Proyecto de IA Aplicada:
Predicción de Educación Primaria en Niños con Autismo

Ruben Dario Sabogal U, rdsabogal@hotmail.com
Cristian Camilo Quebrada, cristian_q_90@hotmail.com
Edwin Pérez L., edwinandperez@gmail.com

Resumen— Este proyecto aborda el subregistro y la inconsistencia en la categorización de estudiantes con Trastorno del Espectro Autista (TEA) en el Sistema de Matrícula (SIMAT) de Cali, así como la débil articulación con los Planes Individuales de Ajustes Razonables (PIAR) y los soportes del sector salud. Se plantea como objetivo general fortalecer la identificación, el registro y el seguimiento de la población con TEA para mejorar el acceso, la permanencia y la pertinencia de los apoyos educativos. La metodología combina: (i) evaluación de calidad de datos (completitud, coherencia y homologación de variables críticas), (ii) estandarización de criterios y evidencias para la categorización TEA y su alineación PIAR↔SIMAT, (iii) diseño de tableros e indicadores por año, comuna, IEO y grado, con panel de calidad, y (iv) un pipeline analítico opcional (EDA + métricas ROC/PR, sesgos, interpretabilidad) como apoyo al tamizaje, con resguardo ético y de privacidad. El plan de trabajo (septiembre–noviembre de 2025) incluye entregables de datos, análisis, documentación operativa, capacitación y cierre con plan de sostenibilidad y KPIs. Se espera disminuir tiempos de acceso a apoyos, reducir riesgo de deserción y mejorar la focalización territorial, mediante una gobernanza clara de roles, validaciones y plazos intersectoriales.

Palabras clave— Educación inclusiva, TEA, SIMAT, PIAR, calidad de datos, analítica educativa, ética de datos.

I. INTRODUCCION

La identificación y el acompañamiento oportuno de estudiantes con Trastorno del Espectro Autista (TEA) constituyen un reto técnico y de gestión para los sistemas de información educativa. En el contexto de Cali, se ha detectado subregistro y falta de consistencia en la categorización de TEA dentro del Sistema Integrado de Matrícula (SIMAT), además de una débil articulación con los Planes Individuales de Ajustes Razonables (PIAR) y con los soportes del sector salud; esta situación afecta la planificación de apoyos, el seguimiento a la permanencia y la focalización territorial por comuna e institución (IEO).

Fig.1. Que es el autismo

Fig. 2. Sistema Integrado de matricula

Este informe propone un marco de intervención para mejorar la calidad del registro y del seguimiento de estudiantes con TEA en SIMAT, mediante: (i) evaluación de la calidad y coherencia de variables críticas; (ii) estandarización de criterios y evidencias para la categorización (incluida la clasificación como discapacidad psicosocial cuando aplique) y definición clara de responsabilidades entre IEO y entidad territorial certificada (ETC); y (iii) articulación operativa entre SIMAT y PIAR con respaldo del sector salud.

Desde el componente analítico-tecnológico, se plantea la implementación de análisis exploratorio de datos (EDA) sistemático, tableros con filtros por año, comuna, IEO y grado—incluyendo un panel de calidad—y, de manera

opcional, un pipeline de “screening” con métricas y validación ética para apoyar decisiones sin reemplazar criterios pedagógicos ni clínicos.

Fig. 3. ¿Qué es el PIAR?

Fig. 4. El PIAR

Las contribuciones de este trabajo son: (1) un diagnóstico estructurado del problema (árbol de causas-efectos) que explica demoras en apoyos, riesgo de deserción y desigualdades de inclusión; (2) un conjunto de objetivos y lineamientos operativos para alinear datos, procesos y responsabilidades intersectoriales; y (3) un plan de trabajo con entregables y cronograma (septiembre–noviembre de 2025) que integra datos, análisis, tableros, gestión del cambio y cierre con plan de sostenibilidad e indicadores clave.

El resto del documento se organiza así: la Sección 2 sintetiza el estado del arte y el marco normativo aplicable; la Sección 3 describe la metodología de evaluación y estandarización de datos; la Sección 4 presenta el diseño de tableros y el pipeline analítico opcional; la Sección 5 detalla el plan de trabajo y entregables; y la Sección 6 discute resultados esperados, consideraciones éticas y limitaciones.

1. Diagnóstico del Problema:

Existe un subregistro y falta de consistencia en la categorización de estudiantes con Trastorno del Espectro Autista (TEA) en el sistema SIMAT, acompañado de una débil articulación con el PIAR y los soportes del sector salud. Esta situación afecta negativamente la planificación educativa, la asignación de apoyos, el seguimiento a la permanencia escolar y la focalización territorial (comuna/IEO). **2. Objetivos del Proyecto**

Objetivo general:

Mejorar la identificación, el registro y el seguimiento de estudiantes con TEA en SIMAT en la ciudad de Cali, fortaleciendo el acceso, la permanencia y la pertinencia de los apoyos educativos (PIAR), mediante una articulación efectiva entre los sectores de educación y salud. **Objetivos específicos:**

- Evaluar la calidad, completitud y coherencia de las variables relevantes en SIMAT (discapacidad/TEA, apoyos, IEO, comuna, grado, etc.) y su alineación con la ruta de registro.

