinco años por barrio, de esto resulto el cociente de atención, este coeficiente fue encontrado a partir del número de CDI y el número de niños en cada barrio. Partiendo de este supuesto, pudimos inferir sobre la calidad de servicio que pueden recibir los niños y por consiguiente se convirtió en una variable más dentro del modelo. Se pudo evidenciar que a medida que la variable se acerca a 1 los niños acceden a un mejor servicio al haber menos niños ubicados en más CDI, sin

embargo, a medida que se acerca a cero, los niños tienen menor calidad en el acceso al servicio al haber más niños en menos CDI.

Desde el análisis de número de niños versus número de CDI por comuna, se visualizó mediante una gráfica como ciertas comunas tienen una diferencia importante entre el número de niños relacionado en los CDI en cada comuna, en comunas como la 14 y la 21 el número de niños es demasiado extenso para el número de CDI que hay en disponibles, mientras que existen comunas como la 2 y la 19 que se encuentran pocos niños en relación con los CDI que tiene la comuna, esto indica que quizá algunos CDI están sobrando en determinadas zonas y están faltando en otras.

Por otra parte, se analiza el porcentaje de participación de un niño en determinado barrio, con el objetivo de encontrar características que puedan ser similares en determinadas comunas, esto permitió evidenciar que existen barrios con más CDI que otros.

7.3 Modelamiento y algoritmos de *clustering*

Se realizó la exploración del marco teórico para la selección del modelo y algoritmo a utilizar, durante este proceso, se identificaron varios modelos potenciales que podrían abordar el objetivo general N.3. Se decide realizar una exploración de los posibles resultados de estos modelos y de los diferentes algoritmos, las características y funcionalidades de cada uno de ellos, inicialmente se procedió a codificar de manera experimental con tres de estos modelos, así:

- Modelo 1: K-MEANS
- Modelo 2: HDBSCAN
- Modelo 3: *OPTICS*

La elección de estos tres modelos se fundamentó en su capacidad para manejar las particularidades de los datos y en su relevancia para el contexto de estudio, densidad y

espacialidad. Se obtuvieron los siguientes resultados:

7.3.1 Modelo 1: K-MEANS.

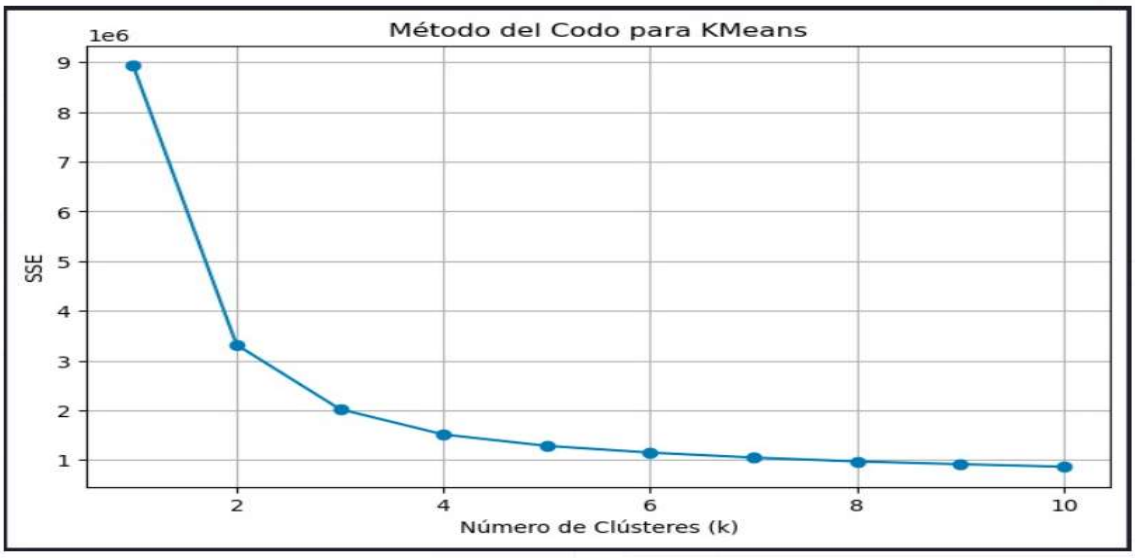
- **Preliminares del modelo**

El algoritmo de *Clustering* K-means es uno de los más usados para encontrar grupos ocultos, o sospechados en teoría sobre un conjunto de datos no etiquetados. Con este modelo se podría confirmar o desterrar- algunas teorías que teníamos asumida de los datos y también reflejar relaciones entre conjuntos de datos que, de manera manual, no hubiéramos reconocido.

- **Modelo y Resultados:**

Mediante la prueba de Hopkins se evidencio que los datos estaban agrupados, por tanto, se podía usar un algoritmo de agrupamiento como K-means, ya que el valor de la prueba fue de 0,89 garantizando la existencia de una estructura de agrupamiento. Adicionalmente, se realizó la prueba del Codo para determinar cuántos clústers eran los eficientes para trabajar en este modelo y los resultados del algoritmo fueron 3 grupos o clusters.

- **Visualización gráfica #1 Método del Codo para K- Means realizada en Python:**



Al usar el modelo K-means desde Python, el algoritmo agrupo la data en 3 clúster de la siguiente manera:

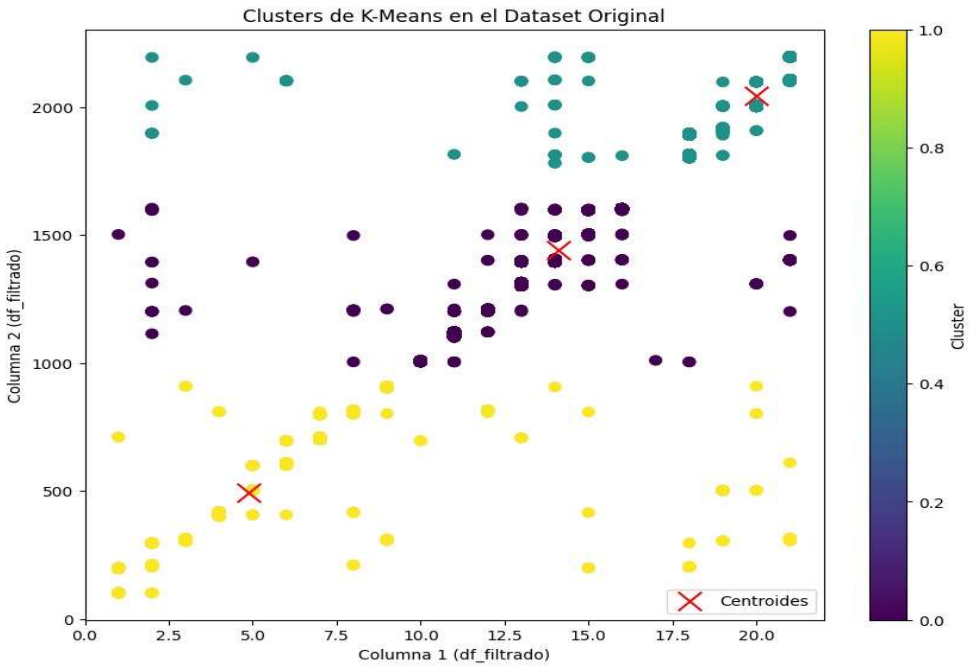
Tabla 11. **Resumen de datos por cluster para el modelo Kmeans:**

#CLÚSTER	DATOS
0	34607
1	18187
2	595

Fuente: Elaboración Propia

Se realiza grafica para observar como el modelo realizo la agrupación de los datos en función de las características analizadas, cada grupo o clúster está identificado por un color distinto, lo que permite visualizar patrones, similitudes y diferencias entre las observaciones. Esta representación facilita la interpretación y validación de los resultados obtenidos del modelo.

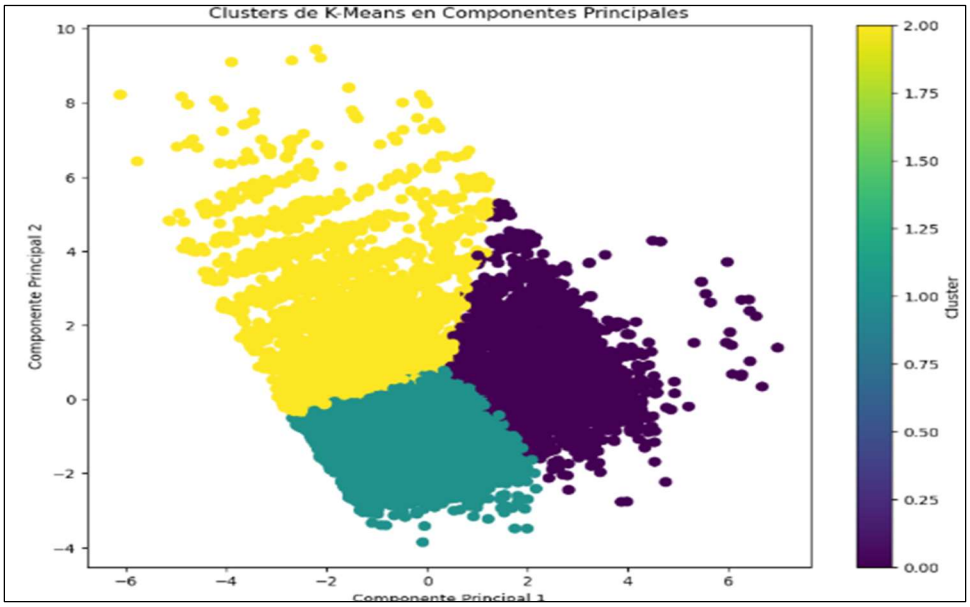
- Visualización gráfica #2 agrupamiento modelo K-Means realizada en python.



La siguiente gráfica presenta los resultados del análisis de componentes principales (PCA), donde

los datos han sido proyectados en un espacio de menor dimensión, esta representación permite visualizar la variabilidad más significativa de las características originales, simplificando la estructura de los datos mientras se preserva la mayor cantidad de información posible. Cada punto en la gráfica refleja una observación, facilitando la identificación de patrones, relaciones y agrupaciones.

- Visualización gráfica #3 agrupamiento con PCA modelo K-Means realizada en python.



7.3.2 Modelo 2: HDBSCAN

- **Preliminares del modelo**

El modelo HDBSCAN es un método de agrupamiento cuyo algoritmo esta creado para encontrar grupos de puntos en áreas densamente pobladas dentro de un conjunto de datos, posee gran habilidad para gestionar de manera efectiva el ruido, es decir, aquellos puntos de datos que no cumplen con los criterios mínimos de densidad para ser asignados a un clúster. Estos puntos se consideran atípicos o irrelevantes en relación a las estructuras densamente pobladas que forman los clústeres, además, resulta especialmente adecuado

para conjuntos de datos que presentan formas irregulares o carecen de una estructura claramente definida, permitiendo una agrupación más flexible y precisa en comparación con otros métodos tradicionales, características que se ajustan a la data que se posee al momento del presente proyecto.

- **Modelo y resultados**

Se ejecuto el método HDBSCAN variando el hiperparámetro "*min_cluster_size*" (Definido como el tamaño mínimo de cada *cluster*), dado que así se evidencio las variaciones respecto al número de *clusters* generados y el ruido en cada ejecución, se destaca que que entre mayor ruido genere el algoritmo, menos estable o confiable es el modelo ejecutado; a continuación se puede ver la tabla con las diferentes variaciones mencionadas, el número de *cluster* generados, el valor del ruido y el valor porcentual del ruido sobre el total de la data, el modelo se ejecutó sobre python:

Tabla 12. Dimensiones seleccionadas para el modelo:

Min_cluster_size	Número de *Clusters*	Ruido	% Ruido
5	1834	18745	35,1102287
10	341	11645	21,81160913
15	173	10467	19,60516211
20	104	7574	14,1864429
25	33	141	0,264099346
30	34	358	0,670550113
35	31	197	0,368989867
40	33	489	0,91591901
45	32	516	0,966491225
50	31	337	0,631216168
55	29	249	0,466388207
60	29	267	0,500103017

Fuente: Elaboración propia

El tamaño mínimo de *cluster* eficiente para el modelo HDBSCAN es de 25, generando de esta forma 33 *clusters,* con un valor mínimo de ruido de 141 puntos y un porcentaje de ruido de 0,264099346%, entre más pequeño sea el ruido generado en la ejecución del modelo mayor es su factibilidad. A continuación, se detalla el proceso y las particularidades en la ejecución del modelo y su refinamiento:

- La base de datos inicial contenía 63305 filas correspondientes a la primera depuración de información del gran archivo de SISBEN IV.
- Se encuentra que, a esta base de datos de 63.305 registros, se le puede realizar un refinamiento más para dar como resultado un archivo consecuente de 53.389 filas. Esto como consecuencia de diferentes validaciones realizadas entre barrios y CDIs y niños dentro del rango de primera infancia.
- Referente a la ejecución del modelo HDBSCAN, como primera medida se ejecutan 12 iteraciones con diferentes *min_clusters_size*, de 5 en 5 iniciando en 5 y terminando en 60. Para esta primera fase de ejecución del modelo, se encuentra que el *min_cluster_size* de 25, arroja 33 *clusters* y un ruido de 141.
- En revisión se determina que 33 *clusters* son demasiados para los objetivos del proyecto, es así como, se propone realizar un mayor número de iteraciones incrementando el valor del *min_cluster_size*. En esta segunda fase se realizaron 800 iteraciones en un ciclo que a inicio desde 100 a 4000, aumentando en múltiplos de 5. Como consecuencia de estas iteraciones se encuentra que el mejor número de *clusters* es el que corresponde a 8.
- De la ejecución del modelo mencionado anteriormente, 172 de ellas corresponden a 8 *clusters*.
- Ahora como tercera fase de iteraciones del modelo HDBSCAN, es decir en un refinamiento del modelo precitado, se propone realizar, la ejecución de esas 172 dan como resultado 8 *clusters* esto con el fin observar. ¿Cuál (es) de esa (s) iteración (es) es la que menos ruido genera?

