# Informe Ejecutivo de Análisis Exploratorio de Datos (EDA)

**Maestría en Inteligencia Artificial Aplicada**  
**Materia:** Análisis de Datos I  
**Profesor:** Ing. José Armando Ordoñez  
**Alumnos:** Edwin Pérez Lozano, Cristian Camilo Quebrada Bautista, Rubén Darío Sabogal  
**Fecha:** Septiembre de 2025 – Cali

## 1. Objetivo

El presente documento tiene como finalidad realizar un análisis exploratorio de datos (EDA) sobre el dataset *“POBLACIÓN TEA CALI.xlsx”*, extraído del sistema SIMAT. El propósito es comprender las características demográficas, socioeconómicas y educativas de la población estudiantil en Cali, con un énfasis especial en la detección temprana de casos relacionados con el Trastorno del Espectro Autista (TEA).

Este análisis constituye un insumo académico y práctico para la formulación de políticas públicas, el diseño de intervenciones focalizadas y la mejora de procesos de inclusión escolar en contextos urbanos diversos.

## 2. Resumen Ejecutivo

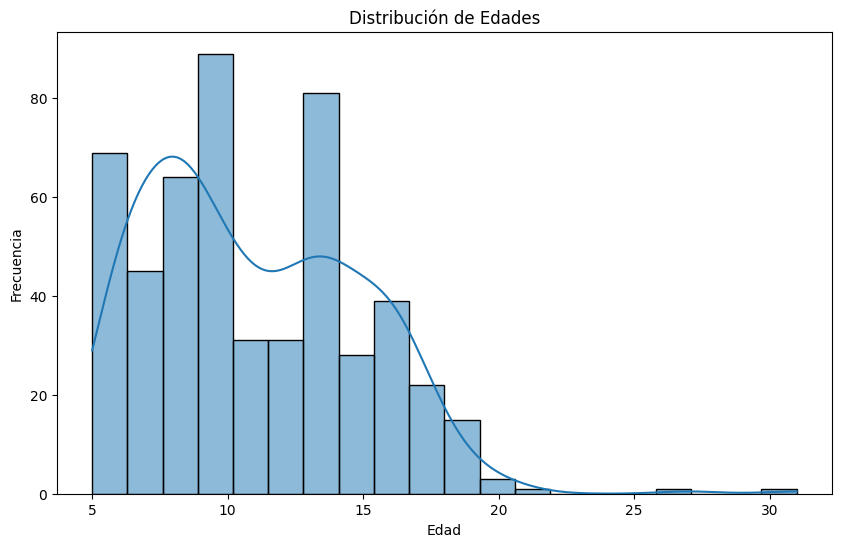
El análisis exploratorio realizado evidencia patrones significativos que aportan valor tanto desde la perspectiva de investigación académica como de política pública. Los resultados destacan la existencia de tres segmentos principales de estudiantes, diferenciados en edad, estrato y nivel escolar. Estos hallazgos reflejan la heterogeneidad de la población y sugieren la necesidad de estrategias diferenciadas para la detección temprana y el acompañamiento de estudiantes con TEA.

En términos de aportes prácticos, el análisis muestra cómo el uso de técnicas de *clustering* y segmentación puede guiar la toma de decisiones en educación y salud, brindando bases sólidas para una planeación estratégica y eficiente en el contexto de Cali.