- Estandarizar criterios y evidencias para la categorización de TEA (como discapacidad psicosocial cuando corresponda), definiendo responsabilidades y validaciones en IEO y ETC.

- Garantizar la articulación entre SIMAT y PIAR (ajustes razonables, ayudas técnicas) con respaldo del sector salud. • Diseñar y validar un pipeline analítico (EDA + indicadores, con opción de ML) para el tamizaje y seguimiento, con enfoque ético y de privacidad.

- Implementar tableros y reportes periódicos por año, comuna, IEO y grado, incluyendo un panel de calidad de datos.

3. Árbol de Problemas

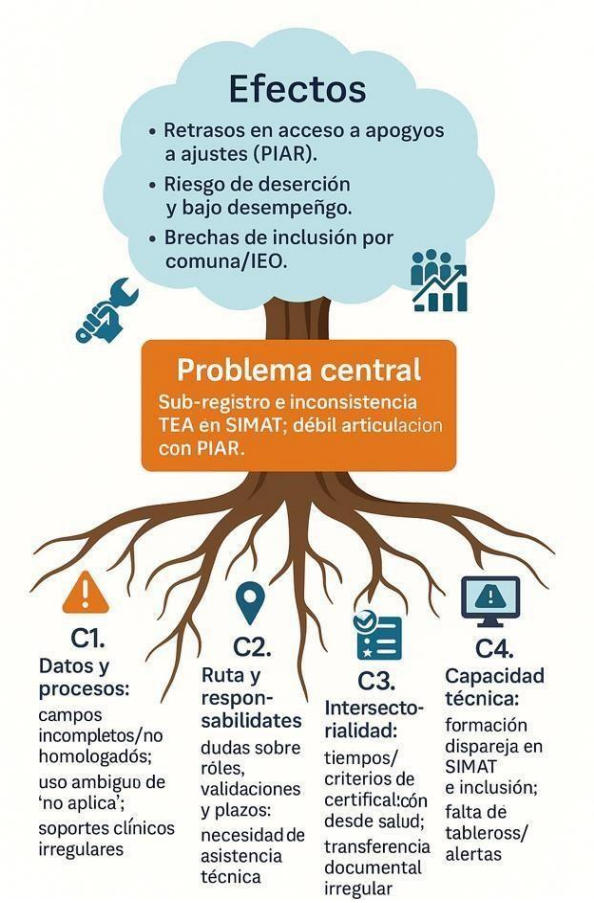


Fig. 5. árbol de Problemas

Efectos:

• Demoras en el acceso a apoyos y ajustes (PIAR).

• Riesgo de deserción escolar y bajo rendimiento académico.

• Desigualdades en inclusión educativa según comuna o IEO.

Causas principales:

• **C1. Datos y procesos**: Campos incompletos o no homologados, uso ambiguo de "no aplica", soportes clínicos irregulares.

- **C2. Ruta y responsabilidades**: Falta de claridad en roles, validaciones y plazos; necesidad de asistencia técnica.

- **C3. Intersectorialidad**: Tiempos y criterios de certificación desde salud; transferencia documental inconsistente.

- **C4. Capacidad técnica:** Formación desigual en SIMAT e inclusión; ausencia de tableros y alertas.

4. Estado del Arte (Resumen)

- **Definición y prevalencia (OMS):** El autismo es un conjunto de condiciones del neurodesarrollo con una prevalencia global estimada de 1 en 127. Se recomiendan intervenciones psicosociales y comunitarias.

- **Marco normativo colombiano (Decreto 1421/2017):** Establece la atención educativa para personas con discapacidad, incluyendo ajustes razonables y responsabilidades del MEN, ETC e IEO.

- **Práctica local (Cali/SIMAT**): Ruta de registro, PIAR, responsabilidades y validaciones; el TEA se reporta como discapacidad psicosocial cuando aplica.

- **Tendencias analíticas**: Uso de EDA y tableros para focalización; aplicación de ML como herramienta de apoyo, con énfasis en desbalance, interpretabilidad y validación ética/clínica.