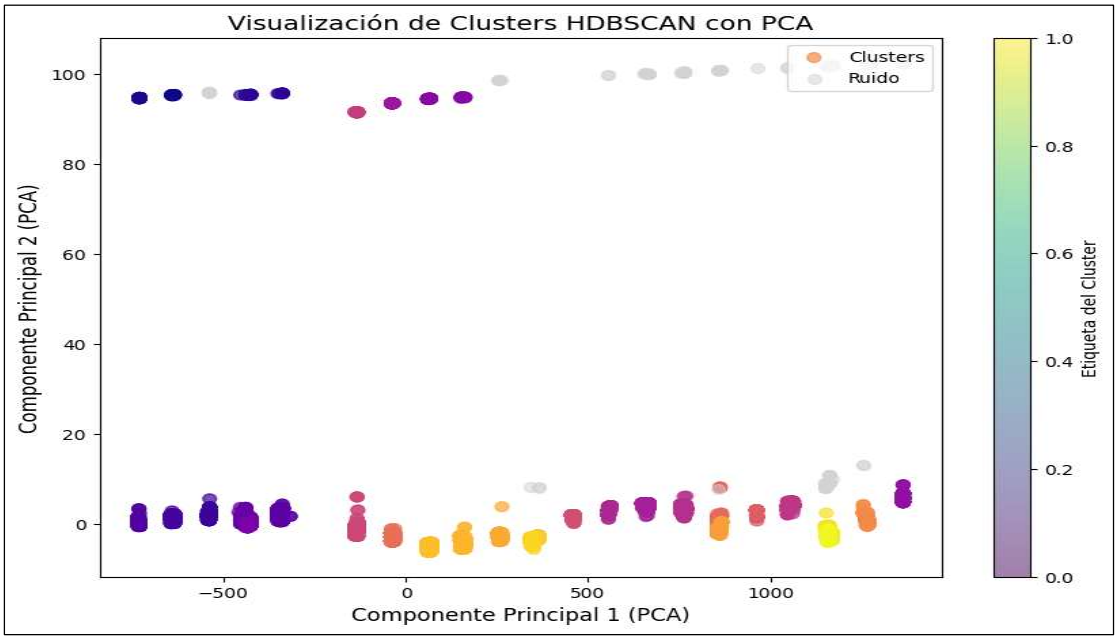
- Se encuentra que el valor del *min_cluster__size* que menor ruido genera es el de 2595, y con la información generada en excel se puede realizar el gráfico de los 8 *clusters*, actualizando la tabla 11, con el refinamiento del modelo se obtuvo, lo siguiente:

Tabla 13. **Dimensiones seleccionadas para el modelo – refinamiento:**

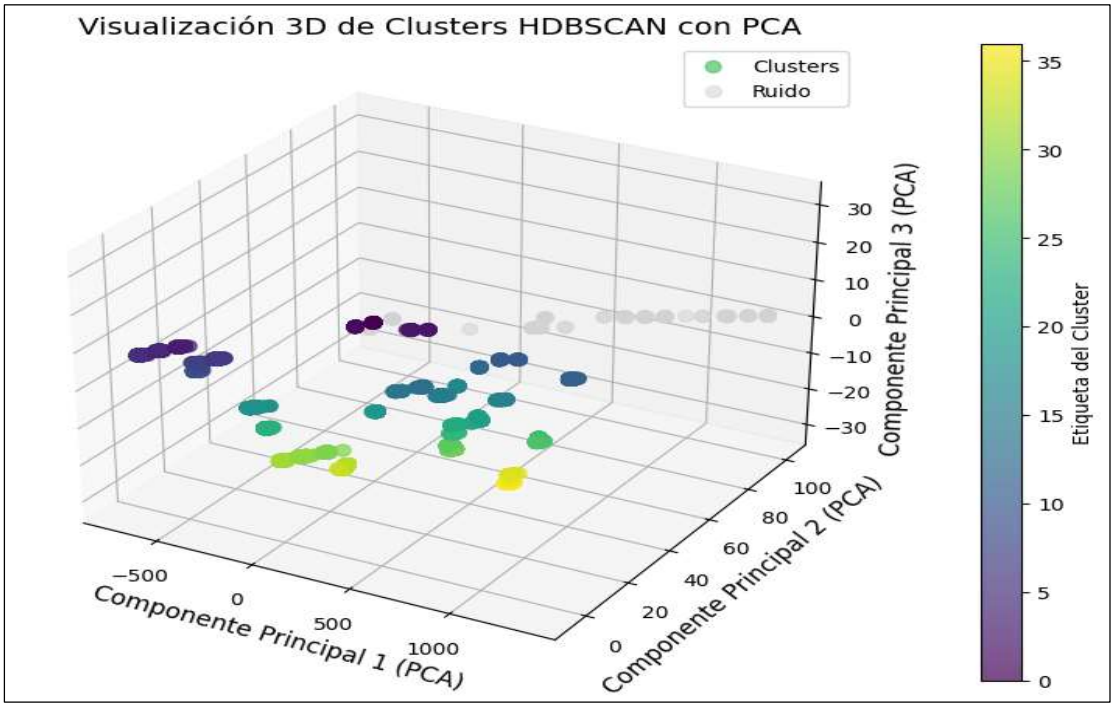
Min_cluster_size	Número de *Clusters*	Ruido	% Ruido
2555	9	5048	9,4551312
2560	8	4481	8,3931147
2565	8	4484	8,3987338
2570	9	5046	9,4513851
2575	8	4479	8,3893686
2580	8	4482	8,3949877
2585	8	4481	8,3931147
2590	8	4483	8,3968608
2595	8	4453	8,3406694
2600	8	4484	8,3987338

Dados los resultados del total de las iteraciones y como punto de partida las comparaciones con el modelo *Optics,* se decide iterar hasta encontrar un agrupamiento y tamaño de error aceptable para el modelo y el proyecto, los resultados finales, aunque generaron un valor de ruido mayor (8.34%), dieron una representación y significancia en el modelo que permitió tomar decisiones sobre el algoritmo a utilizar. A continuación, la salida de la ejecución sobre el programa python:

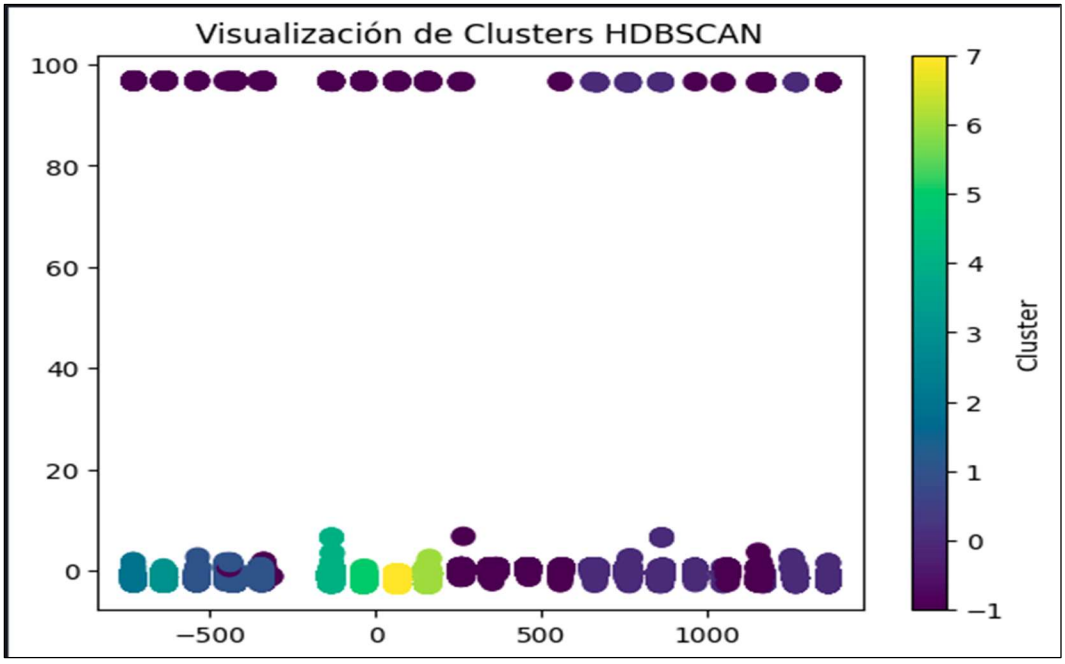
- Visualización gráfica #4 agrupamiento con PCA modelo HDSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) en python.



- Visualización grafica #5 3D en python del agrupamiento con PCA modelo HDSCAN.



- Visualización #6 de Clusters HDBSCAN – con 8 clusters.



7.3.3 Modelo 3: *OPTICS*

- **Preliminares del modelo**

Partimos del postulado que, en un método no supervisado, el algoritmo trabaja con datos sin etiquetas o categorías predefinidas, esto significa que no hay una guía que le diga al algotirmo cómo agrupar los datos o un maestro que le indique como hacerlo, en otras palabras, el algoritmo explora los datos y trata de agrupar patrones similares sin que nadie le diga cómo deben ser esos grupos. En este sentido el algoritmo de densidad en el modelo *Optics* realiza los siguientes pasos:

1. Exploración de datos: El algoritmo examina los datos en busca de patrones o características comunes.
2. Agrupación: Luego, intenta agrupar elementos similares en conjuntos. Por ejemplo, podría agrupar diferentes tipos de legumbres según su tamaño, color o forma, sin saber de antemano cuáles son las categorías correctas.

3. Sin supervisión: Como no hay una instrucción que le indique cómo debe hacer las agrupaciones, el algoritmo debe descubrir las relaciones por sí mismo.

Para empezar la ejecución del código del algoritmo en python, se hace necesaria la transformación de las variables categóricas a numéricas, esto con el fin de evitar que el modelo presente problemas de escala de las variables de tipo entero y de igual forma para las variables categóricas dentro los datos.