## 3. Hallazgos Clave

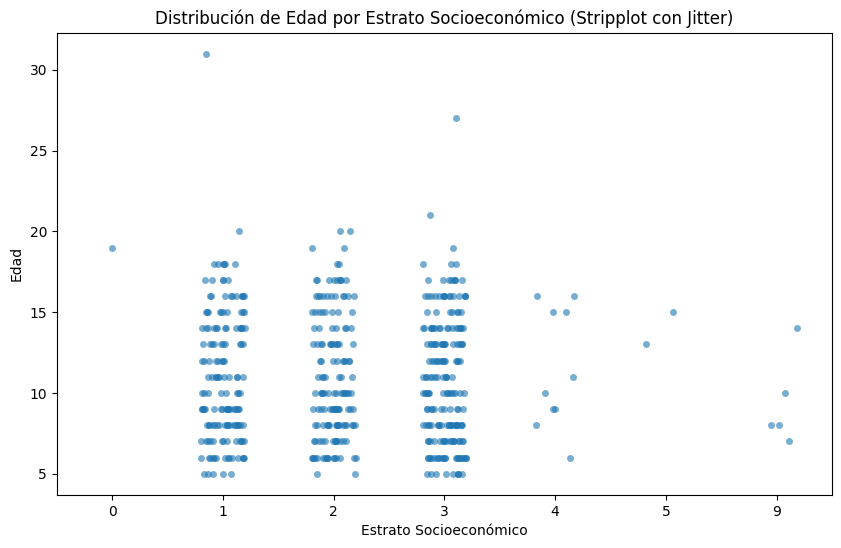
### 3.1 Análisis Univariado (Edad)

* Edad promedio: **10.97 años**.
* Mediana: **10 años**.
* Moda: **8 años**.
* Presencia de **outliers** de 27 y 31 años, que sugieren trayectorias educativas no convencionales o casos especiales.



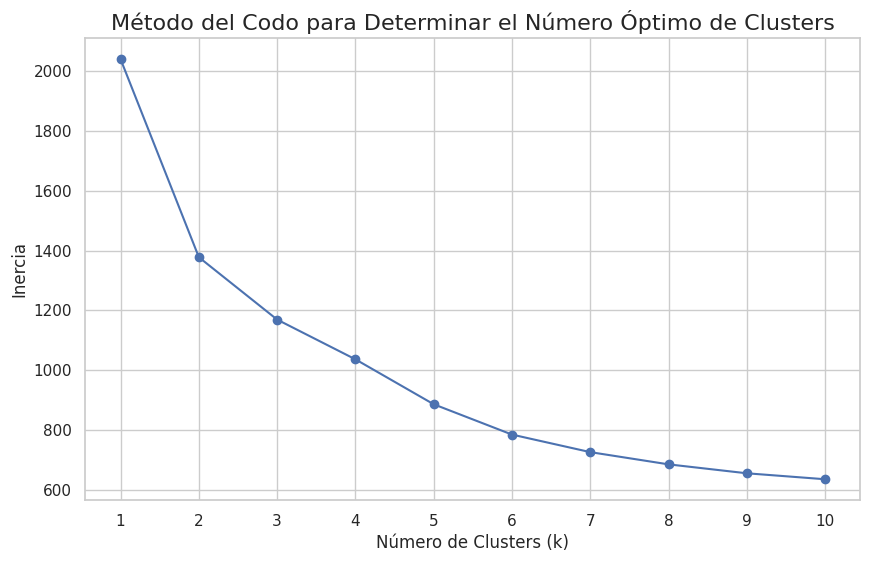
### 3.2 Análisis Bivariado (Estrato vs. Edad)

* La relación entre estrato y edad no es lineal.
* Los estudiantes se concentran en estratos 1, 2 y 3.
* Los casos atípicos de edad se localizan en estratos bajos, lo que refleja una posible desigualdad en trayectorias educativas.