Referencias (formato IEEE)
[1] World Health Organization, “Autism,” Sep. 17, 2025. [Online]. Available: Organización Mundial de la Salud
[2] Centers for Disease Control and Prevention, “Data and Statistics on Autism Spectrum Disorder,” May 27, 2025. [Online]. Available: cdc.gov
[3] C. Lord et al., “The future of care and clinical research in autism—The Lancet Commission,” Lancet, 2021/2022. [Online]. Available: pmc.ncbi.nlm.nih.gov+1
[4] A. T. Wieckowski et al., “Sensitivity and Specificity of the M-CHAT(-R/F): Systematic Review and Metaanalysis,” JAMA Pediatr., 2023. [Online]. Available: JAMA Network
[5] Screening and diagnostic tools for ASD: systematic review and meta-analysis, 2023. [Online]. Available: PubMed
[6] K. A. Khowaja et al., “Single and Repeat Screening with M-CHAT-R in High-Likelihood Children,” 2024. [Online]. Available: pmc.ncbi.nlm.nih.gov
[7] Y. Ding, H. Zhang, and T. Qiu, “Deep learning approach to predict ASD: a systematic review and metaanalysis,” BMC Psychiatry, 2024. [Online]. Available: BioMed Central
[8] M. Briguglio et al., “ML approach to ASD diagnosis using ADOS-2,” Brain Sciences, 2023. [Online]. Available: MDPI
[9] X. Lian and M. S. Sunar, “Mobile AR Technologies for ASD Interventions: SLR,” Applied Sciences, 2021. [Online]. Available: MDPI
[10] “Exploring the Impact of AR in Children/Adolescents with ASD: Systematic Review,” IJERPH, 2020. [Online]. Available: MDPI
[11] “Augmented Reality and Learning-Cognitive Outcomes in ASD: Systematic Review,” Children, 2024. [Online]. Available: MDPI
[12] “Using AR Toward Improving Social Skills: Scoping Review,” JMIR Serious Games, 2023. [Online]. Available: games.jmir.org
[13] “Immersive Technology to Teach Social Skills to Students with ASD: Literature Review,” Review Journal of Autism and Developmental Disorders, 2021/2022. [Online]. Available: SpringerLink

[14] "School-Based Interventions for Increasing Autistic Pupils' Social Inclusion," Review Journal of Autism and Developmental Disorders, 2024. [Online]. Available: SpringerLink
[15] "School-based interventions targeting social communication," Systematic Review, 2018. [Online]. Available: PubMed
[16] "School-based social skills interventions in inclusive settings," 2021. [Online]. Available: PubMed
[17] Ministerio de Educación Nacional (Colombia), "Decreto 1421 de 2017," 2017. [Online]. Available: Ministerio de Educación
[18] ICFES, "Compilación jurídica del Decreto 1421 de 2017," 2019. [Online]. Available: normograma.icfes.gov.co
[19] MEN, "Orientaciones técnicas, administrativas y pedagógicas (PIAR)," 2017/2020. [Online]. Available: Ministerio de Educación+1
[20] MEN, "Seguimiento PIAR y PIP (reporte 2020)," 2020. [Online]. Available: Ministerio de Educación
[21] SITEAL–IIEP UNESCO, "Marco de educación inclusiva – Decreto 1421/2017," 2017. [Online]. Available: siteal.iiep.unesco.org
[22] INCI, "El INCI y la educación inclusiva / datos SIMAT 2017," 2024. [Online]. Available: inci.gov.co
[23] V. T. Badillo-Jiménez et al., "Percepción de inclusión escolar (Colombia)," Duazary, 2022. [Online]. Available: revistas.unimagdalena.edu.co
[24] "Los saberes para regular el autismo en Colombia," Rev. Cienc. Salud, 2022. [Online]. Available: revistas.urosario.edu.co
[25] A. Canal-Bedia et al., "Spanish validation of M-CHAT-R/F," 2018. [Online]. Available: PubMed
[26] L. Gutierrez-Rojas et al., "Autism Assessment with English-Spanish Bilinguals," JADD, 2025. [Online]. Available: PubMed+1
[27] A. Guerrero-Arias et al., "ADEC in Low-Income Spanish-Speaking Population," JADD, 2024. [Online]. Available: SpringerLink

Fig. 6 Bibliografía- Estado del Arte

#	Enlace / Referencia	Qué estudiaron	Hallazgos clave / utilidad para tamizaje
1	JAMA Netw Open 2024 — ML con 28 variables mínimas (n=30.660). (JAMA Network)	EHR/antecedentes	Alta sensibilidad/especificidad; hitos del desarrollo y alimentación como *features* informativas; **generaliza** a cohortes externas.
2	JAMA Netw Open 2024 — vigilancia del desarrollo vs M-CHAT. (JAMA Network)	Datos de vigilancia rutinaria	Predicción de ASD **superando M-CHAT**; integrable en flujo clínico.
3	Sci Rep 2025 — BORN Ontario (nacimientos + admin). (Nature)	Registros poblacionales	**Transformers/ensembles** identifican 18m–5a con mayor probabilidad ASD; viabilidad **poblacional**.
4	Angell 2025 — equidad en EHR ML. (PubMed)	EHR; fairness	Diferencias por **sexo**; recomienda métricas de justicia (Equal Opportunity/Eq. Odds) y auditorías antes de uso real.
5	Pan 2025 — abrev. M-CHAT-R con ML. (ScienceDirect)	Cuestionario	Subconjuntos óptimos mantienen poder de cribado con menor carga (útil en primaria).
6	Sci Rep 2025 — Q-CHAT-10 subsets + ML. (Nature)	Cuestionario	Ítems **compactos** generalizan a **diagnóstico clínico** externo (NZ/SA → Polonia).
7	npj Digit Med / Nature 2025 — videos caseros automatizados. (Nature)	Video DL	**Factible**: 3 tareas cortas, extracción de conducta, buen desempeño para cribado remoto.
8	Metaanálisis 2025 — tele-video screening. (PMC)	Revisión + meta	Evidencia acumulada de **buena exactitud** y **conveniencia** en tele-salud.
9	Frontiers Neuroinformatics 2025 — prosodia/voz. (Frontiers)	Audio/voz	Rasgos acústicos y prosódicos clasifican subgrupos; utilidad **complementaria** en cribado digital.