- **Modelo y resultados**

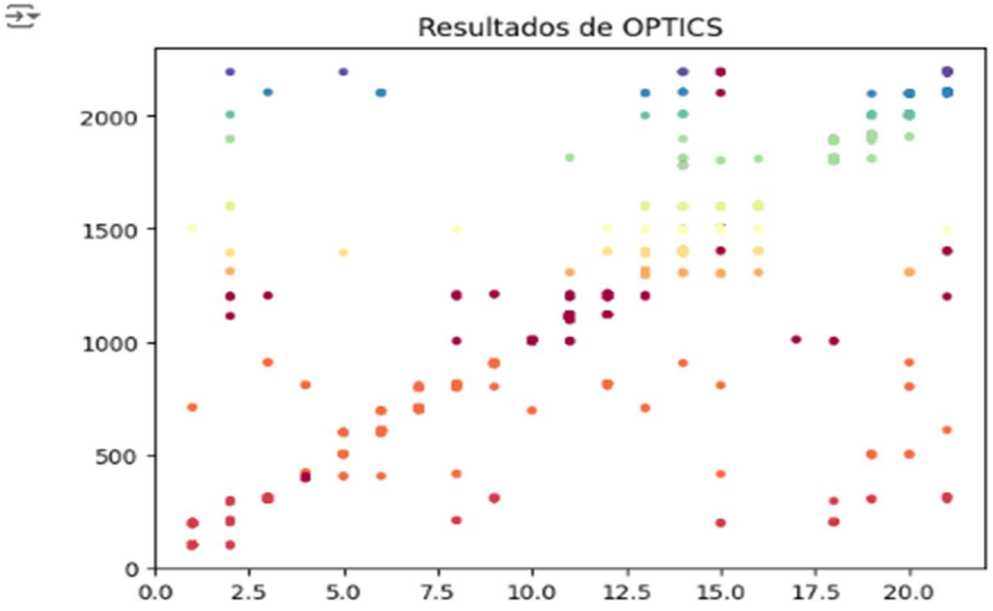
En el modelo *OPTICS*, las características distintivas de cada *cluster* se definen no solo por sus valores promedio, sino también por la densidad local y la distribución de las observaciones (xi) permitiendo explorar *clusters* de diferentes tamaños y formas, como lo podemos ver más adelante en las diferentes iteraciones que se realizaron. Los puntos según el algoritmo en cada *cluster* se agrupan basados en las distancias alcanzables y distancia mínima, parámetros especificados en el modelo y que fueron iterados hasta encontrar el valor de ruido más bajo al igual que el modelo HDBSCAN, con el objetivo de no dejar sesgada o fuera de la clasificación niños en situación de vulnerabilidad. Finalmente, a continuación, las particularidades y el trabajo realizado para obtener los resultados del modelo:

- Gráfico N.06 Tabla de iteraciones en parámetros para encontrar los valores óptimos para aceptar el modelo:

min_samples	min_cluster_size	xi	Ruido	Tamaño de los clusters											
# mínimo de puntos en un vecindario para considerar un punto como núcleo.	Tamaño mínimo de un cluster.	Controla la amplitud de los cambios en la densidad de los puntos.	-1	0	1	2	3	4	5	6	7	8	9	10	11
2000	0,025	0,025	2893	3057	5088	4804	7731	6970	7855	3716	4323	3557	X	X	X
3000	0,025	0,025	6561	3057	5226	4804	7731	6970	7855	3305	4323	3557	X	X	X
3000	0,05	0,05	7325	3057	4462	4804	7731	6970	7855	3305	4323	X	X	X	X
5000	0,05	0,05	7729	3242	12534	6970	7855	15059	X	X	X	X	X	X	X
4000	0,025	0,025	8126	5896	4804	7731	6970	7855	4126	7881	X	X	X	X	X
3000	0.2	0.2	8699	29609	15081	X	X	X	X	X	X	X	X	X	X
15000	0.1	0.1	8926	31411	13052	X	X	X	X	X	X	X	X	X	X
15000	0.2	0.2	8926	31411	13052	X	X	X	X	X	X	X	X	X	X
7000	0.2	0.2	9578	28752	15059	X	X	X	X	X	X	X	X	X	X
1000	0,025	0,025	10096	1434	2013	2013	1970	4804	7731	6970	3138	2417	3305	4323	3557
500	0,05	0,05	11127	4804	7731	6970	7855	3717	3305	4323	3557	X	X	X	X
1000	0,05	0,05	11127	4804	7731	6970	7855	3717	3305	4323	3557	X	X	X	X
300	0,05	0,05	11194	4804	7731	6970	7788	3717	3305	4323	3557	X	X	X	X
10000	0,05	0,05	11194	4804	7731	6970	7788	3717	3305	4323	3557	X	X	X	X
5000	0.1	0.1	11926	11579	6970	7855	15059	X	X	X	X	X	X	X	X
500	0.025	0,025	12718	1434	2013	2013	2056	4804	7731	6970	1369	1478	3305	4323	3557
5000	0,025	0,025	13044	3242	12534	6970	7855	7880	1864	X	X	X	X	X	X
1000	0,05	0,05	13317	28753	11319	X	X	X	X	X	X	X	X	X	X
10000	0.1	0.1	13317	28753	11319	X	X	X	X	X	X	X	X	X	X
10000	0.2	0.2	13317	28753	11319	X	X	X	X	X	X	X	X	X	X
1000	0.1	0.1	15752	7731	6970	7855	15081	X	X	X	X	X	X	X	X
3000	0.1	0.1	15752	7731	6970	7855	15081	X	X	X	X	X	X	X	X
50	0.1	0.1	15819	7731	6970	7788	15081	X	X	X	X	X	X	X	X
100	0.1	0.1	15819	7731	6970	7788	15081	X	X	X	X	X	X	X	X
300	0.1	0.1	15819	7731	6970	7788	15081	X	X	X	X	X	X	X	X
500	0.1	0.1	15819	7731	6970	7788	15081	X	X	X	X	X	X	X	X
50	0,05	0,05	16130	4804	7731	6970	2852	3717	3305	4323	3557	X	X	X	X
5000	0,05	0,05	16130	4804	7731	6970	2852	3717	3305	4323	3557	X	X	X	X
100	0,05	0,05	16248	4804	7731	6970	2634	3717	3305	4323	3557	X	X	X	X
7000	0,05	0,05	16248	4804	7731	6970	2634	3717	3305	4323	3557	X	X	X	X
7000	0,05	0,05	17528	5216	7731	7855	15059	X	X	X	X	X	X	X	X
3000	0,05	0,05	21151	4804	2685	6070	2877	3717	3305	4323	3557	X	X	X	X
7000	0.1	0.1	22744	7731	7855	15059	X	X	X	X	X	X	X	X	X
5000	0.2	0.2	26751	11579	15059	X	X	X	X	X	X	X	X	X	X
50	0.2	0.2	X	33308	15081	X	X	X	X	X	X	X	X	X	X
100	0.2	0.2	X	33308	15081	X	X	X	X	X	X	X	X	X	X
300	0.2	0.2	X	33308	15081	X	X	X	X	X	X	X	X	X	X
500	0.2	0.2	X	33308	15081	X	X	X	X	X	X	X	X	X	X
1000	0.2	0.2	X	38308	15081	X	X	X	X	X	X	X	X	X	X

- Grafica #8 Visualización gráfica *clusters* modelo *OPTICS* dada la salida del modelo final en Python:

```
# Visualización
plt.scatter(X[:, 0], X[:, 1], c=labels, s=10, cmap='Spectral')
plt.title("Resultados de OPTICS")
plt.show()
```

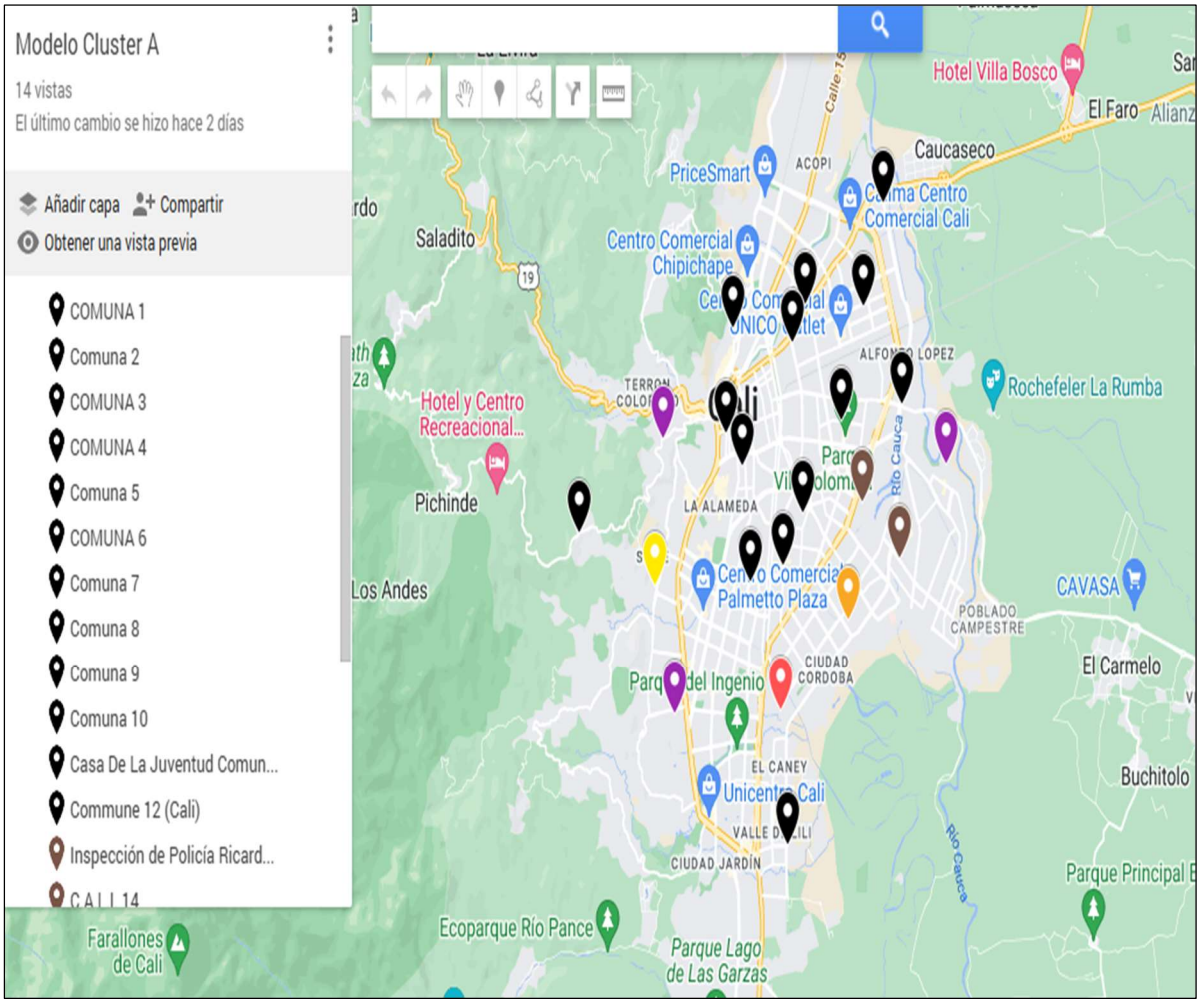


- A continuación, las visualizaciones de 3 diferentes versiones que se realizaron en las etapas del modelo, en el mapa de las comunas de Cali, se puede evidenciar como donde al asignar diferentes medidas a los hiperparámetros los resultados cambiaron y permitieron llegar finalmente a un valor mínimo de ruido.

-

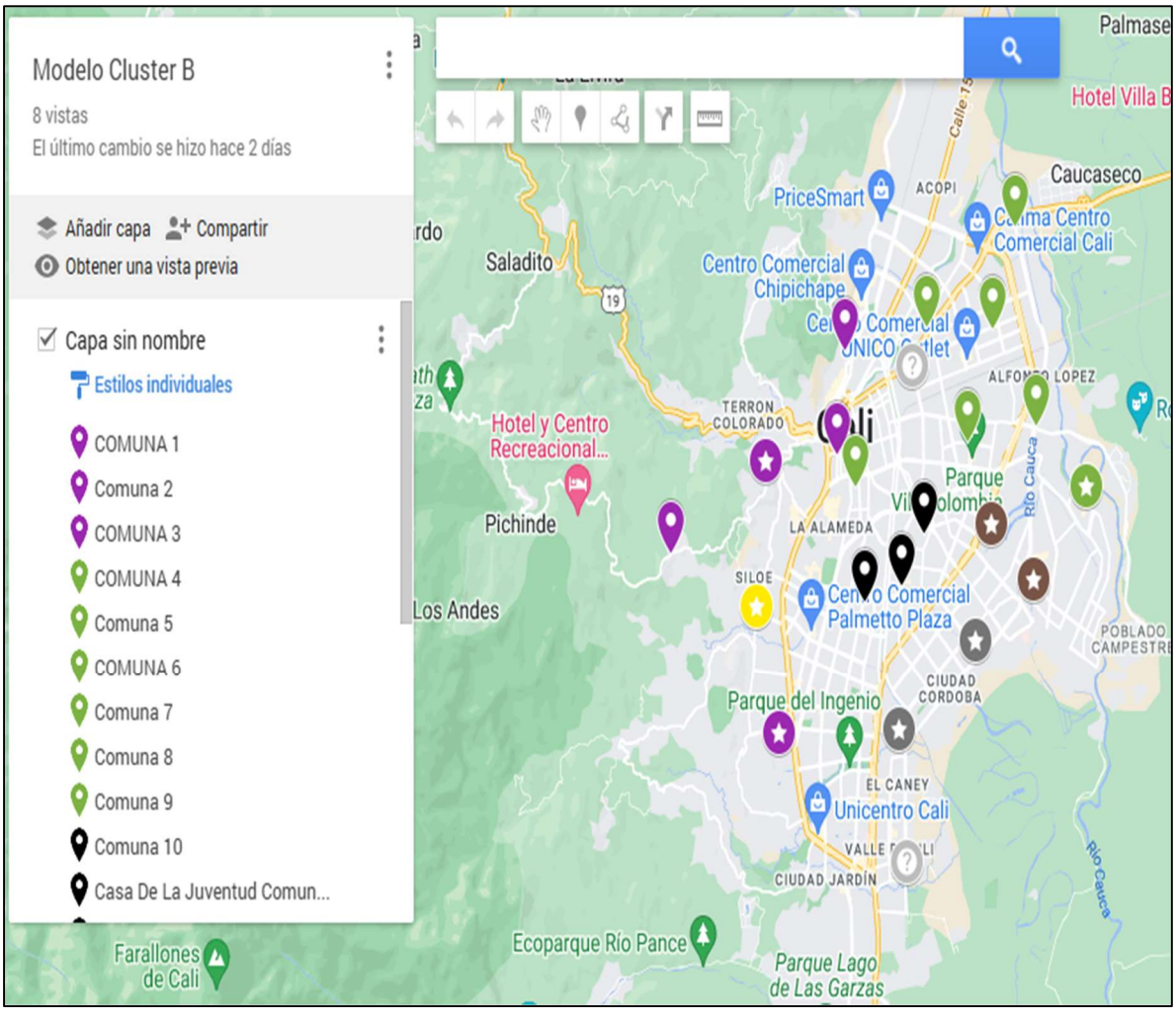
- **Modelo A:** Pametros con refinamiento sobre un vecindario del 1% de la data, así:

MODELO DE APREDIZAJE NO SUPERVISADO OPTICS			Ruido	Tamaño de los clusters											
min_samples	min_cluster_size	xi	-1	0	1	2	3	4	5	6	7	8	9	10	11
5000	0.2	0.2	26751	11579	15059	X	X	X	X	X	X	X	X	X	X



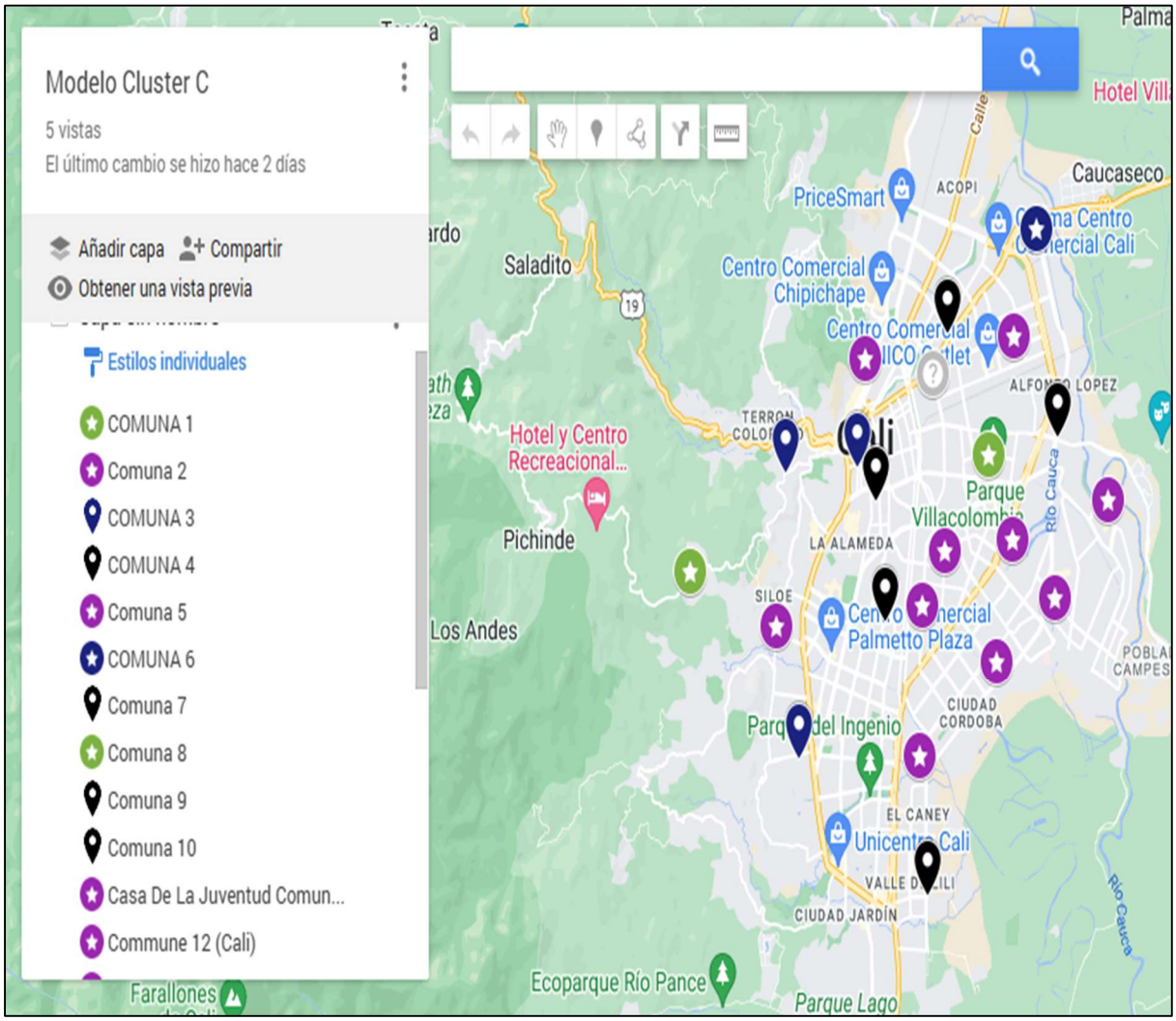
- **Modelo B:** Pametros con refinamiento sobre un vecindario del 2% de la data, así:

MODELO DE APREDIZAJE NO SUPERVISADO OPTICS			Ruido	Tamaño de los clusters											
min_samples	min_cluster_size	xi													
Número mínimo de puntos en un vecindario para considerar un punto como núcleo.	Tamaño mínimo de un cluster.	Controla la amplitud de los cambios en la densidad de los puntos.	-1	0	1	2	3	4	5	6	7	8	9	10	11
50	0.2	0.2	X	33308	15081	X	X	X	X	X	X	X	X	X	X
100	0.2	0.2	X	33308	15081	X	X	X	X	X	X	X	X	X	X
300	0.2	0.2	X	33308	15081	X	X	X	X	X	X	X	X	X	X
500	0.2	0.2	X	33308	15081	X	X	X	X	X	X	X	X	X	X
1000	0.2	0.2	X	38308	15081	X	X	X	X	X	X	X	X	X	X
3000	0.2	0.2	**8699**	29609	15081	X	X	X	X	X	X	X	X	X	X
5000	0.2	0.2	26751	11579	15059	X	X	X	X	X	X	X	X	X	X
7000	0.2	0.2	**9578**	28752	15059	X	X	X	X	X	X	X	X	X	X
10000	0.2	0.2	13317	28753	11319	X	X	X	X	X	X	X	X	X	X
15000	0.2	0.2	**8926**	31411	13052	X	X	X	X	X	X	X	X	X	X



- **Finalmente, el modelo C con parámetros de 0.25% de la data.**

MODELO DE APREDIZAJE NO SUPERVISADO OPTICS			Ruido	Tamaño de los clusters											
min_samples	min_cluster_size	xi													
Número mínimo de puntos en un vecindario para considerar un punto como núcleo.	Tamaño mínimo de un cluster.	Controla la amplitud de los cambios en la densidad de los puntos.	-1	0	1	2	3	4	5	6	7	8	9	10	11
500	0.025	0,025	12718	1434	2013	2013	2056	4804	7731	6970	1369	1478	3305	4323	3557
1000	0,025	0,025	10096	1434	2013	2013	1970	4804	7731	6970	3138	2417	3305	4323	3557
2000	0,025	0,025	2893	3057	5088	4804	7731	6970	7855	3716	4323	3557	X	X	X
3000	0,025	0,025	6561	3057	5226	4804	7731	6970	7855	3305	4323	3557	X	X	X
4000	0,025	0,025	8126	5896	4804	7731	6970	7855	4126	7881	X	X	X	X	X
5000	0,025	0,025	13044	3242	12534	6970	7855	7880	1864	X	X	X	X	X	X



Resultados preliminares del proyecto:

Cada modelo seleccionado fue sometido a procesos de implementación, en los que se analizó el desempeño y resultados en función de los criterios y métricas de evaluación, de igual forma se realizaron comparaciones entre los *clusters* mediante la visualización en tableau, permitiendo determinar cuál de los modelos ofrecía los resultados más precisos y significativos para el objetivo de caracterización de los microterritorios y la vulnerabilidad en la primera infancia de la ciudad de Cali.

- **Exclusión del modelo KMeans**

K-Means es un método que asume que los grupos de datos (c*lusters*) tienen formas redondeadas, por lo cual no fue el más adecuado para los datos de vulnerabilidad objeto de este proyecto, ya que las conexiones entre datos no se encontraban con una continuidad esperada para el modelo, por otra parte Kmeans es un método sensible a valores extremos, y una gran cantidad de estos datos desviaron el centro del grupo y alteraron los resultados, por lo que el modelo requirió determinar los *clusters* con anterioridad, presentando dificultades al no tener claridad sobre la organización de los datos y su forma de medir las distancias entre los puntos, dado que el modelo funciona eficientemente solo cuando los datos están bien separados.

Por otro lado, *OPTICS* y DBSCAN son métodos más flexibles para este tipo de análisis, identificaron grupos con formas irregulares que fue la constante en la data final del modelo, estos modelos manejaron mejor los valores extremos, etiquetándolos como "ruido" (-1) en lugar de influir en los grupos. A diferencia de K-Means, estos modelos no necesitaron determinar la cantidad de grupos con anticipación.

La data del modelo contenía zonas con densidades muy diferentes y tanto *OPTICS* como DBSCAN fueron más efectivos mientras que las métricas de K-Means fueron más confusas, por tanto, *OPTICS* y DBSCAN permitieron tener una interpretación más clara al basarse en la densidad y en cómo están conectados los datos.

5.4 Visualización de microterritorios

En respuesta a este objetivo, que se centra en la visualización en relación con la data del modelo, se han elaborado gráficas utilizando la herramienta tableau. Estas visualizaciones son fundamentales y permiten comprender gráficamente los resultados obtenidos en los modelos y su interpretación. Se adjuntan al final del documento como anexos las visualizaciones logradas no solo con el algoritmo-modelo sino también en el tablero de tableau donde se detallan las siguientes especificaciones:

- Anexo 14 Mapa ciudad Cali por barrios, comunas en la herramienta tableau. (1 Capa de la visualización).

- Anexo 15 Mapa ciudad Cali por CDI, comunas y barrios en la herramienta tableau. (2 Capa de la visualización).

- Las demás historias que hacen parte de la visualización corresponde al comportamiento de los 3 modelos y los agrupamientos definidos por cada uno sobre la ciudad de Cali.

8. VALIDACION DE LA PROPUESTA

La evaluación del modelo se centró en validar la eficacia del algoritmo/modelo seleccionado para determinar la vulnerabilidad de niños en la primera infancia en microterritorios de Cali. Para respaldar esta validación, se realizaron análisis de los datos y de los resultados obtenidos por los diferentes modelos. Además, se tuvo en cuenta la opinión del director de la investigación, aprovechando su experiencia y conocimientos en el área, determinando que los modelos y agrupamiento de estos dos modelos permiten encontrar una similitud entre microterritorios que no eran totalmente visibles, *OPTICS* y DBSCAN finalmente permitieron observar y determinar la concentración de la vulnerabilidad en áreas específicas de la ciudad de Cali, con la que se pudo concluir de forma asertiva este proyecto

de investigación.

8.1 Análisis de dimensiones por agrupamiento Modelo 1: KMEANS

***Cluster* 0:**

Se ubica principalmente en las comunas 6 y 1, con barrios destacados como Ciudadela Floralia y Terrón Colorado, caracterizados por una diversidad socioeconómica. Los hogares pertenecen mayoritariamente a los estratos 2 y 1, con familias de 3 o 4 integrantes y clasificaciones del SISBEN en los niveles A04 y A05, lo que refleja una vulnerabilidad moderada. Además, la interacción con servicios sociales es limitada, evidenciada por un coeficiente de atención bajo (0.0108 y 0.0211). En cuanto a la población infantil, su participación es moderada y predominan los cuidados directos en casa por parte de los padres.

Este grupo, aunque no se encuentra en condiciones críticas, enfrenta desafíos relacionados con la estabilidad económica y social. Se recomienda implementar programas enfocados en educación y formación técnica para mejorar el acceso al empleo, así como fortalecer los servicios de cuidado infantil. Estas intervenciones podrían contribuir a reducir su vulnerabilidad y mejorar la calidad de vida de estas comunidades.

***Cluster* 1:**

Abarca principalmente las comunas 14 y 15, con barrios representativos como Manuela Beltrán y Los Comuneros I Etapa. Los hogares pertenecen mayormente a los estratos 1 y 2, lo que indica niveles socioeconómicos bajos y medio-bajos, con un tamaño familiar típico de 3 o 4 integrantes. Según la clasificación del SISBEN, predominan los niveles A03 y A04, reflejando una mayor vulnerabilidad, aunque el coeficiente de atención es moderado (0.0271 y 0.0241), esto indica cierto grado de interacción con servicios sociales, persisten carencias significativas, especialmente en los hogares con alta participación infantil (4.06 y 3.87 niños por hogar). El cuidado infantil se realiza mayoritariamente en casa, por padres o madres, con acceso limitado a servicios externos.