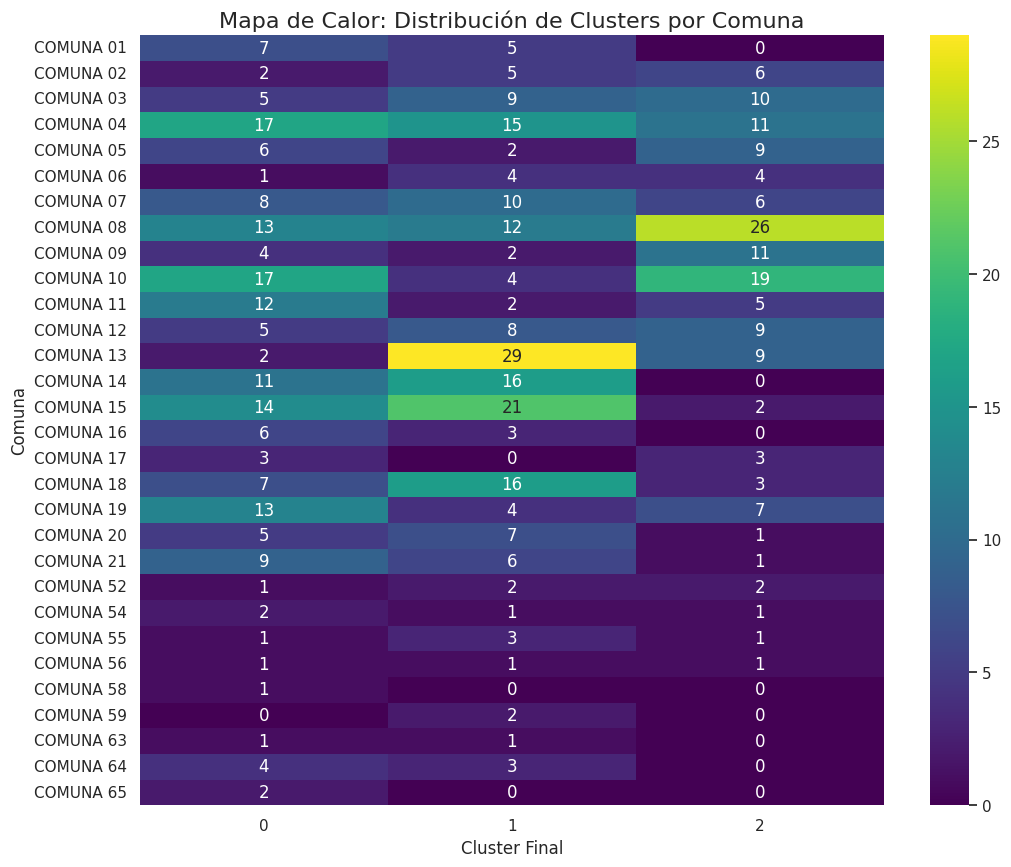
### 3.3 Análisis de Clustering (K-Means)

* **Cluster 0:** Estudiantes adolescentes (~15 años), en grados avanzados (~9), pertenecientes a estratos bajos y medios (~2).
* **Cluster 1:** Estudiantes más jóvenes (~8.5 años), en grados iniciales (~2), mayoritariamente en estratos bajos (~1.5).
* **Cluster 2:** Estudiantes jóvenes (~8.8 años), en grados iniciales (~2.3), con un estrato promedio más alto (~3.3).



## 4. Implicaciones para Política Pública

1. **Detección Temprana Focalizada:** Los Clusters 1 y 2 constituyen la población clave para programas de screening temprano.
2. **Intervenciones Diferenciadas:** El Cluster 0 requiere programas de apoyo académico y social, mientras que Clusters 1 y 2 demandan estrategias preventivas y educativas.
3. **Asignación de Recursos:** La identificación de comunas críticas permite priorizar psicólogos, terapeutas y docentes especializados.
4. **Casos Atípicos:** Los estudiantes fuera del rango típico de edad ameritan investigación particular para identificar barreras de acceso o permanencia escolar.
5. **Equidad y Acceso:** Las diferencias en estrato refuerzan la necesidad de garantizar igualdad de oportunidades en todos los niveles socioeconómicos.



## 5. Aportes Académicos y Científicos

* Este análisis integra técnicas estadísticas y de aprendizaje automático aplicadas al sector educativo, demostrando la pertinencia de la ciencia de datos en problemáticas sociales.
* El uso de K-Means como técnica de segmentación aporta un enfoque práctico para comprender la heterogeneidad en la población estudiantil.
* La detección de outliers no solo tiene relevancia técnica, sino también social, al evidenciar casos que pueden representar brechas de inclusión o rezago educativo.