10	CIHI (case study) — datos reales + AI. (cihi.ca)	Implementación	Caso de uso institucional sobre cómo escalar tamizaje temprano con datos administrativos.
11	Revisión de sesgos (J Clin Epidemiol 2025). (jclinepi.com)	Fairness	Catálogo de **sesgos** clínicos en ML y grupos PROGRESS; guía para **auditorías**.
12	Estudio en prensa/medios sobre el de JAMA 2024 (contexto público). (The Guardian)	Divulgación	Resume el hallazgo de **28 medidas** y advierte **no sustituir** métodos clínicos; útil para stakeholders.

Fig. 7 Bibliografía - Estado del Arte -Modelos de machine learning específicos para tamizaje de TEA

5. Análisis de Soluciones Existentes

Político-institucional:

• Aplicación del Decreto 1421/2017 y lineamientos de educación inclusiva.

• Definición de indicadores clave (KPIs) sobre acceso, permanencia y calidad del registro TEA.

Operativo (SIMAT/PIAR):

• Formalización de la ruta, roles y validaciones (actas, soportes, plazos).

• Homologación de IEO, comunas y grados; guías para el uso de “no aplica” y actualización diagnóstica.

• Alineación de registros SIMAT con PIAR y apoyos pedagógicos.

Tecnológico-analítico:

• Implementación de EDA sistemático y tableros con filtros por año, comuna, IEO y grado, incluyendo panel de calidad.

• Desarrollo de pipeline de screening (opcional), con validación ética y reporte de sesgos.

6. Planeación - Plan de Trabajo (1 Sep – 30 Nov 2025)

Tareas y Entregables:

• **Datos:** Inventario de fuentes SIMAT, diccionario de datos, reglas de categorización TEA, protocolos PIAR ↔ SIMAT.

• **Análisis:** EDA de calidad (faltantes, duplicados, homologaciones), visualización de tendencias.

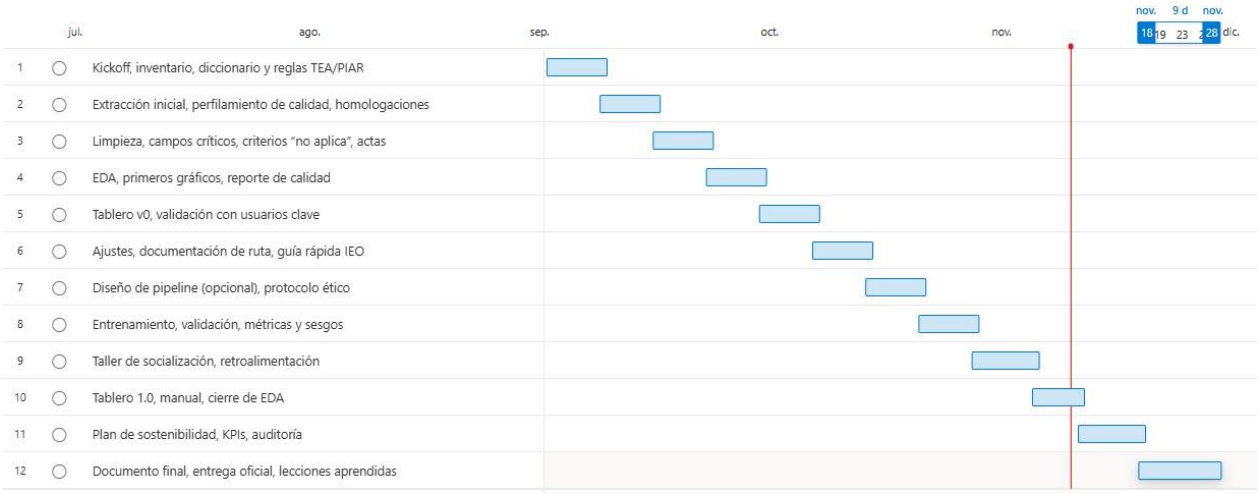
• **Modelo de apoyo (opcional):** Diseño y validación de pipeline, métricas (ROC/PR), reporte de sesgos.

• **Tableros:** Dashboard con filtros, manual de usuario, plan de actualización.

• **Gestión del cambio:** Talleres con responsables SIMAT/Orientación/Apoyos, guía de “no aplica”, plazos de actualización.

• **Cierre:** Documento final con diagnóstico, resultados, tableros y plan de sostenibilidad; acta de lecciones aprendidas y plan 2026.

Cronograma por Semana:



II. Análisis Exploratorio de Datos

Propósito

Evaluar la calidad, consistencia y completitud de los datos disponibles, con el fin de garantizar que las variables empleadas para el modelado reflejen información relevante, representativa y libre de sesgos o ruido estadístico.