Dada la alta vulnerabilidad de este *cluster*, se requiere priorización en políticas sociales enfocadas en la infancia y el fortalecimiento económico de los hogares, la priorización, localización e implementación de CDI´s.

***Cluster* 2:**

Los hogares que más predominan son los que tienen más de 5 personas, se observa un posible hacinamiento, lo que puede generar vulnerabilidad al distribuirse los recursos básicos (espacio, alimentos y atención) entre muchos miembros. Esto se agrava en familias con altos valores de PERSUG, que reflejan condiciones de pobreza según el SISBEN, ya que estas familias enfrentan mayores dificultades para cubrir las necesidades de los niños. Además, los bajos coeficientes de atención en Centros de Desarrollo Infantil (CDI) muestran una falta de servicios comunitarios esenciales, lo que limita el acceso a educación y cuidado temprano. Las condiciones de vivienda también evidencian vulnerabilidad, un gran número de familias alquilan o viven en espacios cedidos, lo que sugiere inseguridad habitacional.

Es importante considerar las diferencias de género en la atención a niños, ya que podrían existir brechas en el acceso a recursos. Asimismo, la alta proporción de niños bajo cuidado exclusivo de sus padres puede ser positiva si los cuidadores tienen recursos suficientes, pero preocupante si no hay acceso a educación inicial o estimulación adecuada.

Estas observaciones resaltan la necesidad de validar datos críticos, como la participación infantil en Cali, y de implementar acciones como ampliar la cobertura de CDI, garantizar viviendas seguras y ofrecer apoyo específico a familias en situaciones de pobreza o vulnerabilidad extrema. Monitorear estas variables ayudará a abordar de manera efectiva las necesidades de la primera infancia.

8.2 Análisis de Modelo 2: *OPTICS*

***Cluster* -1:**

Las comunas más frecuentes son la 6 y la 1, lo que sugiere que las familias se concentran principalmente en estas zonas. Los barrios más comunes son Ciudadela la Floralia y Terron Colorado. Los hogares suelen pertenecer a estratos 2 y 1, reflejando condiciones socioeconómicas de niveles medio-bajos y bajos. Predominan familias con 3 o 4 integrantes, con un cuidado infantil basado principalmente en la asistencia a jardines comunitarios o centros similares.

Sin embargo, este *cluster* se caracteriza por un coeficiente de atención infantil muy bajo (0.0108 y 0.0211), lo que refleja una limitada interacción con programas sociales y una baja participación de los niños en iniciativas de apoyo. Esto sugiere que, aunque los hogares tienen características moderadas, enfrentan desafíos relacionados con el acceso a servicios esenciales y la participación infantil en actividades que podrían favorecer su desarrollo.

***Cluster* 0:**

Las comunas más frecuentes son la 14 y la 13, lo que refleja una concentración significativa en el sur de la ciudad. Los barrios más comunes son Manuela Beltrán y Alfonso Bonilla Aragón. Este *cluster* se caracteriza por hogares pertenecientes a los estratos 1 y 2, lo que indica una población en condiciones socioeconómicas bajas y medio-bajas. Los hogares suelen estar conformados por 3 o 4 personas, y el cuidado infantil más común se realiza a través de jardines comunitarios.

Este *cluster* destaca por tener los coeficientes de atención más altos (0.0272 y 0.0416), lo que indica una fuerte interacción con programas sociales enfocados en la infancia. Además, presenta una participación infantil muy alta, lo que lo convierte en un grupo con una mayor proporción de niños. Aunque existe una fuerte presencia de programas de apoyo, las condiciones socioeconómicas de estos hogares reflejan una alta vulnerabilidad, haciendo

necesario continuar fortaleciendo estas iniciativas para garantizar el bienestar de las familias y los niños en estas áreas.

***Cluster* 1:**

Predomina la comuna 6, específicamente el Barrio ciudadela la Floralia, con hogares clasificados en el estrato 2, lo que indica una leve mejora económica respecto al c*luster* 0. Las viviendas son mayoritariamente propias, con un promedio de un hogar por vivienda y alrededor de 3 personas por hogar. En términos de participación infantil, el porcentaje es bajo (1.89%), en niveles similares al *Cluster* 0. Sin embargo, el acceso a Centros de Desarrollo Infantil (CDI) es menor, con un coeficiente de 0.0108.

Aunque este *cluster* refleja una ligera estabilidad económica comparado con el *cluster* 0, persisten limitaciones significativas en el acceso a servicios infantiles. Esto sugiere la necesidad de fortalecer programas sociales enfocados en ampliar el acceso a servicios de desarrollo infantil, así como políticas que consoliden esta mejora económica, asegurando su sostenibilidad a largo plazo.

***Cluster* 2:**

Predomina la comuna 13, específicamente el Barrio El Vergel, con hogares clasificados en el estrato 2, al igual que el *cluster* 1. Las viviendas son propias, con un promedio de un hogar por vivienda y tres personas por hogar. La participación infantil es 2.31%, ligeramente más alta que en otros *clusters*, mientras que el acceso a Centros de Desarrollo Infantil (CDI) tiene un coeficiente de 0.0161, un nivel intermedio comparado con el *cluster* 0 y el *cluster* 1.

Este *cluster* destaca por una mayor proporción de niños en sus hogares, lo que subraya la importancia de implementar programas enfocados en la atención infantil y el desarrollo temprano. Además, el acceso moderado a los CDI sugiere la necesidad de fortalecer estos servicios para garantizar que las comunidades con mayor presencia infantil puedan cubrir adecuadamente sus necesidades.

***Cluster* 3:**

Predomina la comuna 14, específicamente el Barrio Manuela Beltran, con hogares clasificados en el estrato 1, lo que indica bajos recursos económicos, similar al c*luster* 0. Las viviendas son mayoritariamente propias, con un promedio de un hogar por vivienda y tres personas por hogar. Destaca la participación infantil, que alcanza el 4.06%, siendo la más alta entre todos los *clusters*, y un acceso a Centros de Desarrollo Infantil (CDI) de 0.0271, también el mayor registrado.

Aunque este *cluster* refleja condiciones de bajos recursos, la alta concentración de niños y el acceso relativamente mejor a servicios infantiles sugieren la necesidad de priorizar políticas que fortalezcan aún más estos servicios. Es esencial garantizar que este acceso sea sostenible y se complemente con programas de desarrollo infantil y apoyo a las familias, con el objetivo de mejorar las condiciones generales de la comunidad y reducir las brechas socioeconómicas.

***Cluster* 4:**

Predomina la comuna 15, específicamente el Barrio Los Comuneros, y está compuesto por hogares clasificados en el estrato 1, lo que refleja bajos recursos económicos. Las viviendas son en su mayoría propias, con un promedio de un hogar por vivienda y tres personas por hogar. La participación infantil es del 3.87%, lo que indica una alta concentración de niños en estas comunidades, mientras que el acceso a Centros de Desarrollo Infantil (CDI) tiene un coeficiente de 0.0241, considerado relativamente bueno.

La data permite evidenciar la necesidad de fortalecer los servicios enfocados en la población infantil, dado su peso significativo en la estructura demográfica. Además, es importante garantizar que las comunidades cuenten con programas sociales integrales que combinen acceso a educación, cuidado infantil y apoyo económico, con el objetivo de mejorar la calidad de vida de los hogares y fomentar el desarrollo sostenible de estas áreas.

***Cluster* 5:**

Predomina la comuna 15, específicamente en el Barrio Mojica, con hogares en el estrato 2, lo que refleja una leve mejora económica comparado con el c*luster* 4. Las viviendas son mayoritariamente propias, con un promedio de un hogar por vivienda y tres personas por hogar. La participación infantil es del 3.59%, lo que indica una alta concentración de niños, mientras que el acceso a Centros de Desarrollo Infantil (CDI) tiene un coeficiente de 0.0224, similar al c*luster* 4.

El agrupamiento refleja comunidades con mejores recursos económicos, pero con una alta proporción de niños que necesitan atención específica. Es clave fortalecer el acceso a servicios infantiles y programas educativos, asegurando que esta leve mejora socioeconómica se traduzca en beneficios sostenibles para las familias, especialmente para los niños que constituyen una parte importante de la población en esta área.

***Cluster* 6:**

Predomina la comuna 18, específicamente en el Barrio Sector Alto de los Chorros, compuesto por hogares clasificados en el estrato 1, lo que refleja condiciones económicas limitadas. Las viviendas son en su mayoría propias, con un promedio de un hogar por vivienda y tres personas por hogar. La participación infantil es de 1.64%, considerablemente más baja que en los c*lusters* 4 y 5, mientras que el acceso a Centros de Desarrollo Infantil (CDI) tiene un coeficiente de 0.0148, siendo también menor que en esos *clusters.*

Este *cluster* se caracteriza por una menor proporción de niños y un acceso más limitado a servicios infantiles. Aunque la necesidad de servicios para la población infantil es menos apremiante, se recomienda implementar estrategias para mejorar el acceso a recursos básicos y fortalecer las condiciones de vida de los hogares, especialmente en áreas relacionadas con educación y salud, para garantizar un desarrollo más equilibrado de la comunidad.

***Cluster* 7:**

Predomina la comuna 20, específicamente en el Barrio Siloé, compuesto por hogares de bajos recursos clasificados en el estrato 1. Las viviendas son en su mayoría propias, con un promedio de un hogar por vivienda y tres personas por hogar. La participación infantil es moderada, con un 2.36%, mientras que el acceso a Centros de Desarrollo Infantil (CDI) es bajo, con un coeficiente de 0.0142.

Este *cluster* refleja una comunidad con una proporción moderada de niños, pero con acceso limitado a servicios infantiles, lo que resalta la necesidad de implementar programas que amplíen las oportunidades para el desarrollo temprano. Es crucial fortalecer la infraestructura de servicios básicos y garantizar que las familias con niños puedan acceder a recursos que promuevan su bienestar y desarrollo integral, pese a las condiciones económicas limitadas.

***Cluster* 8:**

Predomina la comuna 21, específicamente en el Barrio Desepaz, con hogares clasificados en el estrato 1, lo que refleja bajos recursos económicos. Las viviendas son en su mayoría propias, con un promedio de un hogar por vivienda y tres personas por hogar. La participación infantil es baja, con un 1.93%, menor que en los c*lusters* 4 y 5. Además, el acceso a Centros de Desarrollo Infantil (CDI) es el más limitado entre todos los *clusters,* con un coeficiente de 0.0096.

Este *cluster* destaca por tener una baja proporción de niños y un acceso muy limitado a servicios infantiles. Las intervenciones deben enfocarse en garantizar el acceso a programas sociales básicos, priorizando la expansión de servicios educativos y de desarrollo infantil. Asimismo, es importante implementar políticas de apoyo para mejorar las condiciones generales de los hogares y aumentar la disponibilidad de recursos esenciales en estas comunidades.

***Cluster* 9:**

Predomina la comuna 21, específicamente en el Barrio Potrero Grande, compuesto por hogares clasificados en el estrato 1, reflejando bajos recursos económicos. Las viviendas son en su mayoría propias y presentan condiciones regulares. En promedio, hay un hogar por vivienda, con tres personas por hogar. La participación infantil es alta, con un 3.49%, lo que indica una significativa concentración de niños en comparación con otros *clusters*. Sin embargo, el acceso a Centros de Desarrollo Infantil (CDI) es limitado, con un coeficiente de 0.0096, uno de los más bajos entre los *clusters.*

Este *cluster* combina una alta proporción de niños con importantes carencias en el acceso a servicios infantiles. Es prioritario enfocar estrategias en el desarrollo de infraestructura social para la infancia, como Centros de Desarrollo Infantil y programas educativos. La ubicación en la comuna 21 refuerza la necesidad de políticas de apoyo social y económico, que puedan brindar mayores oportunidades de desarrollo a las familias y mejorar su calidad de vida.

En síntesis, los *clusters* (*Cluster* -1, *Cluster* 8 y *Cluster* 9) destacan por condiciones críticas, con acceso extremadamente limitado a Centros de Desarrollo Infantil (CDI) y porcentajes variables de niños. El c*luster* -1, con la menor proporción infantil (0.95%) y un coeficiente de CDI de 0.0097, refleja un grupo marginal de bajos recursos y atención insuficiente a la infancia. Los c*lusters* 8 y 9 presentan proporciones de niños más altas (1.93% y 3.49%, respectivamente), pero comparten el acceso más bajo a CDI (0.0096), lo que sugiere una combinación peligrosa de alta concentración infantil y pocos recursos. c*lusters* moderadamente vulnerables (C*lusters* 6 y 7) muestran proporciones de niños de 1.64% y 2.36%, respectivamente, con acceso a CDI algo mejor (0.0148 y 0.0142), aunque todavía insuficiente para garantizar un desarrollo adecuado. En los *clusters* con menor vulnerabilidad relativa (*Clusters* 4 y 5), se observan proporciones altas de niños (3.87% y 3.59%) y un acceso a CDI algo más adecuado (0.0241 y 0.0224), con condiciones ligeramente mejores debido a su clasificación en estrato 2 (*Cluster* 5). Finalmente, los *clusters* de baja vulnerabilidad (*Clusters* 0, 1, 2 y 3) tienen menor concentración infantil (entre 1.8% y 2.3%)

y acceso moderado a CDI (0.0161 a 0.0271), lo que refleja condiciones menos críticas. Es recomendable priorizar intervenciones urgentes en los c*lusters* -1, 8 y 9, fortalecer servicios de los CDI's en los c*lusters* 6 y 7, consolidar avances en los c*lusters* 4 y 5, y apoyar la sostenibilidad de los *clusters* 0, 1, 2 y 3 para prevenir retrocesos en los avances sociales.