## 6. Próximos Pasos Recomendados

* Incorporar nuevas variables relacionadas con apoyos académicos, salud y desempeño escolar
* Integrar datos externos (clínicos, socioeconómicos y culturales) para enriquecer los perfiles estudiantiles.
* Validar los resultados con expertos y actores clave en educación y salud.
* Utilizar los clusters identificados como insumo para modelos predictivos de riesgo en TEA.
* Diseñar planes de acción medibles que permitan aplicar las recomendaciones de forma escalable y sostenible.

## 7. Conclusión

Este informe demuestra el valor de la analítica de datos aplicada al ámbito educativo y de salud. La evidencia presentada refuerza la necesidad de implementar estrategias diferenciadas de detección temprana, asignación eficiente de recursos y diseño de políticas públicas inclusivas.

Desde la perspectiva académica, este trabajo refleja la aplicación de conocimientos adquiridos en la Maestría en Inteligencia Artificial Aplicada, fortaleciendo las competencias en análisis exploratorio de datos, segmentación y extracción de patrones útiles para la toma de decisiones.

## Marco teórico y justificación metodológica

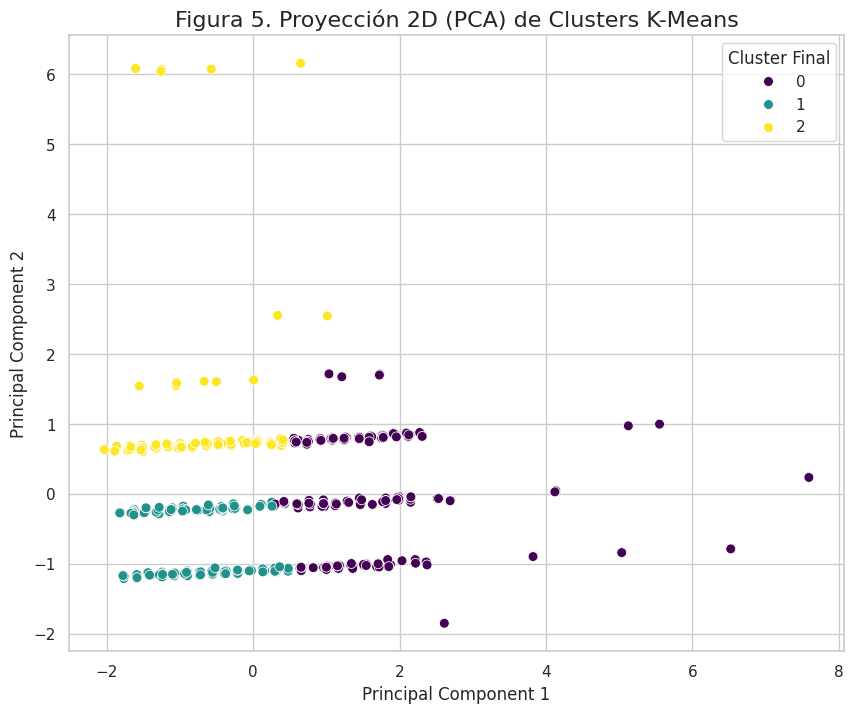
Este estudio se inscribe en la convergencia entre *educational data mining* (EDM) y analítica del aprendizaje para apoyar decisiones de política pública orientadas a la detección temprana del TEA. El uso de K‑Means sobre variables demográficas y educativas (Edad, ESTRATO, COMUNA y GRADO) permite una primera segmentación interpretable, complementada con índices internos de calidad de agrupamiento (Silhouette, Caliński–Harabasz) para elegir *k* y evaluar separación/compacidad. La consideración explícita de outliers es esencial en contextos administrativos como SIMAT, por su impacto en medidas de tendencia y dispersión, así como en la estabilidad de los centroides.