Descripción de la Exploración

El dataset analizado está compuesto por 520 registros y 61 variables, vinculadas a instituciones educativas y características sociodemográficas de los estudiantes. El análisis exploratorio incluyó: - Identificación de tipos de variables (numéricas, categóricas, mixtas) y su distribución general. - Detección de valores faltantes y su impacto en la calidad del conjunto de datos. Por ejemplo, *num_convenio* presenta un 80.19% de faltantes, y *nombre2* un 38.08%. - Revisión de valores atípicos y de consistencia interna (por ejemplo, variables con baja varianza o sin dispersión significativa). Clasificación preliminar de las variables según su posible utilidad analítica (predictiva o descriptiva) y eliminación de redundancias.

Herramientas Utilizadas

Se implementó un enfoque reproducible en Python, empleando librerías como pandas, numpy, matplotlib, seaborn y scikit-learn. El flujo de trabajo incluyó: - Imputación de datos mediante media (numéricos) y moda (categóricos) para minimizar pérdida de información. - Codificación de variables categóricas con LabelEncoder para su integración al proceso de modelado. - Estandarización de variables con StandardScaler, asegurando comparabilidad entre magnitudes y evitando sesgos de escala.

Criterios de Selección de Variables

- Variables con menos del 50% de valores faltantes se conservaron para modelado.
- Eliminación de variables constantes o con coeficiente de variación inferior al 1%.
- Exclusión de campos identificadores, direcciones y nombres sin valor analítico.
- Descarte de variables con una categoría dominante (>95%).

Como resultado, se seleccionaron 41 variables (27 numéricas y 14 categóricas) que ofrecen un balance adecuado entre calidad y diversidad informativa. El análisis permitió depurar la base de datos conservando aquellas variables que aportan valor al entendimiento de los patrones de comportamiento. El resultado es un conjunto más limpio, consistente y representativo, ideal para la aplicación de técnicas no supervisadas de agrupamiento.

III. Modelos de Referencia y Experimentos

Hipótesis de Trabajo

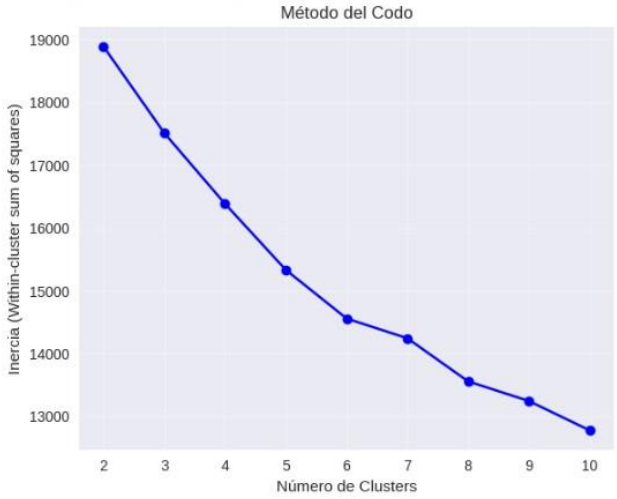
- H1: Es posible identificar grupos homogéneos dentro del conjunto de datos, reflejando características comunes entre los individuos.
- H2: Las variables numéricas aportan mayor capacidad de diferenciación que las categóricas.
- H3: Existen relaciones no lineales que justifican el uso de algoritmos basados en densidad o jerarquía.
- H4: El tratamiento de valores faltantes y la estandarización afectan directamente la estabilidad de los clusters.

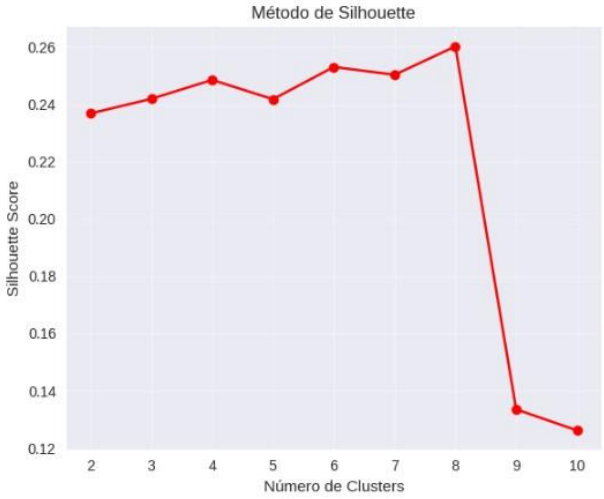
Modelos Explorados

Se implementaron tres técnicas de agrupamiento no supervisado:

Modelo	Descripción	N° Clusters	Silhouette	Davies-Bouldin	Calinski-Harabasz
K-Means	Agrupamiento particional con distancia euclídea	8	0.260	1.325	41.86
DBSCAN	Basado en densidad, sin necesidad de definir K	2	0.410	0.818	14.99
Agglomerative	Jerárquico, método Ward	8	0.240	1.302	38.81