7.3 Análisis de Modelo 3: HDBSCAN

***Cluster* -1:**

Representa áreas vulnerables, principalmente en las comunas 12 y 11, y en barrios como El Rodeo y Los Chorros, que concentran una parte significativa de la población. Los hogares se clasifican mayoritariamente en los estratos socioeconómicos 2 y 3, lo que refleja condiciones económicas bajas o medio-bajas. La mayoría de las viviendas son propias u ocupante de hecho, con clasificaciones 3 y 4, lo que sugiere condiciones habitacionales regulares o mínimas. Los hogares suelen tener entre 1 y 2 unidades familiares, con un promedio de 3 a 4 personas por hogar.

Un aspecto destacado de este *cluster* es la baja participación infantil, con porcentajes de 0.95% y 0.75%, junto con el acceso extremadamente limitado a Centros de Desarrollo Infantil (CDI), con coeficientes de 0.0097 y 0.0025, los más bajos registrados. Esto indica una población con poca presencia de niños y un acceso muy restringido a servicios clave para la infancia. En conclusión, el c*luster* -1 concentra comunidades con importantes carencias sociales, económicas y de infraestructura, lo que exige intervenciones urgentes para mejorar el acceso a servicios básicos y apoyar a las familias en situación de vulnerabilidad.

***Cluster* 0:**

Compuesto por comunidades ubicadas en las comunas 6 y 1, destacando barrios como La Floralia y Terron Colorado. Los hogares pertenecen mayoritariamente al estrato socioeconómico 1, seguido del estrato 2, lo que refleja condiciones de bajos recursos. La mayoría de las viviendas son propias o son Ocupante de hecho, con clasificaciones habitacionales 4 y 3, lo que sugiere mejores condiciones de vivienda en comparación con el

c*luster* -1. Los hogares suelen tener 1 o 2 unidades familiares, con un promedio de 3 o 4 personas por hogar.

En términos de participación infantil, los valores son ligeramente mayores que en el c*luster* -1, con porcentajes de 1.89% y 1.86%. Además, el acceso a Centros de Desarrollo Infantil (CDI) muestra una ligera mejora, con coeficientes de 0.0108 y 0.0211, indicando un acceso moderado a servicios infantiles. En conclusión, el c*luster* 0 refleja comunidades de estrato bajo que presentan una mejor proporción infantil y un acceso algo más adecuado a servicios esenciales, aunque persisten retos significativos que requieren atención para fortalecer el desarrollo infantil y mejorar las condiciones de vida.

***Cluster* 1:**

Compuesto por comunidades ubicadas en las comunas 20 y 18, destacando los barrios Siloé y Alto de Los Chorros. Los hogares pertenecen mayoritariamente al estrato socioeconómico 1, seguido del estrato 2, lo que indica condiciones económicas bajas o medio-bajas. Las viviendas son en su mayoría propias o son ocupantes de hecho, con clasificaciones habitacionales 3 y 2, reflejando condiciones regulares o marginales. Los hogares suelen estar compuestos por 1 o 2 familias, con un promedio de 3 a 4 personas por hogar.

En cuanto a la población infantil, la proporción es moderada, con valores de 2.36% y 1.64%. El acceso a Centros de Desarrollo Infantil (CDI) sigue siendo limitado, con coeficientes de 0.0143 y 0.0148, lo que indica carencias en los servicios para la primera infancia. En conclusión, el *Cluster* 1 refleja comunidades con condiciones intermedias, caracterizadas por una proporción infantil moderada y acceso restringido a servicios infantiles, lo que sugiere la necesidad de programas sociales enfocados en mejorar la infraestructura y el desarrollo infantil en estas áreas.

***Cluster* 2:**

Compuesto por comunidades ubicadas en las comunas 21 y 14, con barrios destacados como Potrero Grande y Las Garzas. Los hogares pertenecen principalmente al estrato socioeconómico 1, seguido del estrato 2, lo que indica condiciones económicas bajas o medio-bajas. La mayoría de las viviendas son propias o en ocupación de hecho, con clasificaciones habitacionales 3 y 2, similares al c*luster* 1, lo que sugiere condiciones regulares o marginales. Los hogares tienen entre 1 y 2 familias, con un promedio de 3 a 4 personas por hogar.

Este *cluster* se caracteriza por una alta concentración infantil en algunos sectores, con una participación de niños del 3.49% en ciertos barrios y del 1.35% en otros. Sin embargo, el acceso a Centros de Desarrollo Infantil (CDI) es extremadamente limitado, con coeficientes de 0.0096 y 0.0055, los más bajos registrados. En conclusión, el c*luster* 2 combina una proporción infantil significativa con un acceso muy restringido a servicios esenciales, lo que resalta la necesidad de intervenciones urgentes en infraestructura y programas sociales dirigidos a la infancia.

***Cluster* 3:**

Compuesto por comunidades ubicadas en las comunas 21 y 20, con barrios principales como Calimio e Invicali. Los hogares pertenecen mayoritariamente al estrato socioeconómico 1, seguido del estrato 2, lo que indica condiciones económicas bajas. La mayoría de las viviendas son propias o son ocupantes de hecho, con clasificaciones habitacionales 3 y 4, reflejando mejores condiciones que el **Cluster** 2. Los hogares están compuestos por 1 o 2 familias, con un promedio de 3 a 4 personas por vivienda.

La participación infantil es baja a moderada, con valores de 1.93% y 1.41%, mientras que el acceso a Centros de Desarrollo Infantil (CDI) es desigual, con coeficientes de 0.0096 y 0.0265, mostrando una mejora relativa en comparación con *clusters* más vulnerables. En conclusión, el *Cluster* 3 representa comunidades con proporciones infantiles moderadas y

un acceso variable a servicios infantiles, lo que sugiere la necesidad de fortalecer la equidad en el acceso a programas y servicios de desarrollo infantil en estas áreas.

***Cluster* 4:**

El c*luster* 4 está compuesto por comunidades ubicadas en las comunas 15 y 16, destacando los barrios Mojica y Bajos Ciudad Córdoba. Los hogares pertenecen mayoritariamente al estrato socioeconómico 2, seguido por el estrato 1, lo que refleja condiciones económicas moderadas. Las viviendas son en su mayoría propias o por ocupación de hecho, con clasificaciones habitacionales 4 y 3, que indican condiciones de vida moderadas. Los hogares están compuestos por 1 o 2 familias, con un promedio de 3 a 4 personas por vivienda.

Este *cluster* se caracteriza por una alta proporción infantil, con participación de niños de 3.59% y 3.51%, y un acceso a Centros de Desarrollo Infantil (CDI) considerado moderado, con coeficientes de 0.0224 y 0.0138. En conclusión, el c*luster* 4 refleja comunidades con una significativa presencia de población infantil y un acceso razonable a servicios infantiles, aunque se sugiere continuar mejorando la cobertura y calidad de los programas destinados al desarrollo infantil en estas áreas.

***Cluster* 5:**

Compuesto por comunidades ubicadas en las comunas 15 y 14, con barrios principales como Comuneros y Promociones Populares. Los hogares pertenecen mayoritariamente al estrato socioeconómico 1, seguido por el estrato 0, lo que indica condiciones económicas muy limitadas. La mayoría de las viviendas son propias (ocupación 1) o tienen otras formas de ocupación (ocupación 4), con clasificaciones habitacionales 3 y 4, similares al c*luster* 4, que reflejan condiciones moderadas. Los hogares suelen tener entre 1 y 2 familias, con un promedio de 3 a 4 personas por vivienda.

Este *cluster* presenta una alta concentración infantil, con porcentajes de participación de niños de 3.87% y 2.58%, y un acceso a Centros de Desarrollo Infantil (CDI) relativamente

bueno, con coeficientes de 0.0241 y 0.0238. En conclusión, el c*luster* 5 combina una elevada proporción de población infantil con un acceso adecuado a servicios infantiles, lo que sugiere que las intervenciones deben enfocarse en fortalecer y mantener estas condiciones para garantizar el desarrollo integral de la infancia en estas comunidades vulnerables.

***Cluster* 6:**

Compuesto por comunidades ubicadas en las comunas 13 y 15, con barrios principales como El Vergel y El Poblado. Los hogares pertenecen mayoritariamente al estrato socioeconómico 2, seguido por el estrato 1, lo que refleja condiciones económicas bajas o medio-bajas. Las viviendas son predominantemente propias (ocupación 1) o presentan otras formas de ocupación (ocupación 4), con clasificaciones habitacionales 4 y 6, lo que sugiere condiciones más variadas que en otros *clusters*. Los hogares están compuestos por 1 o 2 familias, con un promedio de 3 a 4 personas por vivienda.

En cuanto a la participación infantil, esta es menor en comparación con los *clusters* 4 y 5, con porcentajes de 2.31% y 1.55%. El acceso a Centros de Desarrollo Infantil (CDI) es moderado, con coeficientes de 0.0161 y 0.0289, lo que indica una cobertura razonable en comparación con *clusters* más vulnerables. En conclusión, el c*luster* 6 se caracteriza por una menor proporción de niños y un acceso relativamente adecuado a servicios infantiles, lo que sugiere que las intervenciones pueden enfocarse en reforzar estos servicios y garantizar su sostenibilidad en el tiempo.

***Cluster* 7:**

Compuesto comunidades ubicadas en las comunas 14 y 13, destacando los barrios Manuela Beltrán y Alfonso Bonilla Aragón. Los hogares pertenecen mayoritariamente al estrato socioeconómico 1, seguido por el estrato 2, lo que refleja condiciones económicas bajas o medio-bajas. La mayoría de las viviendas son propias (ocupación 1) o presentan otras formas de ocupación (ocupación 4), con clasificaciones habitacionales 3 y 4, similares a las

del *Cluster* 4, lo que indica condiciones moderadas. Los hogares están compuestos por 1 o 2 familias, con un promedio de 3 a 4 personas por vivienda.

Este *cluster* se distingue por tener la mayor proporción infantil entre los analizados, con participaciones de 4.06% y 2.87%, así como el mejor acceso a Centros de Desarrollo Infantil (CDI), con coeficientes de 0.0271 y 0.0416. En conclusión, el c*luster* 7 destaca por su alta concentración de población infantil y un acceso notablemente superior a servicios infantiles, lo que lo posiciona como un referente en la atención a la infancia. Es clave garantizar la sostenibilidad y ampliación de estos servicios para mantener y mejorar las condiciones de desarrollo infantil.

En síntesis, se evidencia a razón del modelo la vulnerabilidad en los *clusters* -1 y 6, debiéndose tratar enfocado en mejorar la cobertura de Centros de Desarrollo Infantil (CDI) e implementar programas sociales que incentiven el cuidado infantil temprano y fortalezcan las condiciones socioeconómicas de las familias. Para los c*lusters* 0 y 5, es necesario fortalecer el acceso y ampliar la capacidad de los servicios infantiles, adaptándolos a la alta concentración infantil en estas comunidades. Finalmente, en los c*lusters* 4 y 7, se sugiere mantener y expandir la calidad de los servicios existentes, utilizando estos *clusters* como modelos para replicar estrategias exitosas que promuevan el desarrollo infantil integral en otras áreas.

9. IMPACTOS DEL PROYECTO

Este proyecto permite una mejor focalización de programas sociales, logrando que la asignación de recursos sea más eficiente y reduciendo significativamente la exclusión social. Esto permitirá diseñar políticas adaptadas a las necesidades específicas de cada microterritorio, garantizando mayor precisión en la planificación territorial y abordando las particularidades de cada comunidad. Además, la implementación de estrategias focalizadas contribuirá a la reducción de la pobreza y la desigualdad, especialmente mediante políticas dirigidas a combatir la pobreza extrema y la distribución equitativa de ayudas. Este enfoque asegura que las comunidades más vulnerables reciban la atención prioritaria que necesitan.