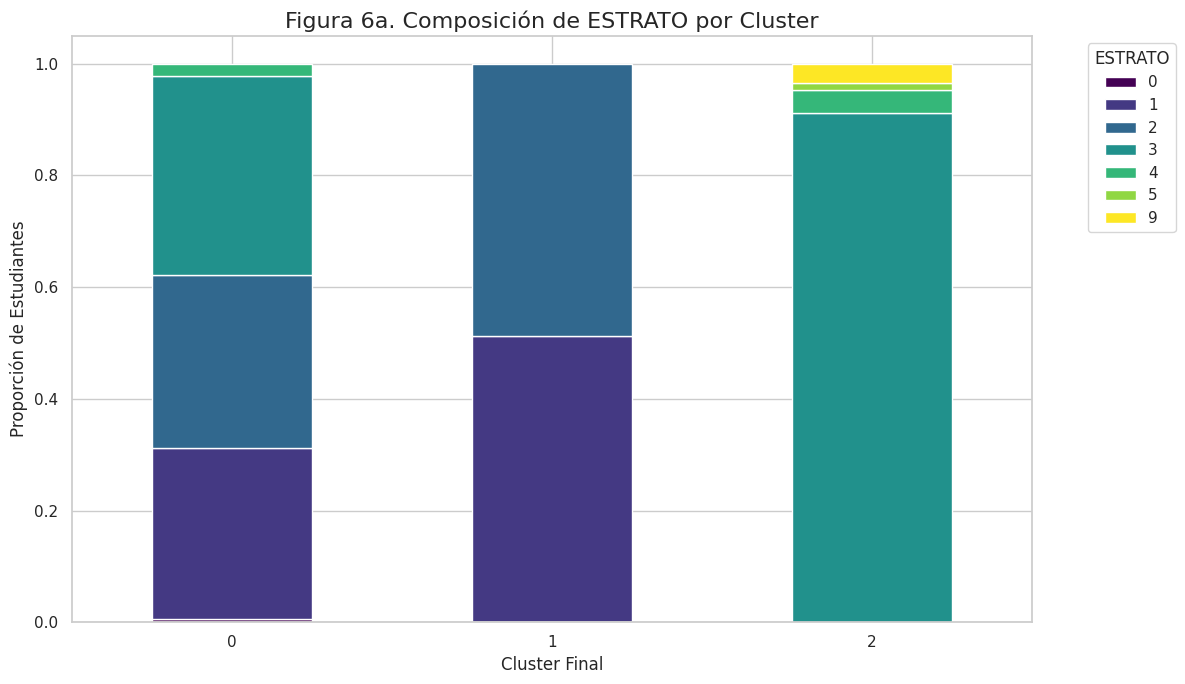
## Metodología ampliada (para reproducibilidad)

* **Depuración de datos:** perfilado de tipos, manejo de faltantes, estandarización de etiquetas de *ESTRATO/COMUNA/GRADO*, detección y tratamiento de outliers (reglas resistentes tipo Iglewicz–Hoaglin y análisis de influencias).
* **Selección de variables:** priorización de variables observables en SIMAT relevantes para detección temprana (edad, grado inicial, estrato, localización). Se propone incorporar en trabajos futuros apoyos pedagógicos, asistencia y reportes de necesidades educativas especiales.
* **Escalamiento:** estandarización (*z-score*) para variables numéricas y codificación para categóricas ordinales (GRADO) cuando aplique.
* **Clustering:** búsqueda de *k* en {2,3,4,5}; selección según Silhouette promedio, CH y *gap statistic* (si está disponible). Inicialización K‑Means++ con 25 *random states* para estabilidad.
* **Validación:** robustez de perfiles bajo *bootstrapping* estratificado por COMUNA/ESTRATO.

### Lugares reservados para figuras (insertar desde simat\_grupo.py)

1. **Figura 2.** Proyección 2D (PCA) con clusters K‑Means.



1. **Figura 3.** Barras apiladas: composición de ESTRATO/GRADO por cluster. 