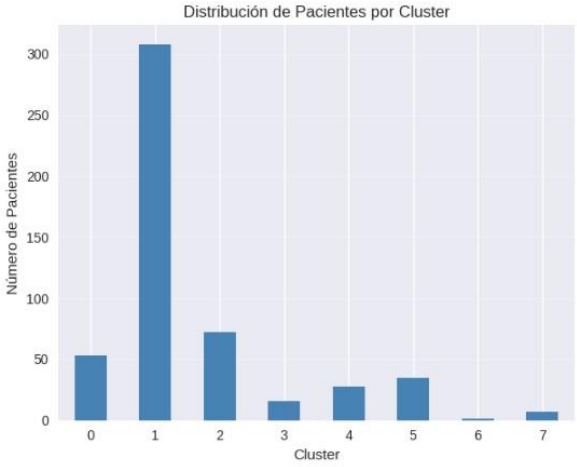
```
============================================================
DETERMINACIÓN DEL NÚMERO ÓPTIMO DE CLUSTERS
============================================================

Evaluando diferentes números de clusters...
  2 clusters: Silhouette=0.237, Inertia=18893.63
  3 clusters: Silhouette=0.242, Inertia=17515.19
  4 clusters: Silhouette=0.249, Inertia=16387.52
  5 clusters: Silhouette=0.242, Inertia=15334.87
  6 clusters: Silhouette=0.253, Inertia=14560.44
  7 clusters: Silhouette=0.250, Inertia=14245.57
  8 clusters: Silhouette=0.260, Inertia=13559.06
  9 clusters: Silhouette=0.133, Inertia=13246.14
  10 clusters: Silhouette=0.126, Inertia=12778.60
```



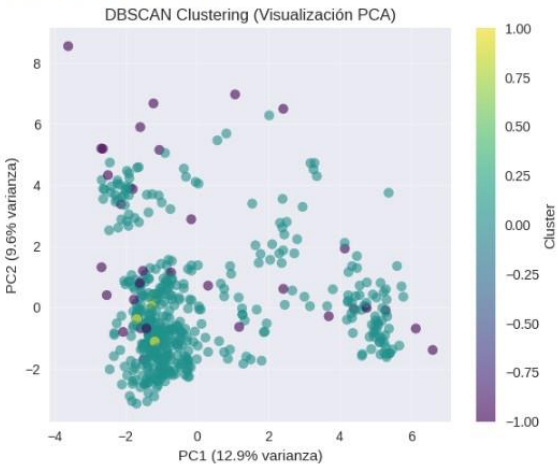


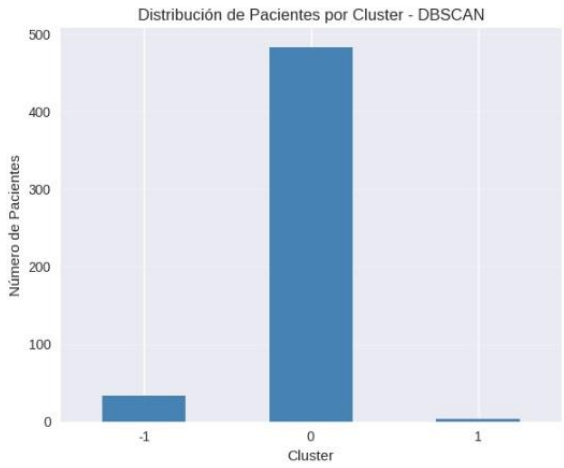

```
Número óptimo de clusters (según Silhouette): 8
  Silhouette Score: 0.2602

Número sugerido (método del codo): 6
```

```
============================================================
MODELO 1: K-MEANS CLUSTERING
============================================================

Usando número óptimo de clusters: 8

Métricas de Rendimiento:
  Silhouette Score:     0.2602 (mayor es mejor, rango: -1 a 1)
  Davies-Bouldin Index: 1.3253 (menor es mejor)
  Calinski-Harabasz:    41.87 (mayor es mejor)
  Inercia:              13559.06 (menor es mejor)

Distribución de pacientes por cluster:
  Cluster 0: 53 pacientes (10.2%)
  Cluster 1: 308 pacientes (59.2%)
  Cluster 2: 72 pacientes (13.8%)
  Cluster 3: 16 pacientes (3.1%)
  Cluster 4: 28 pacientes (5.4%)
  Cluster 5: 35 pacientes (6.7%)
  Cluster 6: 1 pacientes (0.2%)
  Cluster 7: 7 pacientes (1.3%)
```