También fomenta la calidad de vida en la primera infancia, promoviendo el acceso equitativo a servicios de salud y educación y reduciendo indicadores críticos como la desnutrición infantil. La implementación de políticas basadas en el monitoreo continuo permitirá evaluar el impacto de los CDI, ajustando estrategias en tiempo real para maximizar resultados. La planificación estratégica del Estado se verá fortalecida al priorizar inversiones a largo plazo y optimizar recursos, lo que también reducirá la migración motivada por la falta de bienestar y oportunidades. Finalmente, este enfoque fomentará la transparencia y rendición de cuentas, promoviendo la confianza pública y alentando la investigación e innovación social, generando políticas basadas en la ciencia de datos para un desarrollo más equitativo y sostenible.

Por último, es importante mencionar que el fortalecimiento de las UDS (Unidades de Servicio) del ICBF es clave para garantizar el bienestar integral de los niños, niñas y adolescentes, mejorando su calidad de vida mediante el acceso equitativo a servicios esenciales como nutrición, educación y protección. Es indispensable identificar las áreas de mejora en la gestión y atención, permitiendo ajustes oportunos que incrementen la eficacia de la atención de la población. Finalmente, este enfoque también optimiza el uso de los recursos disponibles, orientando las inversiones hacia soluciones sostenibles que impacten positivamente en las comunidades atendidas, asimismo, fomenta la transparencia y la innovación en la toma de decisiones, utilizando herramientas basadas en evidencia para desarrollar políticas más inclusivas y con impacto a largo plazo, fortaleciendo la confianza en las instituciones.

10. CONCLUSIONES

El método K-means demostró ser poco útil para el desarrollo del objetivo del proyecto en cuanto al análisis de vulnerabilidad en la primera infancia, ya que los *clústers* generados no presentaron una delimitación clara, muchos puntos quedaron cercanos a los límites entre los grupos, lo que dificultó la identificación de patrones bien definidos. Además, el Coeficiente de Silhouette indicó que la calidad del agrupamiento fue moderada,

evidenciando que los datos no se separaron de manera efectiva en grupos distintos. Esto se debió a que el modelo *K-means* no logró capturar la estructura compleja de los datos, como relaciones no lineales o distribuciones desiguales. Por esta razón nos centramos en los modelos de *OPTICS* Y HDBSCAN y se decide no continuar con este modelo para continuar con el desarrollo del proyecto.

Es importante resaltar que la teoría estudiada dentro del contexto del marco teórico, indica que los modelos no supervisados de mayor rendimiento para trabajar con datos georefenciados son *Optics* y HDBSCAN, teoría que fue acertada en los resultados que se obtuvieron en cuanto a la formación de *clusters* y tamaño de puntos ruido, adicionalmente a que están diseñados para superar las limitaciones de otros algoritmos más simples, como K-means, ya que adaptarse a la naturaleza no uniforme de los datos, ofrecen soluciones más robustas frente a datos ruidosos o con densidades variables que es la particularidad de la data trabajada, y dado que se contaba con data georreferenciada, los algoritmos seleccionados finalmente se comportan mejor que K-means puesto que suelen ser más adecuados debido a su capacidad para detectar *clusters* con formas y densidades arbitrarias.

Para concluir sobre los modelos finalmente seleccionados sobre HDBSCAN se obtuvieron resultados iniciales de 33 *clusters* con parámetros Min_cluster_size (25) y un ruido de (141) equivalente a un 0.26% de la data toral, se realizó refinamiento del modelo obteniendo 8 *clusters* con Min_cluster_size de (2.595) y un ruido de (4.453) equivalente a un 8.34% de la data total. Por consiguiente, este último resultado permite realizar una comparativa en las similitudes en los resultados del modelo *Optics*. El resultado inicial de los 33 *clusters*, aunque generaba el menor nivel de ruido de todas las iteraciones generadas, el resultado gráfico de los grupos en mapa de la ciudad de Cali, con la herramienta tableau es el solapamiento de los *clusters*, situación que reduce la confiabilidad en la conformación de los grupos y se hace difícil la diferenciación de estos.

En cuanto al modelo *Optics*, inicialmente se tuvieron como resultado 4 *clusters* con un ruido de 11.926 puntos equivalente a 22.34% de la data, posterior a ello después de realizar un refinamiento en los parámetros y refinamiento de la data, se obtiene un resultado de 9

clusters con un ruido de 2.893 equivalente a 5,4% de la data; se observa que este modelo genera un valor más pequeño de puntos ruido, finalmente por los resultados de agrupamiento y su visualización sobre el mapa de la ciudad de Cali sin solapamiento entre los grupos, es el modelo más acorde y recomendado para la solución del proyecto.

Decidir evaluar estos 3 modelos como K-means, *Optics* y HDBSCAN para el desarrollo de este proyecto, no fue solo por orientación de la dirección del proyecto, sino que también fue indispensable para identificar las fortalezas y debilidades de cada uno de estos, dado que podría impactarse directamente la en la calidad del análisis y los resultados finales del proyecto. Mediante esta evaluación se asegura que los modelos finalmente elegidos, están alineados con las prioridades y limitaciones del proyecto, evitar errores en el análisis, justificar las decisiones que se tomaron y, sobre todo, obtener los resultados útiles y confiables que finalmente están expuestos en este documento.

11. TRABAJOS FUTUROS

Este proyecto permitió evidenciar falencias en la prestación de los servicios asociados a los CDI'S, teniendo en cuenta la información que se obtuvo no solo de los modelos finales sino de la data en general, a modo de ejemplo se evidencio que una cantidad importan de CDI´S no tienen niños de la primera infancia asignados dentro de la ubicación cercana a ellos, por lo que a futuro puede desarrollarse un proyecto que se centre en la optimización de la asignación de recursos en Centros de Desarrollo Infantil (CDI) mediante el uso de la ciencia de datos para la mejora de la atención en la primera infancia, permitiendo la eficiencia de los recursos asignados a los territorios, ya que se parte del postulado que son limitados y deben administrarse de manera óptima para maximizar el alcance a las comunidades más vulnerables no solo de la ciudad de Cali sino del país en general. Actualmente, la asignación de recursos se realiza de manera general, sin un análisis detallado de las necesidades específicas de cada territorio, región o centro y las poblaciones objetivo a su alrededor.

Bibliografía

[1] «UNICEF,» El desarrollo en la primera infancia: El momento más importante de la vida. Fondo de las Naciones Unidas para la Infancia., 2017. [En línea]. Available: https://www.unicef.org/.

[2] S. J. P y D. A. Phillips, From Neurons to Neighborhoods: The Science of Early Childhood Development, Washington, DC: National Academies Press, 2000.

[3] K. P. Murphy, Machine Learning: A Probabilistic Perspective, Cambridge, MA: MIT Press, 2012.

[4] F. Provost y T. Fawcett, Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking, Sebastopol, CA: O'Reilly Media, 2013.

[5] G. James, D. Witten, T. Hastie y R. Tibshirani, An Introduction to Statistical Learning: With Applications in R, Nueva York: Springer, 2013.

[6] C. O'Neil y R. Schutt, Doing Data Science: Straight Talk from the Frontline, Sebastopol, CA: O'Reilly Media, 2013.

[7] F. d. l. N. U. p. l. I. (UNICEF), «Estado de la Niñez en Colombia: Avances y Desafíos,» 2019. [En línea]. Available: https://unicef.org.co/sitan/assets/pdf/sitan.pdf?utm_source=chatgpt.com.

[8] M. d. E. N. d. Colombia, «Estrategia de atención integral a la primera infancia.,» 2009. [En línea]. Available: https://www.mineducacion.gov.co/1759/articles-177829_archivo_pdf_fundamentos_ceroasiempre.pdf.

[9] I. C. d. B. F. (ICBF), «Lineamiento Técnico para la Atención a la Primera Infancia,» 2020. [En línea]. Available: https://www.icbf.gov.co/system/files/procesos/lm5.pp_lineamiento_tecnico_para_la_atencion_a_la_primera_infancia_v7.pdf.

[10] M. d. E. N. d. Colombia, «Actividades rectoras de la primera infancia y de la educación inicial,» [En línea]. Available: https://www.mineducacion.gov.co/primerainfancia/1739/w3-article-178032.html.

[11] P. u. J. d. Cali, «El mejor lugar para cre-ser,» [En línea]. Available: https://elmejorlugarparacreser.javerianacali.edu.co/.

[12] UNESCO, «Invertir en la atención y educación de la primera infancia proporciona beneficios para toda la vida,» 2024. [En línea]. Available: https://www.unesco.org/es/articles/invertir-en-la-atencion-y-educacion-de-la-primera-infancia-proporciona-beneficios-para-toda-la-vida.

[13] C. E. Maya escolbar y M. A. García Pérez, «Accesibilidad a Servicios de Salud y Determinantes Sociales en la Primera Infancia en Cali,» 2021. [En línea]. Available: https://repositorio.uniajc.edu.co/bitstreams/4ad605b6-8e6b-42d0-8453-c0b441d577f8/download.

[14] Ministerio de Educación Nacional, «Ministerio de Educación Nacional - Prosperidad para todos,» Gobierno de Colombia, [En línea]. Available: https://www.mineducacion.gov.co/primerainfancia/1739/article-177827.html. [Último acceso: 29 11 2023].

[15] Gobierno de Colombia - Corte Constitucional de Colombia, Constitución Política de Colombia, Bogotá D.C.: Consejo Superior de la Judicatura, Sala Administrativa.

[16] MInisterio de Salud, «Colombia Potencia de Vida,» Gobierno de Colombia, [En línea]. Available: https://www.minsalud.gov.co/atencion/Paginas/transparencia-acceso-informacion.aspx. [Último acceso: 19 11 2023].

[17] Instituto Colombiano de Bienestar Familiar (ICBF), «Bienestar Familiar,» Gobierno de Colombia, 2021. [En línea]. Available: https://www.icbf.gov.co/portafolio-de-servicios-icbf/centro-de-desarrollo-infantil. [Último acceso: 20 11 2023].

[18] M. d. S. y. p. social, «Lineamientos para la organización y operación de los equipos básicos de salud,» Febrero 2023. [En línea]. Available: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/PP/lineamientos-operacion-equipos-basicos-salud-resolucion-2788-2022.pdf. [Último acceso: 8 Octubre 2024].

[19] AMAZON, «AWS-AMAZON,» Amazon Web Services, 2023. [En línea]. Available: https://aws.amazon.com/es/what-is/geospatial-data/#:~:text=Los%20datos%20geoespaciales%2C%20tambi%C3%A9n%20conocidos,la%20superficie%20de%20la%20Tierra. . [Último acceso: 20 11 2023].

[20] evaluandosoftware, «evaluandosoftware,» Ingenima, 2023. [En línea]. Available: https://www.evaluandosoftware.com/bpm/la-geolocalizacion-funciona/. [Último acceso: 19 11 2023].

[21] SI-GEO, «SI-GEO Sistema de Información Geográfico del Sector Educativo,» Ministerio de Educación , [En línea]. Available: https://www.mineducacion.gov.co/1621/article-190610.html. [Último acceso: 23 04 2024].

[22] ArcGis, «ArcGis,» esri, 2021. [En línea]. Available: https://desktop.arcgis.com/es/arcmap/latest/manage-data/shapefiles/what-is-a-shapefile.htm. [Último acceso: 19 11 2023].

[23] DANE, «DANE,» 2018. [En línea]. Available: https://www.dane.gov.co/files/sen/lineamientos/manual-uso-marco-geoestadistico-nacional-en-proceso-estadistico.pdf. [Último acceso: 22 11 2023].

[24] Arimetrics, «Arimetrics,» Arimetrics, 2022. [En línea]. Available: https://www.arimetrics.com/glosario-digital/analitica-descriptiva. [Último acceso: 21 11 2023].

[25] IBM, «Topics,» IBM, [En línea]. Available: https://www.ibm.com/mx-es/topics/predictive-analytics. [Último acceso: 21 11 2023].

[26] IBM, «ibm.com,» IBM, [En línea]. Available: https://www.ibm.com/es-es/topics/machine-learning#Machine%20Learning%20Supervisado. [Último acceso: 01 12 2023].

[27] J. A. Manrique, «Predicción de la demanda de Smartphone de introducción al mercado Colombiano mediante modelos de Machine Learning.,» Fundación Universitaria Konrad Lorenz,, 2022.. [En línea]. Available: https://repositorio.konradlorenz.edu.co/.

[28] J. L. R. Pérez, «Técnicas de aprendizaje automático para la detección de intrusos en redes de computadoras,» *Revista Cubana de Ciencias Informáticas,* vol. 8, nº 4, 2014.

[29] C. E. Guilcapi Lopez y M. E. Montero Arias, «Clusters espaciales de la economía de los socios de la Coop. "Educadores de Chiborazo" Ltda (2016-2020),» *Escuela Superior Politécnica de Chimborazo,* 2022.

[30] J. Gillard, E. O'Riordan y A. Zhigljavsky, «Simplicial and minimal-variance distance in multivariate data analysis,» *Journal of Statistical Theory and practice 16.1,* pp. 9-11, 2022.

[31] L. Kaufman y P. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis., John Wiley & Sons., 2009.