## 

## Amenazas a la validez y consideraciones de equidad

* **Validez interna:** posibles errores de digitación en SIMAT, heterogeneidad en protocolos de registro y rezagos temporales.
* **Validez externa:** los clústeres describen la población SIMAT‑Cali y no se extrapolan automáticamente a otras ciudades sin recalibración.
* **Equidad:** diferencias de acceso y diagnóstico por estrato/territorio pueden sesgar la identificación temprana. Se recomienda auditoría de sesgos y monitoreo de métricas de equidad (p. ej., *equal opportunity*) en futuros modelos predictivos.

## Implicaciones operativas (detalle)

* **Búsqueda activa en clusters jóvenes (1 y 2):** jornadas de *screening* (M‑CHAT‑R/F) en grados de transición a primaria, con derivación oportuna a valoración clínica.
* **Soporte a Cluster 0 (adolescentes):** acompaño psicoeducativo y adaptaciones curriculares; articulación con servicios de salud mental y orientación vocacional.
* **Asignación de recursos:** priorización de comunas con alta concentración de perfiles de riesgo; fortalecimiento de capacidades docentes (detección y apoyos en aula).

## Plan de evaluación (KPIs e indicadores SMART)

* **Cobertura de *screening*:** ≥90% en grados iniciales en 12 meses.
* **Tiempo a valoración diagnóstica:** mediana <90 días desde *screening* positivo.
* **Retención escolar:** reducción de abandono en estudiantes TEA identificados vs. línea base.
* **Equidad territorial:** brecha de cobertura <10 p.p. entre comunas de mayor y menor estrato medio.

## Limitaciones y trabajos futuros

* Incorporar variables adicionales de SIMAT y enlazar con salud (SISPRO) y desempeño académico.
* Validar cualitativamente perfiles con equipos interdisciplinarios locales.
* Explorar métodos de clustering alternativos (GMM, k‑prototypes para datos mixtos) y detección de outliers basada en densidad.

## Referencias (formato IEEE)

Robins, K. Casagrande, M. Barton, C. M. Chen, T. Dumont‑Mathieu, and D. Fein, "Validation of the Modified Checklist for Autism in Toddlers, Revised With Follow‑Up (M‑CHAT‑R/F)," Pediatrics, vol. 133, no. 1, pp. 37–45, Jan. 2014, doi: 10.1542/peds.2013‑1813.

A. L. Siu and U.S. Preventive Services Task Force, "Screening for Autism Spectrum Disorder in Young Children: USPSTF Recommendation Statement," JAMA, vol. 315, no. 7, pp. 691–696, Feb. 2016, doi: 10.1001/jama.2016.0018.

C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," IEEE Trans. Syst., Man, Cybern., Part C (Appl. Rev.), vol. 40, no. 6, pp. 601–618, Nov. 2010, doi: 10.1109/TSMCC.2010.2053532.

C. Romero and S. Ventura, "Educational Data Mining and Learning Analytics: An Updated Survey," WIREs Data Mining Knowl. Discov., vol. 10, no. 3, p. e1355, May 2020, doi: 10.1002/widm.1355.

M. Ahmed, R. Seraj, and S. M. S. Islam, "The k‑means Algorithm: A Comprehensive Survey and Performance Evaluation," Electronics, vol. 9, no. 8, art. 1295, Aug. 2020, doi: 10.3390/electronics9081295.

K. A. Shaw et al., "Prevalence and Early Identification of Autism Spectrum Disorder Among Children Aged 4 and 8 Years — ADDM Network, United States, 2022," MMWR Surveill. Summ., vol. 74, no. SS‑2, pp. 1–22, Mar. 2025, doi: 10.15585/mmwr.ss7402a1.

R. Ferguson, "Learning Analytics: Drivers, Developments and Challenges," Int. J. Technol. Enhanced Learn., vol. 4, no. 5/6, pp. 304–317, 2012, doi: 10.1504/IJTEL.2012.051816.