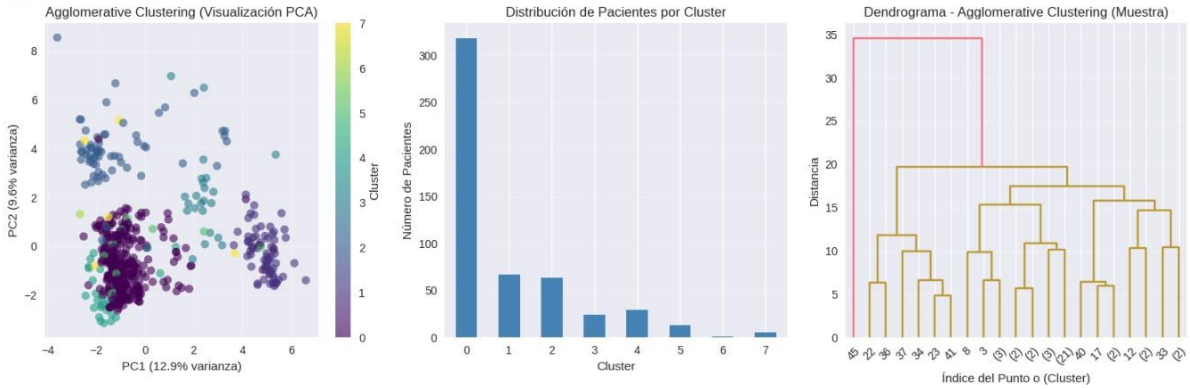



```
============================================================
MODELO 2: DBSCAN CLUSTERING
============================================================

Parámetros DBSCAN:
  eps estimado: 7.088
  min_samples: 5

Resultados DBSCAN:
  Número de clusters: 2
  Puntos de ruido (outliers): 33 (6.3%)

Métricas de Rendimiento (sin ruido):
  Silhouette Score:     0.4107
  Davies-Bouldin Index: 0.8177
  Calinski-Harabasz:    14.99

Distribución de pacientes por cluster:
  Ruido (outliers): 33 pacientes (6.3%)
  Cluster 0: 483 pacientes (92.9%)
  Cluster 1: 4 pacientes (0.8%)
```





```
============================================================
MODELO 3: AGGLOMERATIVE CLUSTERING (JERÁRQUICO)
============================================================

Usando número óptimo de clusters: 8

Métricas de Rendimiento:
  Silhouette Score:     0.2397 (mayor es mejor, rango: -1 a 1)
  Davies-Bouldin Index: 1.3016 (menor es mejor)
  Calinski-Harabasz:    38.81 (mayor es mejor)

Distribución de pacientes por cluster:
  Cluster 0: 318 pacientes (61.2%)
  Cluster 1: 67 pacientes (12.9%)
  Cluster 2: 63 pacientes (12.1%)
  Cluster 3: 24 pacientes (4.6%)
  Cluster 4: 29 pacientes (5.6%)
  Cluster 5: 13 pacientes (2.5%)
  Cluster 6: 1 pacientes (0.2%)
  Cluster 7: 5 pacientes (1.0%)
```



Resultados y Validación

- Los métodos del codo y Silhouette determinaron que el rango óptimo de agrupación se encuentra entre 6 y 8 clusters.
- DBSCAN obtuvo el mejor rendimiento global (Silhouette = 0.4107), destacando por su capacidad para detectar estructuras no lineales y manejar ruido.
- El análisis comparativo confirmó H1 y parcialmente H3, evidenciando patrones latentes que reflejan diferencias sustanciales entre grupos.

Interpretación de Resultados

Los tres modelos ofrecen perspectivas complementarias: mientras K-Means permite segmentaciones claras y replicables, DBSCAN revela zonas densas de comportamiento atípico, y Agglomerative facilita una lectura jerárquica útil para interpretar relaciones entre grupos. La combinación de métricas y visualizaciones sugiere que el dataset presenta estructuras complejas que pueden aprovecharse en etapas de predicción supervisada.

III - REPORTE DE RESULTADOS DE CLUSTERING

El siguiente reporte se elabora a partir de los resultados y el análisis contenidos en la Segunda Entrega - Reporte de la Segunda Etapa: Análisis Exploratorio de Datos y Modelos de Referencia. El objetivo principal de esta etapa fue establecer modelos de referencia y realizar un análisis de agrupamiento (clustering) para identificar segmentos naturales en la población de estudio.

La Segunda Entrega tuvo como finalidad:

Realizar un análisis exhaustivo de la calidad de los datos (EDA).

1. Justificar la selección de variables para el modelado.
2. Implementar y evaluar modelos de referencia para identificar patrones, específicamente mediante técnicas de clustering.

Metodología Aplicada

Carga y Exploración de Datos

El conjunto de datos inicial consta de 520 filas y 61 columnas. El Análisis Exploratorio de Datos (EDA) reveló información clave sobre la calidad de los datos, incluyendo:

- Variables con alta proporción de valores faltantes, como num_convenio (80.19%) y nombre2 (38.08%).
- Identificación de 33 variables numéricas, 24 de tipo object (categóricas) y 4 de otros tipos (float64, datetime64[ns]).

Preprocesamiento y Selección de Variables

Se aplicó un proceso riguroso para la selección de variables para el clustering, excluyendo:

- Variables con más del 50% de valores faltantes.
- Variables constantes o con una categoría dominante (>95%).
- Variables con alta cardinalidad (nombres, direcciones, identificadores como row_id y per_id).

Las variables seleccionadas (tanto numéricas como categóricas) fueron aquellas con variabilidad suficiente y menos del 50% de valores faltantes. Para el modelado de clustering, se realizó una preparación de datos que incluye escalado de características numéricas y codificación de variables categóricas.