[32] P. N. Tan, S. Michael y K. Vipin, Introduction to Data Mining., Pearson., 2006.

[33] D. Maesschalck, R. Jouan y M. D, «The Mahalanobis distance,» *Chemometrics and Intelligent Laboratory Systems,* pp. 1-18, 2000.

[34] J. Ward, «Hierarchical Grouping to Optimize an Objective Function,» *Journal of the American Statistical Association,* vol. 58, nº 301, pp. 236-244, 1963.

[35] N. &. A. G. Andrienko, Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach., Springer., 2006.

[36] Z. D. B. H. J. &. K. R. Li, «Swarm: Mining Relaxed Temporal Moving Object Clusters.,» Proceedings of the VLDB Endowment, 2010.

[37] L. McInnes, J. Healy y S. Astels, «Hierarchical Density-Based Clustering.,» *Journal of Open Source Software,* vol. 2, nº 11, p. 205, 2017.

[38] R. Campello, D. Moulavi y J. Sander, «Density-Based Clustering Based on Hierarchical Density Estimates,» *Pacific-Asia Conference on Knowledge Discovery and Data Mining,* pp. 160-172, 2013.

[39] J. C. M. Betancourt, «Free Code Camp,» 24 Abril 2021. [En línea]. Available: https://www.freecodecamp.org/espanol/news/8-algoritmos-de-agrupacion-en-clusteres-en-el-aprendizaje-automatico-que-todos-los-cientificos-de-datos-deben-conocer/. [Último acceso: 09 10 2024].

[40] J. Jan, M. Kamber y J. Pei, Data Mining: Concepts and techniques, Morgan, 2011.

[41] L. Kaufman y P. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, 1990.

[42] A. E. Ezugwu, A. K. Shukla, M. B. Agbaje, O. N. Oyelade, A. J. García y J. O. Agushaka, «Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature,» *Neural Computing and Applications,* vol. 11, pp. 6247-6306, 2020.

[43] J. Bergstra y Y. Bengio , «Random Search for Hyper-Parameter Optimization,» *Journal of Machine Learning,* vol. 13, pp. 281-305, 2012.

[44] F. Hutter, H. H. Hoos y K. Leyton, «Sequential Model-Based Optimization for,» *University of British Columbia.*

[45] P. Haya, «Instituto de ingenieria del conocimiento,» [En línea]. Available: https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/. [Último acceso: 09 octubre 2024].

[46] Google, «Google for developers,» [En línea]. Available: https://developers.google.com/machine-learning/clustering/overview?hl=es-419. [Último acceso: 1 12 2023].

[47] Universidad Nacional de La Plata, «OAS,» Facultad de Informática , 2016. [En línea]. Available: http://163.10.22.82/OAS/Agrupamiento_Kmedias/definicin.html. [Último acceso: 02 12 2023].

[48] Pontificia Universidad Javeriana Cali - Propacífico, «El mejor lugar para creSer,» Riqueza Completa, Centro de Investigación Aplicada - PUJC, [En línea]. Available: https://elmejorlugarparacreser.javerianacali.edu.co/mapa-del-valle/#:~:text=El%20Mejor%20Lugar%20para%20creSER,y%20el%20Norte%20del%20Cauca. [Último acceso: 15 11 2023].

[49] J. P. Álvarez, L. Giraldo Huertas y A. Cano , «Desarrollo socio-cognitivo en la primera infancia: los retos por cumplir en salud pública en la zona Sabana Centro y Boyacá,» *Rev. salud pública 19 (4) Jul-Aug 2017,* vol. 19, nº 4, 2017.

[50] Cobaleda Martínez, Diego Andrés - Pontificia Universidad Javeriana Bogotá, «Repositorio Javeriana,» Pontificia Universidad Javeriana - Bogotá D.C.- Colombia, 2016. [En línea]. Available: https://repository.javeriana.edu.co/handle/10554/21024. [Último acceso: 17 11 2023].

[51] Secreatría Distrital de Planeación - Bogotá, «Secretaría de Planeación,» Gobierno de Colombia, [En línea]. Available: https://www.sdp.gov.co/transparencia/informacion-interes/glosario/entidad-promotora-de-salud-eps. [Último acceso: 21 11 2023].

[52] IBM, «IBM.com,» IBM, [En línea]. Available: https://www.ibm.com/mx-es/topics/knn. [Último acceso: 02 12 2023].

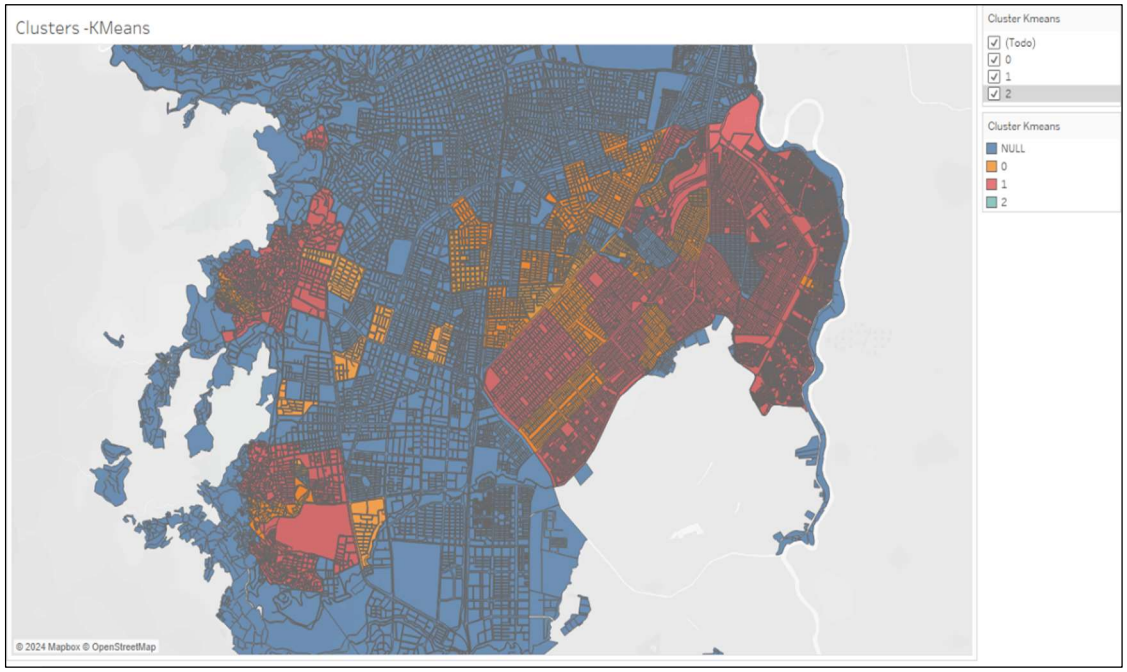
[53] R. J. G. B. M. D. &. S. J. Campello, «Density-Based Clustering Based on Hierarchical Density Estimates.,» *Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD).,* 2013.

Herramientas de programación:

- Python es un lenguaje de programación de alto nivel, interpretado y de propósito general.
    - Bibliotecas de Python que permitieron realizar un análisis eficiente de los datos. Entre ellas se encuentran Pandas, NumPy y Matplotlib.
    - Pandas: es una biblioteca poderosa y flexible que brinda herramientas para manipular y analizar datos de manera eficiente. Permite trabajar con estructuras de datos como DataFrames, lo cual facilita la limpieza, transformación y exploración de los datos. Con Pandas, se realizaron operaciones como la selección y filtrado de datos, cálculos estadísticos, y fusionar conjuntos de datos para un análisis más completo.
    - NumPy se enfoca principalmente en la manipulación de matrices y cálculos numéricos. Esta biblioteca es fundamental para el análisis de datos, ya que proporciona funciones y métodos eficientes para realizar operaciones matemáticas y estadísticas en nuestros datos. Con NumPy, se realizaron cálculos numéricos complejos, trabajar con matrices multidimensionales y realizar manipulaciones y transformaciones de datos de manera eficiente.
    - Matplotlib para la visualización de datos. Matplotlibes una biblioteca ampliamente utilizada y poderosa que permite crear una variedad de gráficos y visualizaciones para explorar y comprender mejor los datos. Con Matplotlib, se pudo crear gráficos de líneas, gráficos de barras, histogramas, diagramas de dispersión y muchas otras visualizaciones que ayudaron a identificar patrones, tendencias y relaciones en los datos.
    - Autoviz es una biblioteca de visualización automática basada en Python que puede crear una variedad de tipos de gráficos según un conjunto de datos. El objetivo de su diseño es mejorar la eficiencia y la comodidad del proceso de exploración y análisis de datos, eliminando la necesidad de escribir manualmente un código de visualización extenso.
    - Sweetviz es una biblioteca poderosa de Python destinada a mejorar el proceso de análisis exploratorio de datos. Sweetviz crea visualizaciones detalladas que muestran información sobre diferentes aspectos de su conjunto de datos con solo dos líneas de código. Sweetviz ofrece una solución de exploración de datos fácil de usar que puede usarse para analizar valores objetivo, comparar conjuntos de datos o examinar características de funciones.
    - DTale es una biblioteca de Python que nos permite ver un DataFrame de Pandas. D-Tale produce una interfaz gráfica que es interactiva. D-Tale proporciona una variedad de detalles a los datos proporcionados. La mayoría de los formatos de

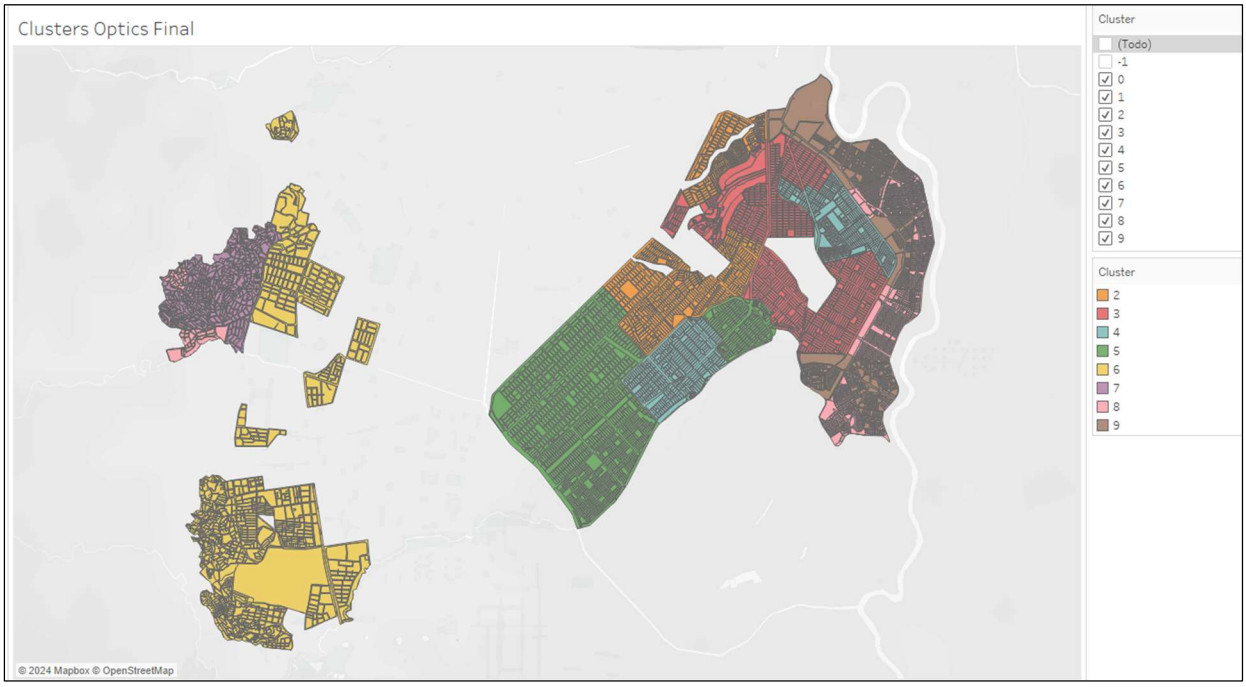
archivo que admite son CSV, TSV, XLS y XLSX. Es una biblioteca basada en Python que utiliza Flask como frontend y React como backend.

ANEXO 1: Herramienta de visualización Objetivo 4: Gráfica en tableau de *clusters* generados por el modelo K-means.



ANEXO 2: Herramienta de visualización Objetivo 4: Gráfica en tableau de *clusters* generados por el modelo HDBSCAN #33 Clusters.

ANEXO 3: Herramienta de visualización Objetivo 4: Gráfica en tableau de *clusters* generados por el modelo HDBSCAN #8 *Clusters.*

ANEXO 4: Herramienta de visualización Objetivo 4: Gráfica Visualización en tableau de *clusters* generados por el modelo *OPTICS Modelos A y B.*

ANEXO 5 Herramienta de visualización Objetivo 4: Visualización *Clusters* Modelo *Optics* con refinamiento de información en el modelo C.