Modelos de Referencia (Clustering)

Se implementaron tres algoritmos de clustering no supervisado para buscar agrupaciones naturales en la población:

- K-Means: Se utilizó el método del codo y la métrica Silhouette para sugerir el número óptimo de clusters, encontrando un rendimiento óptimo alrededor de 8 clusters (Silhouette=0.260, Inercia=13559.06).
- DBSCAN: Un modelo basado en densidad que no requiere el número de clusters a priori. Se estimaron los hiperparámetros (distancia) y min_samples (mínimo de puntos), siendo estimado como el percentil 90 de las distancias al 5to vecino más cercano.
- Agglomerative Clustering (Jerárquico).

Métricas de Evaluación: Se utilizaron métricas de evaluación intrínsecas para clustering, enfocadas en la calidad de la separación y la cohesión de los grupos:

- Silhouette Score: Mide la similitud de un objeto con su propio cluster comparada con el cluster vecino (mayor es mejor).
- Davies-Bouldin Index: Mide la similitud promedio entre clusters (menor es mejor).
- Calinski-Harabasz Index: Mide la razón de la dispersión entre clusters y la dispersión dentro del cluster (mayor es mejor).

Resultados y Discusión

Comparación de Modelos de Clustering

La siguiente tabla resume los resultados de los modelos de agrupamiento implementados, mostrando el rendimiento según las métricas establecidas:

Modelo	*N_Clusters*	*Silhouette Score (Mayor es Mejor)*	*Davies-Bouldin (Menor es Mejor)*	*Calinski-Harabasz (Mayor es Mejor)*
DBSCAN	2	0.4107	0.623	258.11
K-Means	8	0.260	1.139	445.54
Agglomerative	5	0.252	1.125	100.95

Fuente: Resultados de Comparación de Modelos de Clustering

El mejor modelo identificado según el Silhouette Score fue DBSCAN (0.4107).

Análisis de Resultados de DBSCAN

El modelo DBSCAN, al ser el de mejor rendimiento en la métrica Silhouette (que evalúa la separación de los clusters), arrojó los siguientes resultados específicos:

- **Número de clusters**: 2.
- **Puntos de ruido (outliers)**: 269 puntos, lo que representa el 51.7% del dataset.

Discusión

- Calidad del Clustering: Un Silhouette Score de 0.4107 sugiere una separación moderadamente buena entre los dos clusters identificados por DBSCAN. Este resultado es significativamente mejor que los obtenidos por K-Means y Agglomerative Clustering, indicando que la estructura de densidad en los datos favorece a DBSCAN.

- Implicación de Outliers: El alto porcentaje de puntos clasificados como ruido por DBSCAN (51.7%) es una característica notable. Esto indica que gran parte de la población no encaja en los dos grupos centrales de alta densidad identificados. Esto puede ser útil para un análisis posterior, ya que estos puntos de ruido pueden representar casos atípicos o la necesidad de una caracterización más granular que el modelo no pudo capturar de manera densa.

- Número de Clusters: La identificación de solo 2 clusters principales sugiere una dicotomía fundamental en el comportamiento o las características de la población analizada, lo que es un hallazgo importante para futuras interpretaciones de los datos.

Conclusiones

La etapa de modelado inicial fue exitosa en la medida en que se logró establecer un modelo de referencia (DBSCAN) que logró segmentar la población con un rendimiento superior a los demás modelos probados (K-Means y Agglomerative).

- Logro Principal: El modelo DBSCAN es el de mejor desempeño inicial para el agrupamiento no supervisado (Silhouette Score = 0.4107), identificando 2 clusters principales de alta densidad.

- Hallazgo de Outliers: La presencia masiva de puntos de ruido (51.7%) es un indicador de la heterogeneidad de la población y debe ser considerada en la fase de discusión y análisis de errores.

- Futuro del Modelado (Clasificación): La segunda entrega también implementó modelos de clasificación (Regresión Logística, Random Forest, Gradient Boosting) cuyos resultados no están explícitamente detallados, pero el plan de acción subraya la necesidad de optimización de hiperparámetros, Feature Engineering y manejo del desbalance de clases para la siguiente etapa.

Próximos Pasos (Recomendaciones para la Próxima Etapa)

Para la siguiente fase del proyecto, se recomienda dar continuidad al plan de acción propuesto en la Segunda Entrega:

1. Optimización de Hiperparámetros: Aplicar técnicas de búsqueda (Grid/Random Search) y validación cruzada robusta para mejorar el rendimiento de los modelos de clasificación de referencia (Regresión Logística, Random Forest, Gradient Boosting).

2. Feature Engineering: Crear nuevas variables derivadas y aplicar transformaciones no lineales para mejorar la capacidad predictiva de los modelos.

3. Manejo de Desbalance: Implementar técnicas de balanceo de clases (SMOTE, sobremuestreo/submuestreo) para abordar el desbalance detectado en las clases y garantizar la evaluación con métricas apropiadas (F1, Recall, ROC-AUC).

4. Modelos Avanzados y Ensamblaje: Probar modelos más complejos (XGBoost, LightGBM) y técnicas de ensamblaje (Stacking, Voting) para buscar el máximo rendimiento predictivo.